UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
"MAT. LUIS MANUEL RIVERA GUTIÉRREZ"

PREDICCIÓN DE SERIES DE TIEMPO UTILIZANDO MÉTODOS DE INTELIGENCIA ARTIFICIAL

T E S I S

PARA OBTENER EL GRADO DE:

LICENCIADO EN CIENCIAS FÍSICO MATEMÁTICAS

PRESENTA:

DULCE NATALY SILVA MENDOZA

DIRECTOR DE TESIS:

DR. JOSÉ ANTONIO GONZÁLEZ CERVERA

MORELIA MICH. MARZO DE 2021

A mi familia.

AGRADECIMIENTOS

Al ver este trabajo de tesis concluido, vienen a mi mente cada suceso que me trajo hasta este punto, las decisiones, aquello que se sintió como éxito o como fracaso, las emociones y sentimientos que se apoderaban de mí durante el transcurso de la elaboración de esta tesis. Por supuesto también todas esas personas que estuvieron ahí apoyándome de diferentes maneras, sacándome una sonrisa, dándome un consejo, compartiendo su tiempo y conocimiento conmigo, apoyándome económica y emocionalmente, motivándome y creyendo en mí a pesar de los tropiezos. Es por esto que me gustaría agradecer a mi familia, amigos, compañeros, profesores y a esas personas que estaban de paso, pero que dejaron algo en el camino y que ahora son parte de mi persona, aunque no me sea posible mencionar a todas.

En primer lugar quiero agradecer a mis papás, hermanos y familia, porque siempre han estado ahí incondicionalmente a pesar de los malos ratos, nunca las palabras serán suficientes para agradecer tanto.

Sin lugar a duda, la elaboración de esta tesis no hubiera sido posibles sin las enseñanzas, el tiempo, la paciencia y los conocimientos impartidos por todos mis profesores. Agradezco especialmente a mi asesor el Dr. José Antonio González Cervera por aceptarme como tesista y por tanta paciencia. A mis sinodales la Dra. Karina Mariela Figueroa Mora, el Dr. Francisco Javier Dominguez Mota, el M.C. Jesús Ortiz Béjar y el Dr. Héctor Igor Pérez Aguilar por tomarse el tiempo de revisar esta tesis y por sus sugerencias. Al Dr. Jorge Luis López López, a la Dra. María Luisa Pérez Seguí y al Dr. Armando Sepúlveda por desarrollar mi gusto por las matemáticas. Al Dr. Petr Zevandrov, al Dr. Héctor Tejeda, al Dr. Francisco Siddhartha Guzmán, al Dr. Osvaldo Osuna, al Dr. Fernando Hernández, al Dr. Luis Valero y al Dr. Roberto García. Todos son ejemplo a seguir, los aprecio y los admiro.

No podían faltar las risas y el compañerismo que dan ánimo y sabor a la vida. Agradezco a Ricardo Chávez, Manuel Alejandro Espinoza y Jennifer López, mis compañeros y amigos de generación más apreciados. A Jorge Antonio Morales, Santiago Medrano, Iván Avilés, Isaac López y Ernesto Herrera, por su inigualable amistad y su apoyo constante a pesar de la distancia. A Guillermo Estrada, Ricardo Ochoa, Manuel Alejandro Romo de Vivar e incluso a Juan Salvador Alvarado, los quiero y los admiro. A Venecia Chávez, Sofía Zavala, Itzayana Guzmán y Melany Higuera, porque su sola presencia motiva e inspira. Agradezco también a Mauricio Carrillo, Francisco Rivera y Miguel Gracia, por su amistad, ayuda y recomendaciones que me brindaron. A mi par de amigos y compañeros colombianos,

Jeisson y Alejandro, porque no se cansan de darme ánimos y nunca dudan en ayudarme en lo que pueden. A Joel Chacón y a Ángel Balderas, porque siempre atienden a mis molestias a pesar de sus propias ocupaciones.

Morelia MEX., 7 de marzo de 2021

Índice general

RESUMEN

Durante este trabajo de tesis se estudiarán diferentes métodos y herramientas para analizar series de tiempo bajo la suposición de que éstas cumplen con el modelo clásico de la descomposición. Además, se dará una introducción a una clase de modelos lineales para series de tiempo estacionarias y no estacionarias, los modelos ARIMA. El objetivo de este trabajo es generar predicciones de la dirección del movimiento de series de tiempo, por lo que se describirán los predictores lineales para los modelos ARIMA y se estudiarán dos modelos de inteligencia artificial con los que se podrán realizar predicciones de series de tiempo. Los modelos de inteligencia artificial que se estudiarán en este trabajo son las redes neuronales artificiales (RNAs) perceptrón de tres capas de alimentación hacia adelante y las máquinas de soporte vectorial (MSV) para clasificación binaria no lineal, ambos modelos basados en el aprendizaje estadístico o supervizado. Finalmente, se analizarán un conjunto de series de tiempo, se realizarán predicciones de éstas con los diferentes modelos estudiados e implementados y se hará una comparación del desempeño de los modelos en cada problema.

Palabras clave: Modelos ARIMA, RNA, MSV, aprendizaje estadístico y ajuste del modelo.

ABSTRACT

This thesis will focus in the study of differents methods and tools dedicated to analyze time series under the assumption that these satisfy the classical descomposition model. Also, a class of linear models for stacionary and no stacionary time series, well known as ARIMA models, will be introduced. The aim of this work is to predict direction of time series movement using the ARIMA models and two kind of artificial inteligence models, the feed-forward three-layer perceptron artificial neural networks (ANN) and the support vector machines (SVM) for no linear binary classification, which are based in statistical or supervised learning. Finaly, by implementing the models under study, the predictions of the time series will be analyzed as well as the performance of the models.

Capítulo 1

Introducción

El análisis de *series de tiempo*, o bien, el análisis de datos experimentales que se han observado en diferentes puntos en el tiempo conduce a problemas de modelado e inferencia estadística, donde se busca proporcionar modelos matemáticos que den una descripción compacta y acertada de los datos que generan la serie de tiempo. Este tipo de problemas supone una correlación en el muestreo de puntos adyacentes en el tiempo, por lo que puede restringir severamente la aplicabilidad de los muchos métodos estadísticos convencionales que tradicionalmente dependen del supuesto de que estas observaciones adyacentes son independientes y están distribuidas de manera idéntica. Entre algunas de las aplicaciones más comunes del análisis de series de tiempo están: la eliminación de ruido en las series, el control de valores futuros ajustando parámetros, generar simulaciones y predecir valores futuros de la serie de tiempo.

El impacto del análisis de series de tiempo en aplicaciones científicas puede ser documentado en un conjunto particular de campos en la ciencia, donde han surgido problemas importantes. Por ejemplo, en economía surgen diferentes series de tiempo que continuamente se están exponiendo, como la cotizaición del mercado de valores o la cantidad de desempleados cada mes. Las ciencias sociales estudian las series de la población, como la tasa de natalidad o las inscripciones escolares. En epidemiología se interesan en la cantidad de infectados de una enfermedad en un periodo de tiempo. Algunos otros campos donde surgen importantes problemas con series de tiempo es en medicina, en física, en meteorología, en geología, en ingeniería, entre otros, ver [1].

El análisis de series de tiempo que será estudiado en este trabajo propone modelos para describir series de tiempo *estacionarias*, es decir, que describen una naturaleza *estocástica* o *no determinista* del experimento, pero que sus propiedades estocásticas como la media y la correlación entre los datos no dependen del tiempo.

Una de las tareas principales del análisis de series de tiempo es encontrar la distribución del proceso estocástico que generó la serie estacionaria. Por lo regular las series de tiempo de datos muestrales que encontramos de un experimento no son estacionarias, es decir, pueden contener alguna componente *tendencial*, *periódica* o estar bajo el efecto de alguna

transformación. Identificar si una series es estacionaria y transformar una serie de tiempo en una serie estacionaria, son otras de las tareas del análisis de series de tiempo, las cuales se valen de algunos métodos y herramientas, como la estimación de *correlaciones* para revisar la dependencia de los datos, el *ajuste de modelos*, el *suavizamiento* y la diferenciación de series, que son utilizados para llevar a cabo la estimación y eliminación de *componentes* que impiden que una serie sea estacionaria.

Algunos de los modelos que se proponen para modelar una serie estacionaria son: los modelos *autorregresivo de orden p* (AR(p) acrónimo del inglés autoregressive), *de promedios móviles de orden q* (MA(q) acrónimo del inglés moving averange), el compuesto por los dos anteriores, *autorregresivo de promedios móviles de orden p y q* (ARMA(p, q) del inglés autoregressive moving averange) y para series no estacionarias, el modelo *autorregresivo integrado de promedios móviles de orden p, d y q* (ARIMA(p, d, q) por sus siglas del inglés autoregressive integrated moving averange), los cuales pueden representar diferentes tipos de series de tiempo. Estos modelos tienen la característica de ser lineales y con una distribución gaussiana, sin embargo, las series de tiempo no siempre pueden ser descritas con modelos lineales, un ejemplo de este tipo de series son las series en finanzas, según [2, 3] *"La predicción del mercado de valores se considera una tarea desafiante del proceso de predicción de series de tiempo financieras, ya que el mercado de valores es esencialmente dinámico, no lineal, complicado, no paramétrico y catótico por naturaleza"*[*].

Para abordar este problema y aprovechando el desarrollo que ha tenido el cómputo científico, se propone hacer uso de algunos modelos de inteligencia artificial, como las redes neuronales artificiales (RNAs) y las máquinas de soporte vectorial (MSVs), pues según [4] *"La principal ventaja de las RNAs es su capacidad de modelado no lineal flexible"*[**]. Las RNAs y las MSVs son utilizadas para realizar diferentes tareas como clasificación, reconocimiento de patrones, regresiones, entre otros, ver [5, 6]. Estos modelos se caracterizan por el uso de funciones altamente no lineales conocidas como *Kerneles* y de algoritmos de aprendizaje estadístico supervisado que utilizan métodos de optimización numérica.

El objetivo de este trabajo de tesis es estudiar sobre el análisis de series de tiempo y diferentes métodos de hacer predicciones sobre las series temporales, con la intención de poder analizar, generear predicciones, y comparar el desempeño de los distintos métodos estudiados e implementados para desempeñar la tarea de predecir series de tiempo, tomando en cuenta solo los valores de pasados de la serie como se hace en [4]. En este caso se estará trabajando con los modelos en forma de clasificadores, como se hace en [2, 7], es decir, nuestro interés principal será poder predecir si una serie de tiempo incrementará o disminuirá su valor en el siguiete paso de tiempo con respecto a la posición actual.

Tanto el análisis de series de tiempo como los modelos que propone la inteligencia artificial son temas que requieren de un conocimiento previo de estadística, según [8] *"Usar*

[*]Traducido del inglés al español.
[**]Traducido del inglés al español.

herramientas sofisticadas como redes neuronales, boosting y máquinas de soporte vectorial sin comprender estadísticas básicas es como hacer una cirugía cerebral antes de saber cómo usar un curita"[***]. Es por esto que se dará una breve introducción en el resto de este capítulo a algunos temas de estadística que están basados en el contenido de [9], con la intención de ir familiarizando al lector con los términos y notación estadística que se estarán utilizando durante el desarrollo de esta tesis.

En el Capítulo 2 se hablará sobre métodos para analizar series de tiempo, ver si cumplen con el modelo de descomposición clásica y la obtención de series estacionarias. En el Capítulo 3 se describirán algunos de los modelos para series estacionarias y no estacionarias, formas de identificar qué modelo es recomendable utilizar y las predicciones que se logran con cada modelo. En los Capítulos 4 y 5 se estudiarán los modelos de RNAs y MSVs respectivamente. En el Capítulos 6 se encuentran los resultados de un conjunto de series de tiempo que fueron analizadas y modeladas, así como un problema de clasificación. Las conclusiones de este trabajo se pueden encontrar en el Capítulo 7. Finalmente, en el Apéndice A se dará una breve introduccion a temas de probabilidad y en el Apéndice B se abordarán algunos conceptos de optimización.

1.1. Estadística

El propósito de la estadística es hacer inferencias acerca de una población con base en información contenida en una muestra tomada de esa población, haciendo uso de herramientas de probabilidad. Debido a que las poblaciones están caracterizadas por medidas descriptivas numéricas llamadas *parámetros*, el objetivo de muchas investigaciones estadísticas es calcular el valor de uno o más parámetros relevantes. El cálculo, estimación o aproximación de un parámetro se lleva a cabo por medio de los que se conocen como *estimadores*, estos serán descritos a continuación.

1.2. Estimadores

Bajo la suposición de que una población está descrita por algún modelo que depende de ciertos parámetros, al parámetro de interés le llamaremos *parámetro objetivo* en el experimento. Se podría dar la estimación de los parámetros en dos formas distintas, la primera sería con solo un número, al que se llama *estimación puntual* y la segunda consta de dos números con los que se construye un intervalo, el cual tiene la intención de encerrar el parámetro de interés, a esta forma se le denota *estimación de intervalo*.

Las estimaciones se obtienen mediante los *estimadores*. Un *estimador* es una regla a menudo expresada como fórmula que indica como calcular el valor de una estimación con base en las mediciones contenidas en una muestra. En otras palabras, un estimador es un

[***]Traducido del inglés al español.

estadístico que se define como una función de *variables aleatorias* (v.a.s) observables en una muestra y de costantes conocidas. Por lo tanto, un estimador es también una variable aleatoria (v.a.) y tiene una distribución que llamaremos *distribución muestral*. En las secciones A.3 y A.4 del Apéndice A se pueden consultar las definiciones de v.a. y distribución de una v.a. respectivamente.

Cuando las observaciones se obtienen mediante muestreo aleatorio sin restitución a partir de una población finita, resulta que las observaciones son dependientes, sin embargo, se hacen esencialmente independientes si la población es grande comparada con el tamaño de la muestra. Para hacer inferencia estadística se supone que las poblaciones son grandes en comparación con el tamaño de la muestra, es decir, que las variables aleatorias obtenidas a través de muestreo aleatorio $Y_1, Y_2, \ldots, Y_n$ son independientes e identicamente distribuidas (IID). Para revisar el concepto de independencia de v.a.s ir a la sección A.9.

Se pueden obtener diferentes estimadores para un mismo parámetro objetivo. Cada estimador representa una regla subjetiva para obtener una sola estimación. Es por esto que se debe establecer criterios de bondad que nos permita decidir si el estimador es bueno o malo.

1.3. Estimadores puntuales

Supongamos que queremos especificar una estimación puntual para un parámetro objetivo que llamaremos θ. El estimador de θ estará expresado como $\hat{\theta}$, que se leé como "$\theta - gorro''$. El gorro indicará que se está estimando el parámetro que tiene debajo de él. A continución definiremos algunas propiedades deseables para los estimadores puntuales.

Una forma de evaluar la bondad de cualquier procedimiento de estimación puntual es en término de la distancia que existe entre la estimación y el parámetro objetivo, esta distancia se denomina *error de estimación* y está dada por

$$\epsilon = |\hat{\theta} - \theta|.$$

Por supuesto que nos gustaría que el error de estimación fuera tan pequeño como sea posible, sin embargo, esta cantidad varía aleatoriamente en muestreo repetido y la bondad de un procedimiento de estimación puntual no puede ser calculada con base en el valor de una sola estimación, más bien, debemos observar los resultados cuando el procedimiento de estimación se usa en innumerables veces, es decir, debemos evaluar la bondad del estimador puntual a partir de una distribución de frecuencia de los valores de las estimaciones obtenidas en muestreo repetido y observar cómo se agrupa esta distribución alrededor del parámetro objetivo.

Es altamente deseable que la distribución muestral se agrupe alrededor del parámetro objetivo de forma que

$$E(\hat{\theta}) = \theta. \tag{1.1}$$

Un estimador con la propiedad de la ecuación (1.1) se les llama *insesgado*, si $E(\hat{\theta}) \neq \theta$ entonces se dice que es *sesgado*. El *sesgo* de un estimador se define como

$$B(\hat{\theta}) = E(\hat{\theta} - \theta).$$

Además de estimadores insesgados, buscamos que la varianza del estimador $Var(\hat{\theta})$ sea lo más pequeña posible, es decir, si tengo dos estimadores insesgados diré que el mejor es el que tiene menor varianza. Ir a la sección A.8 para revisar el concepto de valor esperado, varianza y otros momentos.

A parte de usar el sesgo y la varianza de un estimador puntual para caracterizar su bondad, se podrían emplear otras medidas de bondad como el *error cuadrático medio* de un estimador $\hat{\theta}$, definido como

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]. \tag{1.2}$$

El siguiente teorema viene como ejercicio en [9] y aquí será demostrado.

Teorema 1.3.1. *El MSE está en función del sesgo y la varianza del estimador y cumple con la siguiente igualdad*

$$MSE(\hat{\theta}) = Var(\theta) + B(\hat{\theta})^2.$$

Demostración. Usando la identidad

$$(\hat{\theta} - \theta) = [\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta] = [\hat{\theta} - E(\hat{\theta})] + B(\hat{\theta}),$$

tenemos que

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= E[([\hat{\theta} - E(\hat{\theta})] + B(\hat{\theta}))^2] \\ &= E([\hat{\theta} - E(\hat{\theta})]^2) + 2E(\hat{\theta} - E(\hat{\theta}))B(\hat{\theta}) + B(\hat{\theta})^2 \\ &= Var(\hat{\theta}) + B(\hat{\theta})^2. \end{aligned}$$

□

Algunos ejemplos importantes de estimadores insesgados son la media muestral $\bar{Y}$ y la varianza muestral S^2 de la muestra de variables aleatorias $Y_1, Y_2, \ldots, Y_n$ que están dados respectivamente como sigue

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{y} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

1.4. Estimadores de intervalo

Un *estimador de intervalo* o *intervalo de confianza* es una regla que especifica el método para usar las mediciones muestrales en el cálculo de dos números que forman los puntos extremos de un intervalo. Idealmente un intervalo de confianza debe contener al parámetro objetivo y tener una amplitud relativamente pequeña. Dado que la amplitud y ubicación del intervalo son cantidades aleatorias, no podemos asegurar que el parámetro caerá dentro del intervalo calculado con una muestra en particular. Es por esto que el objetivo es encontrar estimadores capaces de generar intervalos estrechos con alta probabilidad de contener a θ.

Los puntos extremos superior e inferior de un intervalo de confianza se denominan *límites de confianza* superior e inferior, respectivamente. La probabilidad de que un intervalo de confianza (aleatorio) incluya a θ (una cantidad fija) se llama *coeficiente de confianza*. Suponga que los límites de confianza inferior y superior son $\hat{\theta}_I$ y $\hat{\theta}_S$ respectivamente, para un parámetro θ. Entonces, si

$$P(\hat{\theta}_I \leq \theta \leq \hat{\theta}_S) = 1 - \alpha, \tag{1.3}$$

la probabilidad $1 - \alpha$ es el coeficiente de confianza.

1.5. Método de los momentos

El *método de los momentos* es uno de los más antiguos para encontrar estimadores puntuales, a este método lo caracteriza la facilidad de su uso, pues parte de la idea de que los momentos poblacionales $\mu'_k = E(Y^k)$ de una variable aleatoria Y, están bien aproximados por su correspondiente k-ésimo momento muestral,

$$m'_k = \frac{1}{n} \sum_{i=1}^{n} Y_i^k.$$

Dado que los momentos poblacionales están en función de los parámetros poblacionales que se quieren estimar, el método consiste en que las estimaciones a dichos parámetros sean las soluciones del sistema de ecuaciones

$$m'_k = \mu'_k,$$

para $k = 1, 2, \ldots, t$, donde t es el número de parámetros a estimar. Se pueden revisar varios ejemplos de este método en la sección 9.6 de [9]. Las desventajas de este método es que muchas ocasiones proporciona estimadores con sesgo.

1.6. Método de máxima verosimilitud

El *método de máxima verosimilitud* es un método para obtener estimadores. Se dice que es un método útil puesto que con frecuencia se logran obtener estimadores insesgados y con varianza pequeña.

Sean $y_1, y_2, \ldots, y_n$ observaciones muestrales de las v.a.s correspondientes $Y_1, Y_2, \ldots, Y_n$, cuya distribución depende explícitamente de un parámetro θ. Entonces, si la distribución de las v.a.s $Y_1, Y_2, \ldots, Y_n$ es discreta, la *verosimilitud*, $L(y_1, y_2, \ldots, y_n|\theta)$, se define como la probabilidad conjunta de $y_1, y_2, \ldots, y_n$. Si la distribución de las variables aleatorias $Y_1, Y_2, \ldots, Y_n$ es continua, entonces la *verosimilitud*, $L(y_1, y_2, \ldots, y_n|\theta)$, es la densidad conjunta evaluada en $y_1, y_2, \ldots, y_n$. El concepto de distribuciones conjuntas puede ser revisado en la sección A.9.

Si el conjunto de variables aleatorias $Y_1, Y_2, \ldots, Y_n$ denota una muestra aleatoria de una distribución, es decir, que las variables aleatorias son IDD, entonces, en el caso de que la distribución sea discreta tenemos que

$$L(y_1, y_2, \ldots, y_n|\theta) = p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2)\ldots p(y_n),$$

mientras que si la distribución es continua

$$L(y_1, y_2, \ldots, y_n|\theta) = f(y_1, y_2, \ldots, y_n) = f(y_1)f(y_2)\ldots f(y_n).$$

Supongamos que que la función de verosimilitud depende de k parámetros $\theta_1, \theta_2, \ldots \theta_k$. El *método de máxima verosimilitud* escoge como estimaciones los valores de los parámetros que maximizan la verosimilitud $L(y_1, y_2, \ldots, y_n|\theta_1, \theta_2, \ldots \theta_k)$.

Para destacar que la función de verosimilitud es una función de los parámetros $\theta_1, \theta_2, \ldots \theta_k$ y se busca optimizar sobre ellos, a veces se expresa la verosimilitud como $L(\theta_1, \theta_2, \ldots \theta_k)$. A los estimadores $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_k$ obtenidos con este método se les llama *estimadores de máxima verosimilitud* y se abrevia MLE por sus siglas en inglés. Revisar los Ejemplos 9.14 y 9.15 de [9] para ilustrar el procedimieno del método. Los MLE para la media y la varianza de una distribución normal están dados por

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \text{y} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

respectivamente, en este caso la $\hat{\mu}$ es insesgado pero $\hat{\sigma}^2$ no.

1.7. Modelos determinísticos y probabilísticos

Hasta ahora hemos trabajado bajo la suposición de que las variables aleatorias $Y_1, \ldots, Y_n$ son IDD. Una implicación de esta suposición es que el valor esperado de Y_i es constante,

es decir, que $E(Y_i) = \mu$ para todo $i = 1, \ldots, n$. Esta suposición no es válida en muchos problemas inferenciales, como es el caso en problemas de series de tiempo. Ahora estudiaremos procedimientos inferenciales que pueden ser usados cuando tenemos una variable aleatoria Y llamada *variable dependiente* para la cual su valor esperado es una función de otras variables no aleatorias $x_1, \ldots, x_k$ llamadas *variables independientes* (hablando en el sentido matemático y no probabilístico).

Muchos tipos diferentes de funciones matemáticas se pueden usar para modelar una respuesta que sea una función de una o más variables independientes. Éstas se pueden clasificar en dos categorías: modelos determinísticos y probabilísticos.

Los *modelos determinísticos* son aquellos que no toman en cuenta ningún error para predecir o pronosticar el valor de la variable dependiente en términos de las variables dependientes. Un ejemplo es el modelo descrito en la siguiente ecuación

$$y = \beta_0 + \beta_1 x, \tag{1.4}$$

donde β_0 y β_1 son parámetros desconocidos, y y x son la variable dependiente e independiente respectivamente. En este tipo de modelos una vez fijando el valor de la variable independiente, el valor de la variable dependiente siempre será el mismo.

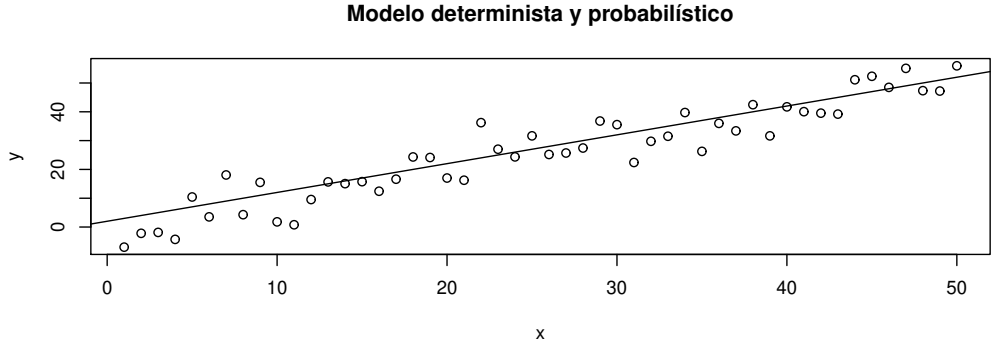

Figura 1.1: En esta figura se muesta como un modelo determinista intenta describir el comportamiento de un conjunto de datos.

En muchas ocasiones un modelo determinístico puede estar muy alejado de la realidad. Por ejemplo en la Figura 1.1 se muestra un conjunto de datos al que se le ajustó el modelo dado en la ecuación (1.4), aunque el valor esperado de los datos aumenta como el modelo, este no se ajusta totalmente. Los *modelos probabilísticos* buscan modelar la naturaleza que no se puede predecir en un fenómeno, agregan un factor aleatorio. Por ejemplo, un modelo más realista para los datos de la Figura 1.1 sería suponer que

$$E(Y) = \beta_0 + \beta_1 x,$$

o bien,

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

donde ϵ es una variable aleatoria con media cero y varianza σ^2.

1.8. Modelo lineal

Aun cuando se puede usar un número infinito de funciones diferentes para modelar el valor medio de la variable de respuesta Y como función de una o más variables independientes, nos concentraremos en un conjunto de modelos llamados *modelos estadísticos lineales*.

Un *modelo estadístico lineal* o *modelo de regresión lineal* que relaciona una respuesta aleatoria Y con un conjunto de variables independientes $x_1, x_2, \ldots, x_k$, es de la forma

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon, \tag{1.5}$$

donde $\beta_0, \beta_1, \ldots, \beta_k$ son parámetros desconocidos, ϵ es una variable aleatoria y $x_1, x_2, \ldots, x_k$ toman valores conocidos. Supondremos que $E(\epsilon) = 0$, por lo tanto

$$E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \tag{1.6}$$

Cuando hablamos de un modelo estadístico lineal, queremos decir que es lineal en función de los parámetros desconocidos $\beta_0, \beta_1, \ldots, \beta_k$. Desde un punto de vista práctico, se reconoce nuestra incapacidad para dar un modelo exacto por naturaleza.

En la siguiente sección usaremos el método de mínimos cuadrados para obtener estimadores de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$.

1.9. Método de mínimos cuadrados

Supongamos que el conjunto de observaciones $y_1, y_2, \ldots y_n$ está descrito por un modelo Y que está en función de un conjunto de parámetros $\beta_0, \beta_1, \ldots, \beta_k$ y una variable aleatoria. Nos gustaría encontrar estimadores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ para los parámetros desconocidos, donde $\hat{y}_1, \hat{y}_2, \ldots \hat{y}_n$ son las estimaciones hechas por el modelo estimador de Y para las observaciones $y_1, y_2, \ldots y_n$ respectivamente. El *método de mínimos cuadrados* genera aquellos estimadores que minimizan la *suma de errores cuadrados* (SSE por sus siglas en inglés) o *error empírico* (EE), es decir, que minimizan la ecuación

$$SSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Para ilustrar mejor este método, encontraremos los estimadores de una regresión lineal.

Suponga que tenemos el modelo lineal

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon \tag{1.7}$$

y hacemos n observaciones independientes $y_1, y_2, \ldots y_n$, en Y. Podemos escribir cada observación y_i como

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \tag{1.8}$$

donde x_{ij} es el ajuste de la j-ésima variable independiente para la i-ésima observación, $i = 1, 2, \ldots, n$. Ahora definamos las siguientes matrices

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_0 & x_{12} & x_{13} & \cdots & x_{1k} \\ x_0 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

donde $x_0 = 1$. Entonces, las n ecuaciones que representan las y_i en términos de $\mathbf{X}$,β y ϵ se pueden escribir en su forma matricial como

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \text{ o bien, } \hat{\mathbf{Y}} = \mathbf{X}\hat{\beta},$$

entonces, notemos que

$$n \cdot SSE = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta),$$

para optimizar esta función y encontrar los estimadores $\hat{\beta}$ que minimizan el SSE definamos $g(\beta) = (\mathbf{Y} - \mathbf{X}\beta)$ y $h(v) = v'v$ entonces $n \cdot SSE(\beta) = h \circ g(\beta)$ y por regla de la cadena o la Propiedad B.1.1 tenemos que

$$\begin{aligned} n \cdot DSSE(\beta) &= Dh(g(\beta))Dg(\beta) \\ &= -2g(\beta)'\mathbf{X} \\ &= -2(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{X}. \end{aligned}$$

Si $\hat{\beta}$ es el estimador que minimiza el SSE, entonces $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, pues

$$\begin{aligned} -2(\mathbf{Y} - \mathbf{X}\hat{\beta})'\mathbf{X} &= 0 \\ -2(\mathbf{Y}' - \hat{\beta}'\mathbf{X}')\mathbf{X} &= 0 \\ -2\mathbf{Y}'\mathbf{X} - 2\hat{\beta}'\mathbf{X}'\mathbf{X} &= 0 \\ \hat{\beta}'\mathbf{X}'\mathbf{X} &= \mathbf{Y}'\mathbf{X} \\ (\hat{\beta}'\mathbf{X}'\mathbf{X})' &= (\mathbf{Y}'\mathbf{X})' \\ \mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'\mathbf{Y}. \end{aligned}$$

En el caso en el que $\{\epsilon_i\} \sim N(0, \sigma^2)$ los MLE para la regresión lineal coinciden con los estimadores de mínimos cuadrados (LSE por sus siglas en inglés).

1.10. Máquinas de aprendizaje

Al igual que en estadística, en la inteligencia artificial se busca ajustar un modelo a un conjunto de datos, en este caso al conjunto de observaciones, $\{(x_i, y_i) \mid x_i \in \mathbb{R}^J, y_i \in \mathbb{R}, i = 1, \ldots, p\}$ se le llama *conjunto de etrenamiento* y a los vectores x_i y y_i *vectores de entrada* y *vectores de salida* respectivamente. Al modelo propuesto para representar los datos se le conoce como *máquina de aprendizaje*, se piensa como una función $f(x, \alpha)$ que está definida por un conjunto de posibles asignaciones $x \to f(x, \alpha)$ y está determinada por el parámetro ajustable α.

Al proceso elección de α que en estadística se le conoce como ajuste del modelo, aquí se le conoce como *entrenamiento* y una vez fijado el parámetro α la función $f(x, \alpha)$ recibirá el nombre de *máquina entrenada*. Al igual que en el modelo de regresión lineal, las máquinas de aprendizaje aprenden o se entrenan minimizando el error empírico

$$EE = \frac{1}{p}\sum_{i=1}^{p} (y_i - f(x_i, \alpha))^2 .$$

Así como en estadística, decimos que una máquina es determinista si dado un x y fijo un α el valor de $f(x, \alpha)$ siempre es el mismo.

En el siguiente capítulo estudiaremos más a detalle los modelos probabilísticos para series de tiempo, aunque a diferencia de este capítulo no estaremos suponiendo que el conjunto de $\{\epsilon_i\}$ son IID.

Capítulo 2

Análisis de series de tiempo

La recopilación de datos y el estudio o análisis de una serie temporal tienen como objetivo resolver problemas de modelación e inferencia, con el propósito de lograr una mejor comprensión del comportamiento del fenómeno que genera la serie temporal. A diferencia de la inferencia estadística, el análisis de series de tiempo supone dependencia entre los datos muestrales que integran la serie temporal. Existen diversas aplicaciones del análisis de series de tiempo, una de ellas es hacer predicciones de series temporales.

En este capítulo estudiaremos diferentes herramientas que nos ayudarán a obtener modelos adecuados para una serie de tiempo. Partiremos de la suposición de que una serie de tiempo cumple con el modelo de la descomposición clásica que está descrito en la sección 2.6, a grandes rasgos, este modelo supone que una serie temporal está compuesta por una serie que define la tendencia, una serie periódica y una serie irregular o estacionaria. El concepto de estacionariedad será introducido en la sección 2.4, este concepto es de suma importacia en el análisis de series temporales, pues existe diferentes modelos bien estudiados con los que se pueden dar buenas predicciones. Dos herramientas que serán de mucha utilidad son la covarianza y la correlación, ya que estas definen medidas de dependencia de los datos adyacentes y nos ayudará a revisar el comportamiento de los datos. El resto del capítulo se centrará en la obtención se series estacionarias, eliminación y estimación de tendencia y periodos.

Por conveniencia de la facilidad del manejo y disponibilidad de los datos estadísticos, a lo largo de este capítulo estaremos trabajando con las series de tiempo que maneja el paquete **astsa** y algunas de las funciones predeterminadas que vienen disponibles para el análisis de series de tiempo en el lenguaje de programación **R**. El contenido de este y el siguiente capítulo están basados principalmente en el contenido de [10] y [1].

2.1. Series de tiempo

Definición 2.1.1. *Una **serie de tiempo** es una colección de datos que fueron observados y etiquetados en intervalos equidistantes del tiempo, estas pueden ser vistas como una sucesión de valores* $\{x_t\}$*, donde* $t \in \{1, 2, ..., n\}$ *es el índice cronológico y n es el número de elementos que contiene la serie de tiempo.*

Por lo regular n es un número finito ya que en la mayoría de los casos se trabaja con datos experimentales y la obtención de éstos está limitada y aunque existen series de tiempo continuas éstas se tienen que discretizar para poder trabajar con ellas con métodos computacionales. Es muy importante tener la cantidad adecuada de datos que describan el fenómeno, de lo contrario se estará analizando algo que está fuera de la realidad.

Una parte importante del análisis de las series de tiempo es la selección de un modelo probabilístico apropiado para los datos el cual permitirá describir y dar lugar a la posibilidad de una naturaleza impredecible en el fenómeno que se busca describir o modelar.

Definición 2.1.2. *Un **modelo de series de tiempo** para el conjunto de datos observados* $\{x_t\}$*, es la especificación de una distribución conjunta (o posiblemente sólo su media y su varianza) de la secuencia de variables aleatorias* $\{X_t\}$*, postulando que* $\{x_t\}$ *es una realización de* $\{X_t\}$*.*

En general, una sucesión de variables aleatorias $\{X_t\}$ donde $t = 0, \pm 1, \pm 2, \ldots$ o algún subconjunto de los naturales se dice que es un *proceso estocástico*. Para fines prácticos, en ocasiones no se distingue entre la serie de tiempo $\{x_t\}$ y el proceso $\{X_t\}$ que la generó.

Un modelo completo de series de tiempo busca espesificar la distribución conjunta de las variables aleatorias $\{X_t\}$, es decir,

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n], \ \infty < x_1, \ldots, x_n < \infty, \ \ n = 1, \ldots.$$

A menos que los datos hayan sido generados con un mecanismo bastante simple, no es sencillo especificar la distribución de las variables aleatorias, esta es una de las razones por las que se utilizan solo el primero y segundo momento, otra razón es porque si se supone que los datos tienen una distribución normal multivariada sus predictores solo dependen de los primeros dos momentos, al igual que en los predictores lineales como podemos ver en la sección 3.8. Aunque en general se pierde información sobre la distribución bajo esta suposición, esto nos ayuda a tener aproximaciones.

Para comenzar con el análisis de una serie de tiempo se procede a graficar la serie. Una serie de tiempo puede ser graficada indicando el tiempo en el eje horizontal (hora, día, mes, año, etc.) o índice conológico y en el eje vertical el número de la medición que se obtuvo del fenómeno. Por ejemplo, veamos la gráfica de la Figura 2.1, donde se muestra la serie de tiempo de la tasa de desempleo de los Estados Unidos, registrada cada mes del año 1948 al 1978 con un total de 372 muestras.

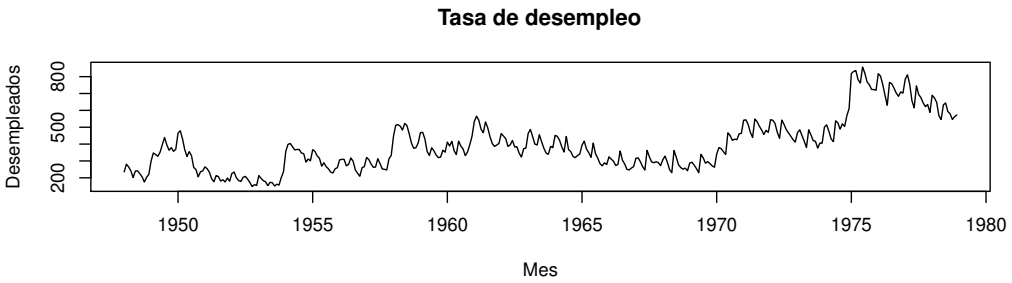

Figura 2.1: En este gráfico se observa en el eje de las ordenadas el número de personas desempleadas en los Estados Unidos mientras que en el eje horizontal vemos la linea de tiempo que señala el mes en el que se registró el dato.

Graficar una serie de tiempo nos permite revisar si tiene alguna característica en particular, como puntos aislados, cambio de comportamiento, si está bajo el efecto de alguna transformación o la presencia de alguna componente tendencial o cíclica como se mencionan en la Sección 2.6.

Como ya hemos mencionado, el análisis de serie de tiempo supone que una serie de tiempo es una realización de un conjunto de variables aleatorias que tienen cierta dependencia entre variables aleatorias que se encuentran a cierta distancia con respecto a los índices cronológicos de la serie tiempo. Además, la noción de serie de tiempo pretende que el comportamiento de esta sea repetitiva o regular a lo largo del tiempo, de aquí es que surge el modelo clásico de descomposición que será explicado en la Sección 2.6.

La falta de independencia entre dos valores X_s y X_t de una serie de tiempo puede evaluarse numéricamente, como en la estadística clásica, utilizando las nociones de covarianza y correlación que están descritas en la Sección A.10.4 del Apéndice A. En la siguiente sección introduciremos algunos modelos básicos de series de tiempo, los cuales modelan dependencia entre variables aleatorias que se encuentran a cierta distancia.

2.2. Modelos estadísticos

Probablemente el modelo más sencillo para una serie de tiempo es el modelo de *Ruido blanco*. Una serie de tiempo que cumple con el modelo de ruido blanco es una sucesión $\{w_t\}$ de variables aleatorias independientes con media 0 y varianza σ^2. Si $\{w_t\}$ es una serie de tiempo de ruido blanco se denota como $\{w_t\} \sim WN(0, \sigma^2)$.

En la Figura 2.2 podemos ver un ejemplo de una serie de tiempo de ruido blanco Gaussiano, esto significa que $\{w_t\} \sim IIDN(0, \sigma^2)$ son variablea aleatorias con media cero, varianza σ^2 e IID con distribución normal (se puede revisar la distrubución normal y otras distribuciones en la sección A.12.2 del Apéndice A).

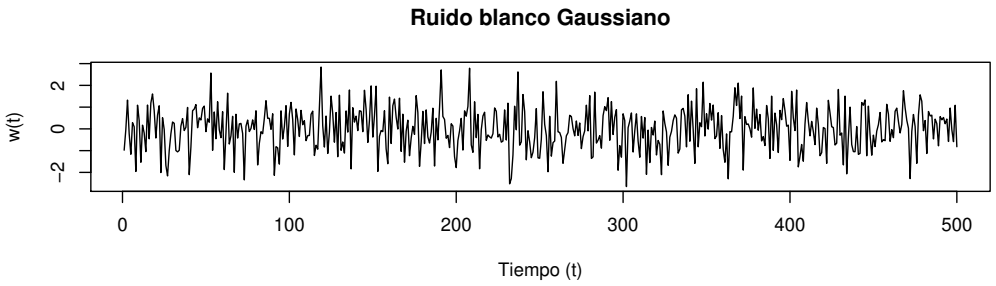

Figura 2.2: En la gráfica podemos ver la serie de tiempo de ruido blanco generada con 500 valores aleatorios independientes con distribución normal, de media cero y varianza uno.

El modelo de *promedios móviles*, es un modelo que se usa regularmente para suavizar series de tiempo, es decir, para eliminar el ruido de las series. El suavisamiento puede ser usado como un método para encontrar tendencias y periodos, como se muestra en la sección 2.7. Dada la serie de tiempo $\{w_t\}$ de ruido blanco, el modelo de promedios móviles usando tres valores está dado por la siguiente ecuación

$$v_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1}). \tag{2.1}$$

En la Figura 2.3 podemos visualizar la gráfica de serie de tiempo $\{v_t\}$, notemos que esta gráfica con respecto a la gráfica de la serie de tiempo de ruido blanco de la Figura 2.2 tiene menos picos, es decir, es más suave.

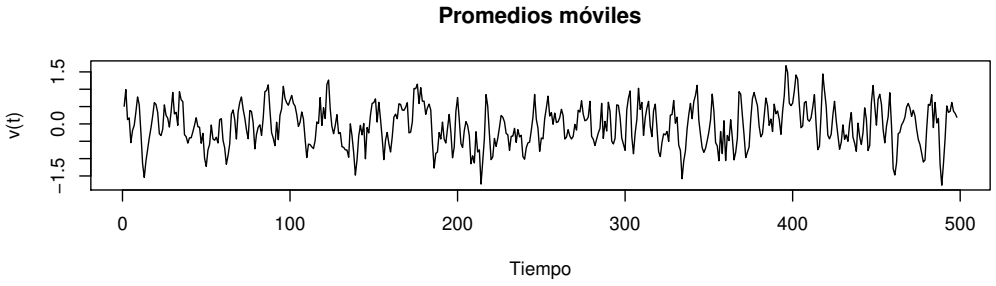

Figura 2.3: En esta gráfica se muestra el suvizamiento de la serie de tiempo de ruido blanco con el modelo ecuación (2.1).

El modelo *autorregresivo* supone que el comportamiento de una serie de tiempo $\{x_t\}$ depende de los valores previos de la serie más un valor aleatorio, este tipo de modelos serán mejor estudiados en el Capítulo 3. Un ejemplo de una serie de tiempo con el modelo

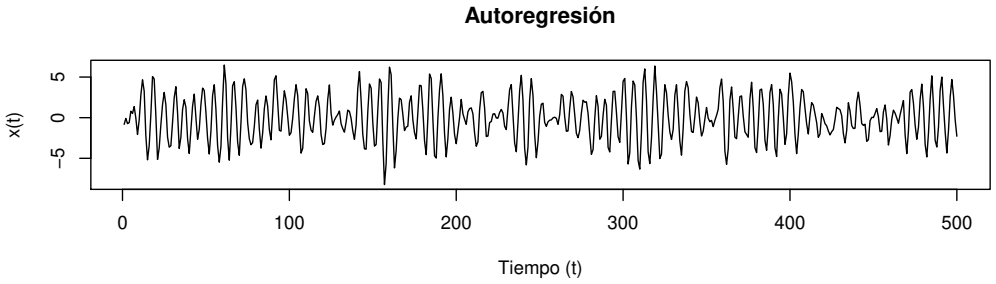

Figura 2.4: En este gráfico podemos ver la serie de tiempo generada con el modelo de la ecuación (2.2), con $x_0 = 0$ para inicializar el modelo.

autorregresivo está dada por

$$x_t = x_{t-1} + ,9x_{t-2} + w_t, \tag{2.2}$$

la gráfica de este modelo lo podemos ver en la siguiente Figura 2.4.

El modelo de *caminata aleatoria con desvío* es otro modelo para series de tiempo, un ejemplo de este modelo está dado en la ecuación (2.3), donde el parámetro δ se le conoce como desvío, este tipo de modelos podría usarse para modelar un comportamiento autorregresivo que tenga una linea recta como tendencia, pues nótese que si $\delta = 0$ entonces el modelo es simplemente un modelo autrorregresivo.

$$x_t = \delta + x_{t-1} + w_t \tag{2.3}$$

El modelo de la ecuación (2.3) se puede reescribir como

$$x_t = t\delta + \sum_{i=1}^{t} w_i, \tag{2.4}$$

suponiendo que $x_0 = 0$. En la Figura 2.5 podemos ver un ejemplo de una serie de tiempo generada con este modelo.

Existen algunos modelos más realistas que suponen algunos ciclos, por ejemplo

$$c_t = 2\cos\left(2\pi\frac{t+15}{50}\right) + w_t. \tag{2.5}$$

En la Figura 2.6 podemos ver tres series de tiempo generadas con el modelo de la ecuación (2.5), en las que solamente varía la varianza del ruido introducido.

En la siguiente sección serán introducidas las medidas de dependencia definidas para series de tiempo y se calculará la dependencia que tienen los datos generados con los modelos mencionados en esta sección.

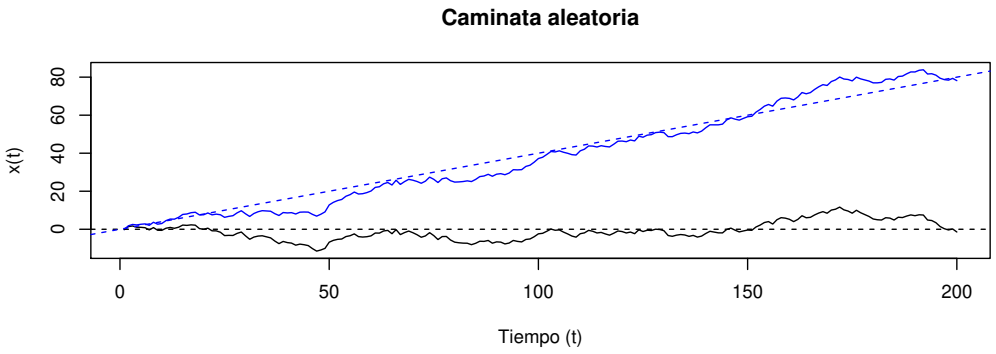

Figura 2.5: En esta gráfica se muestran dos series de tiempo, ambas generadas con el modelo de la ecuación (2.3). En la parte inferior, de color negro, podemos observar la serie de tiempo con $\delta = 0$ e inicializando $x_0 = 0$, mientras que en la parte superior y de azul está la serie de tiempo con parametro δ=0.4 y $x_0 = 0$ para inicializar.

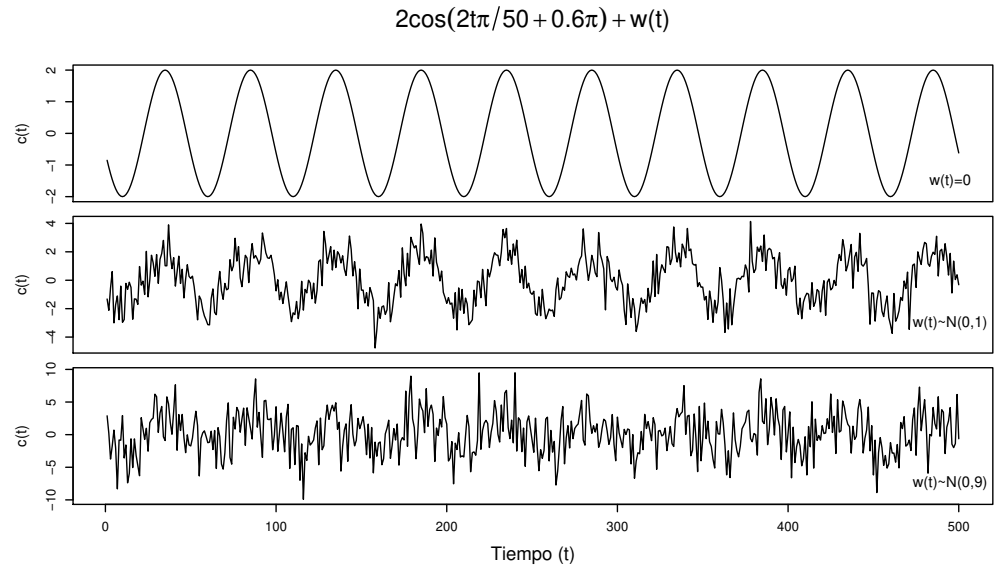

Figura 2.6: En esta figura se muestran tres gráficas de series de tiempo generadas con el modelo de la ecuación (2.5). En la parte superior se muestra la serie de tiempo con ausencia de ruido, mientras que a las de la parte media e inferior se les agregó ruido blanco con $\sigma^2 = 1$ y $\sigma^2 = 9$ respectivamente.

2.3. Medidas de dependencia

En la Sección A.10.4 del Apéndice A se decriben dos funciones para calcular o medir teóricamente la dependencia entre dos variables aleatorias que tienen una distribución conjunta, estas son: la covarianza y la correlación. Para el caso de un proceso estocástico o serie de tiempo $\{X_t\}$ estas medidas de dependencia serán definidas como sigue.

La *función de autocovarianza* entre dos valores de una serie de tiempo $\{X_t\}$, digamos X_s y X_t, se define como la covarianza entre las dos observaciones, es decir,

$$\gamma_X(s,t) = Cov(X_s, X_t).$$

Esta función mide la dependencia lineal que existe entre dos valores de una misma serie observados en diferentes tiempos, pero no muestra si existe dependencia de algún otro tipo. Se dice que X_s y X_t no están correlacionados si $\gamma_X(s,t) = 0$, en particular, si X_s y X_t son v.a.s independientes no existe correlación entre ellas (ver Teorema A.10.2), pero no se puede asegurar independencia solo por no existir correlación entre ellas (ver Ejemplo 5.24 de [9]). A menos que las variables X_s y X_t tuvieran una distribución normal bivariada esta afirmación sería cierta y es fácil de probar. Nótese que la función de autocovarianza en el mismo punto es simplemente la varianza, es decir, $\gamma_X(t,t) = Var(X)$.

Dado que la función de autocovarianza depende de la magnitud de la medición, no se puede determinar si la correlación entre dos valores de una serie de tiempo es grande o pequeña, para resolver ese problema se define la *función de autocorrelación* (ACF) definida para dos valores X_s y X_t de una serie de tiempo $\{X_t\}$ como

$$\rho_X(s,t) = \frac{\gamma_X(s,t)}{\sqrt{\gamma_X(s,s)\gamma_X(t,t)}},$$

la cual cumple que $-1 \leq \rho(X_s, X_t) \leq 1$ (ver Teorema A.10.4 para revisar la demostración). Esta medida muestra qué tan predecible es el valor X_t dado el valor X_s de una serie de tiempo.

Se puede medir la dependenncia entre dos valores de dos series de tiempo o procesos estocásticos diferentes. Sean $\{X_t\}$ y $\{Y_t\}$ series de tiempo, X_s y Y_t valores pertenecientes a las series respectivamente, entonces la *función de covarianza cruzada* se define como

$$\gamma_{XY}(s,t) = Cov(X_s, Y_t),$$

mientras que la *función de correlación cruzada* (CCF) está dada por

$$\rho_{XY}(s,t) = \frac{\gamma_{XY}(s,t)}{\sqrt{\gamma_X(s,s)\gamma_Y(t,t)}}.$$

Para ilustrar las medidas de dependencia veremos algunos ejemplos del cálculo de la media, covarianza y correlación con las series de tiempo mencionadas en el capítulo ante-

rior.

Ejemplo 2.3.1. *Para el caso de la serie de tiempo de ruido blanco* $\{w_t\}$ *tenemos que* $\mu_{w_t} = E(w_t) = 0$ *y* $Var(w_t) = \sigma^2$ *por definición, además como todas las v.a.s de esta serie son independientes, tenemos que*

$$\gamma_w(s,t) = \begin{cases} \sigma^2 & si \;\; s = t \\ 0 & si \;\; s \neq t \end{cases} \quad y \;\; \rho_w(s,t) = \begin{cases} 1 & si \;\; s = t \\ 0 & si \;\; s \neq t \end{cases}.$$

Ejemplo 2.3.2. *Para el caso de la serie de tiempo generada con el modelo de promedios móviles dado por la ecuación (2.1) tenemos que usando el Teorema A.9.1 la media es*

$$\mu_{v_t} = E(v_t) = \frac{1}{3}(E(w_{t-1}) + E(w_t) + E(w_{t-1})) = 0.$$

Haciendo uso del Teorema A.10.3 tenemos que la covarianza para dos valores de la serie de tiempo $\{v_t\}$ *está dada como*

$$\gamma_v(s,t) = Cov(v_s, v_t) = \sum_{i=-1}^{1}\sum_{j=-1}^{1}\frac{1}{9}Cov(w_{s+i}, w_{t+j}).$$

Para terminar el cálculo de la covarianza, es necesario dividir el problema en varios casos.

Cuando $s = t$*, tenemos que:*

$$\gamma_v(t,t) = \sum_{i=-1}^{1}\sum_{j=-1}^{1}\frac{1}{9}Cov(w_{t+i}, w_{t+j}) \tag{2.6}$$

$$= \frac{3}{9}\sigma^2, \tag{2.7}$$

para pasar de la ecuación (2.6) a (2.7) notemos que $Cov(w_{t+i}, w_{t+j}) = \sigma^2$ *si* $i = j$ *y cero en otro caso, por ser independientes, por lo tanto, si* $i = j$ *hay tres componentes de la sumatoria que son igual a* σ^2*.*

Si $s = t + 1$ *entonces*

$$\gamma_v(t+1,t) = \sum_{i=-1}^{1}\sum_{j=-1}^{1}\frac{1}{9}Cov(w_{t+1+i}, w_{t+j}) \tag{2.8}$$

$$= \frac{2}{9}\sigma^2, \tag{2.9}$$

la justificación de pasar de la ecuación (2.8) a (2.6) es análoga al caso $s = t$*.*

Para $s = t - 1$ *y* $s = t + 2$ *el cálculo es análogo a los anteriores, obteniendo que* $\gamma_v(t-1,t) = \frac{2}{9}\sigma^2$*,* $\gamma_v(t-2,t) = \frac{1}{9}\sigma^2$ *y cuando* $|s - t| > 2$ *tenemos que* $\gamma_v(s,t) = 0$*. Por lo tanto tenemos que*

$$\gamma_v(s,t) = \begin{cases} \frac{3}{9} & si \;\; s = t \\ \frac{2}{9} & si \;\; |s-t| = 1 \\ \frac{1}{9} & si \;\; |s-t| = 2 \\ 0 & si \;\; |s-t| > 2 \end{cases} \quad y \;\; \rho_v(s,t) = \begin{cases} 1 & si \;\; s = t \\ \frac{2}{3} & si \;\; |s-t| = 1 \\ \frac{1}{3} & si \;\; |s-t| = 2 \\ 0 & si \;\; |s-t| > 2 \end{cases}.$$

Ejemplo 2.3.3. *Cuando estamos trabajando con la serie de tiempo de caminata aleatoria con desvio, considerando la serie generada con el modelo de la ecuación (2.4) tenemos que por el Teorema A.9.1*

$$\mu_{x_t} = E(x_t) = t\delta + \sum_{i=1}^{t} E(w_i) = t\delta,$$

por lo tanto tenemos que la autocovarianza está dada como

$$\gamma_x(s,t) = Cov\left(\sum_{i=1}^{s} w_i, \sum_{i=1}^{t} w_i\right) = \sum_{i=1}^{s}\sum_{j=1}^{t} Cov(w_i, w_j) = \text{mín}(s,t)\sigma^2$$

y la función de autocorrelación es $\rho_x(s,t) = \text{mín}(s,t)/\sqrt{st}$.

Ejemplo 2.3.4. *Para el caso de la serie de tiempo generada con la ecuación (2.5) tenemos que su valor esperado está dado como*

$$\mu_{c_t} = E(c_t) = E\left(2\cos\left(2\pi\frac{t+15}{50}\right) + w_t\right) = 2\cos\left(2\pi\frac{t+15}{50}\right) \tag{2.10}$$

miestras que las funciones de correlación y covarianza quedan como en el ejemplo de ruido blanco.

Como notamos en algunos ejemplos, existen modelos para los que la media es constante y la función de covarianza solo depende de la distancia entre los valores, a los modelos que cumplen esta característica se les conoce como modelos o series estacionarias. En el siguiente capítulo estudiaremos mejor este tipo de series.

2.4. Series estacionarias

Se dice que una serie de tiempo *estrictamente estacionaria* $\{X_t\}$ es aquella en la que el comportamiento probabilístico para cada colección de valores $\{x_{t_1}, x_{t_2}, ..., x_{t_k}\}$ es exactamente igual en los valores desplazados $\{x_{t_1+h}, x_{t_2+h}, ..., x_{t_k+h}\}$, es decir,

$$P(x_{t_1} \leq c_1, x_{t_2} \leq c_2, ..., x_{t_k} \leq c_k) = P(x_{t_1+h} \leq c_1, x_{t_2+h} \leq c_2, ..., x_{t_k+h} \leq c_k),$$

para cada tiempo $t_1, t_2, ...t_k$, cada constante $c_1, c_2, ..., c_k$ y cada desplazamiento $h \in \mathbb{Z}$.

Es difícil encontrar series estrictamente estacionarias, es por esto que surge el concepto de series de tiempo *débilmente estacionarias* o simplemente *estacionarias*.

Definición 2.4.1. *Una serie de tiempo $\{X_t\}$ es **estacionaria** si sus propiedades estadísticas no cambian con el tiempo, es decir, si se cumplen las siguientes propiedades:*

Propiedad 2.4.1. $\sigma^2_{X_t} < \infty, \ \forall t \in \mathbb{Z}$.

Propiedad 2.4.2. $\mu_{X_t} = \mu, \quad \forall t \in \mathbb{Z}$.

Propiedad 2.4.3. $\gamma_X(s,t) = \gamma_X(s+r, t+r), \quad \forall s,t,r \in \mathbb{Z}$.

Si $\{X_t\}$ es una serie estacionaria, entonces la Propiedad 2.4.3 implica que $\gamma_X(s,t)$ solo depende de la distancia $|t-s| = h$ de los elementos de la serie, entonces podriamos redefinir la función de autocovarianza para las series estacionarias en función de h, como

$$\gamma_X(h) = \gamma_X(s, s+h).$$

De igual manera se puede definir la función de autocorrelación para series estacionarias como

$$\rho_X(h) = \rho_X(s, s+h) = \frac{\gamma_X(s, s+h)}{\sqrt{\gamma_X(s,s)\gamma_Y(s+h, s+h)}} = \frac{\gamma(h)}{\gamma(0)}.$$

Algunas de las propiedades que cumplen las funciones de autocorrelación y autocovarianza para procesos estacionarios las mencionaremos a continuación, la demostración de estas propiedades se pueden encontrar en la sección 2.1 de [11].

Propiedad 2.4.4. $\gamma(0) \geq 0$.

Propiedad 2.4.5. $|\gamma(h)| \leq \gamma(0) \ \ \forall h$.

Propiedad 2.4.6. $\gamma(h) = \gamma(-h) \ \ \forall h$.

Propiedad 2.4.7. *Una función real definida en los enteros es la autocovarianza de una proceso estacionario si y solo si es par y no negativa definida.*

Propiedad 2.4.8. $\rho(0) = 1$.

A partir de la definición de serie estacionaria podemos ver que la serie de tiempo de ruido blanco $\{w_t\}$ es estacionaria, pues como fue calculado en el Ejemplo 2.3.1, para cada t se tiene que $\mu_{w_t} = 0$ y que

$$\gamma_w(h) = \begin{cases} \sigma^2 & si \ \ h = 0, \\ 0 & si \ \ h > 0. \end{cases}$$

También se puede verificar por el Ejemplo 2.3.2 que la serie de tiempo de promedios móviles es estacionaria, pues $\mu_{v_t} = 0$ para todo t y

$$\gamma_v(h) = \begin{cases} \frac{3}{9}\sigma^2 & si \;\; h = 0, \\ \frac{2}{9}\sigma^2 & si \;\; h = 1, \\ \frac{1}{9}\sigma^2 & si \;\; h = 2, \\ 0 & si \;\; h > 2. \end{cases}$$

Sin embargo, la serie de tiempo de caminata aleatoria analizada en el Ejemplo 2.3.3 no es estacionaria pues $\mu_{x_t} = t\delta$ y esta depende del tiempo, para el caso en el que $\delta = 0$ cuando el modelo es autorregresivo tampoco es estacionaria, pues la función de covarianza también depende de t. La estacionariedad de los modelos autorregresivos será mejor estudiada en el Capítulo 3. Otro modelo que genera series no estacionarias es el caso del modelo de la ecuación (2.5), pues como vimos en el Ejemplo 2.3.4 su media μ_{c_t} expresada en la ecuación (2.10) depende de t.

Aunque las funciones teóricas de autocorrelación y correlación cruzada son útiles para describir las propiedades de ciertos modelos hipotéticos, la mayoría de los análisis deben realizarse utilizando datos muestreados $x_1, \ldots, x_n$. A continución se definirán las medidas de dependencia muestral para series de tiempo.

2.5. Correlación muestral

Tomando en cuenta que la media de las v.a.s de un proceso estacionario es constante, entonces la *media muestral* se puede definir como

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

el cual ya se ha mencionado en el Capítulo 1 es un estimador insesgado.

La *función de covarianza muestral* será definida como

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+h} - \bar{x})(x_t - \bar{x}),$$

que aunque no cumple con ser un estimador insesgado, sí cumple con la propiedad de paridad $\hat{\gamma}(-h) = \hat{\gamma}(h)$ y según [11] también es una función no negativa definida.

Análoga a la definición de autocorrelación, la *función de autocorrelación muestral* se define como

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

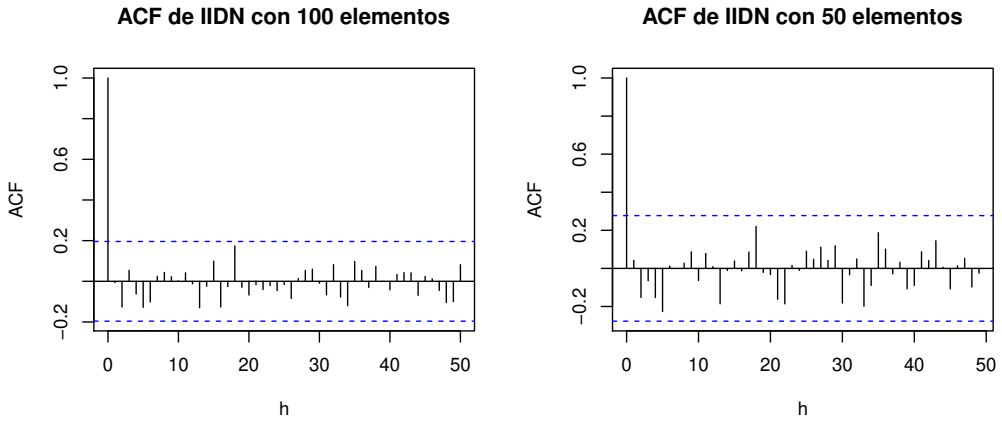

Figura 2.7: Esta figura muestra las gráficas de la función de correlación muestral para una distancia máxima de $h = 50$. En la izquierda vemos la correlación para una serie con 100 valores y en la izquierda las correlaciones de una serie con 50 elementos. Las lineas azules delimitan la región de confianza del 95 %.

Mientras que $\gamma(h) = 0$ para $h \neq 0$ en una serie de tiempo de ruido blanco gaussiano, $\hat{\gamma}(h) \approx 0$ para $h \neq 0$, de hecho, por la Propiedad 1.2 de [1] podemos asegurar que la función de autocovarianza muestral para una serie de tiempo de ruido blanco gaussiano tiene una distribución que se aproxima a una normal con media $\mu = 0$ y varianza $\sigma^2 = 1/n$, por lo tanto, aproximadamente 95 % de de las correlaciones debe caer en el intervalo de confianza $\pm 1{,}96/\sqrt{n}$. Para ilustrar la afirmación anterior, en la Figura 2.7 se muestran dos gráficas de la correlación muestral para la serie de tiempo de ruido blanco gaussiano, la gráfica de la derecha fue realizada tomando una serie con 50 elementos y la de la izquierda se tomó en cuenta una serie con 100 elementos, podemos observar que la región de confianza marcada por las lineas punteadas azules incrementa mientras menos valores tiene la serie de tiempo, es decir, mientras la muestra sea más grande, se tendrá una mejor estimación. Dado que en ambos casos se calculó el valor de la correlación para 50 valores, se esperaría que aproximadamente el 5 % de los valores, es decir 2,5 quedara fuera de la región de confianza, en este caso todos los valores quedaron dentro de su respectiva región de confianza.

Este método será utilizado durante este trabajo para verificar si una serie de tiempo es de ruido blanco, en la sección 1.6 de [11] se pueden consultar algunos métodos más.

El estimador de la función de covarianza cruzada está dado por la siguiente ecuación

$$\hat{\gamma}_{XY}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(y_t - \bar{y}),$$

donde $\hat{\gamma}(-h) = \hat{\gamma}(h)$, para $h = 1, \ldots, n-1$ y la función de correlacieon cruzada muestral

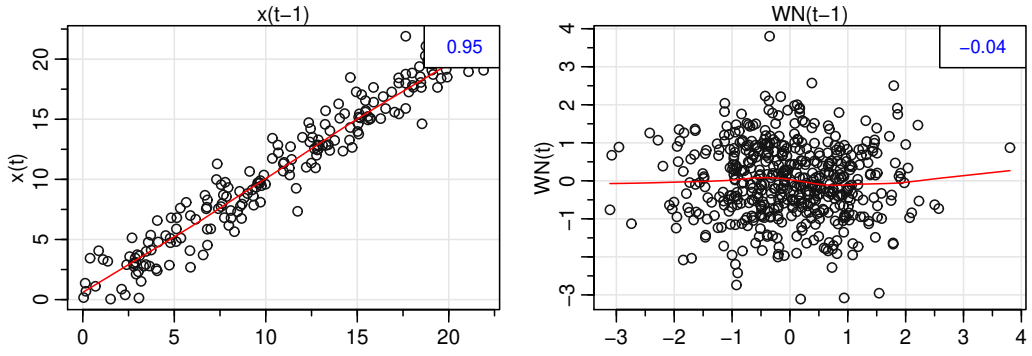

Figura 2.8: Esta figura muestra los diagramas de dispersión de dos series de tiempo, en la esquina superior derecha de cada recuadro se puede ver el coeficiente de correlación. En la izquierda vemos el diagrama de una serie con una dependencia lineal my alta, mientras que a la derecha vemos el diagrama de una serie de ruido blanco.

está dada como

$$\hat{\rho}_{XY}(h) = \frac{\hat{\gamma}_{XY}(h)}{\sqrt{\hat{\gamma}_X(0)\hat{\gamma}_Y(0)}}.$$

Otra forma de verificar si entre los valores que se encuentran a cierta distancia en una serie de tiempo existe dependencia de los datos es revizando los diagramas de dispersión de esta, es decir, graficar la los valores de una serie de tiempo contra los valores de esta misma o de otra pero desplazada, por ejemplo, en la Figura 2.8 se muestra en la parte izquierda el diagrama de dispersión de una serie de tiempo con una dependencia lineal muy alta a un dato de distancia, mientras que en la parte derecha se muestra el diagrama de una serie de ruido blanco a un dato de distancia. Este tipo de diagramas pueden dejar ver otro tipo de dependencia que no muestra la ACF o la CCF, un ejeplo de esto lo podemos ver en el Ejemplo 2.8 de [1].

Por lo regular las series de tiempo que se deben estudiar tienen otras componentes que hacen que la series a tratar no sean estacionarias. En la siguiente sección estudiaremos las componentes de una serie de tiempo.

2.6. Componentes de una serie de tiempo

Existen dos enfoques del análisis de las series de tiempo comúnmente identificados como el enfoque del dominio de tiempo y el enfoque de la dominio de frecuencia.

El *enfoque del dominio del tiempo* está generalmente motivado por la aceptación o consideración de que la correlación entre puntos adyacentes en el tiempo se explica mejor en términos de una dependencia del valor actual sobre los valores pasados de la serie de tiempo, por lo tanto, se enfoca en modelar algún valor futuro de una serie de tiempo como

una función paramétrica de los valores actuales y pasados de esta y otras series de tiempo.

El *enfoque del dominio de la frecuencia* supone que las características primarias de interés en los análisis de series de tiempo se refieren a las variaciones sinusoidales, periódicas o sistemáticas encontradas naturalmente en la mayoría de los datos. Estas variaciones periódicas son a menudo causadas por fenómenos biológicos, físicos o ambientales de interés.

En base a estos dos enfoques es que se supone que las series de tiempo se rigen por el *modelo de la descomposición clásica*, donde la idea es que la serie se puede descomponer en tres componentes o series, las cuales después se combinan aditivamente como

$$X_t = m_t + s_t + Y_t. \tag{2.11}$$

Existen otros modelos en los que se pueden combinar de forma distinta las componente, por ejemplo el modelo multiplicativo que está dado por

$$X_t = m_t \cdot s_t \cdot Y_t. \tag{2.12}$$

Nótese que se puede llegar del modelo de la ecuación (2.12) al de la ecuación (2.11) aplicando logaritmos, el modelo multiplicativo puede ser sugerido cuando la varianza de los datos muestra un notorio crecimiento conforme el tiempo incrementa.

Cada componente representa un tipo de variación en cada momento $t \in \{1, 2, ..., n\}$ de la serie de tiempo. m_t es conocida como componente tendencial, s_t es la componente estacional y Y_t es conocida como variación irregular.

La componente *tendecial* o también conocida como *tendencia* m_t de una serie, describe la evolución lenta y a largo plazo del nivel medio de la serie de tiempo, esta se considera la consecuencia de fuerzas persistentes que afectan el crecimiento o la reducción de la serie.

La componente *cíclica* o *estacional* s_t representa las fluctuaciones de una serie de tiempo como una función no aleatoria periódica. Un patrón cíclico existe cuando una serie está influenciada por factores estacionales (por ejemplo, el trimestre del año, el mes o el día de la semana). En general las fluctuaciones suelen modelarse como sumas se senos y coseno.

La variación o componente *irregular* Y_t se refiere a la variabilidad en el comportamiento de la serie que se debe a pequeñas causas impredecible. La componente irregular se propone que sea modelada por un serie estacionaria, pues estas son estocásticas, pueden modelar más series que solo la de ruido blanco y además existen métodos para predecirlas.

Para ilustrar mejor el modelo de la descomposición clásica tomemos la Figura 2.9 en la que observamos la descomposición de la serie de tiempo del precio de la gasolina en New York, registrado cada semana del 2000 al 2010. Estas componentes fueron calculadas con la función *decompose()* de R, donde dicha función supone que la componente cíclica está implícita en la componente tendencial.

En la siguiente sección veremos varios modelos con los que se puede aproximar la tendencia y la componente estacional de una serie de tiempo.

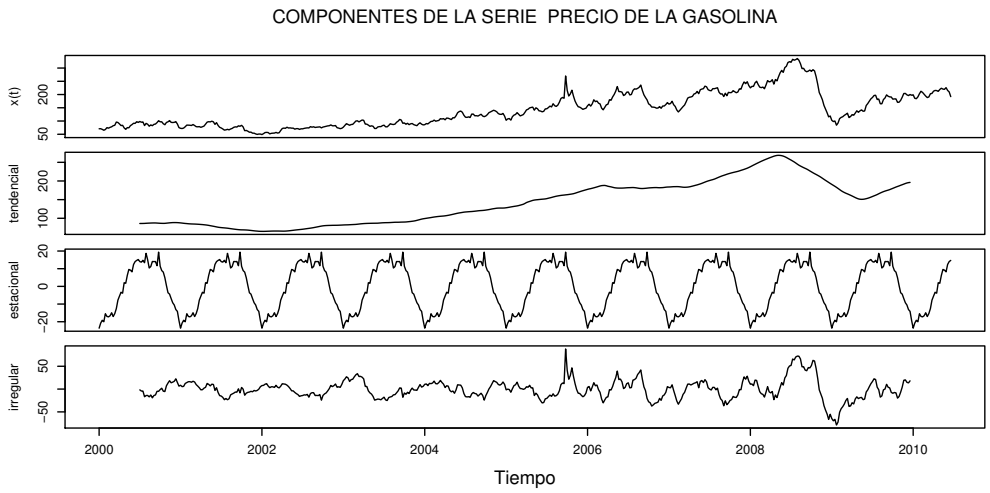

Figura 2.9: En la parte superior vemos la serie de tiempo del precio del gas en New York del 2000 al 2010, las gráficas que se encuentran bajo esta son las series de su tendencia, su variación estacional y su variación irregular respectivamente

2.7. Estimación de componentes

Bajo la suposición de que una serie de tiempo cumple con el modelo de descomposición clásica se tiene la tarea de obtener las componentes tendencial y estacional. Una vez teniendo cada componente se tienen la labor de encontrar un buen modelo probabilístico que describa a la serie irregular y en conjunto con las componentes tendencial y cíclica, lograr buenas predicciones.

Existen dos formas o métodos para obtener una estimación de la componente irregular. La primera forma es estimando las componentes m_t y s_t, sustrayendolas de la serie original X_t, es decir $\hat{Y}_t = X_t - \hat{m}_t - \hat{s}_t$, a $\hat{Y}_t$ se le llama *residuales*. La segunda forma es eliminando directamente las componentes tendencial y estacional de la serie $\{X_t\}$ usando un método el de diferencias, el cual será explicará más adelante.

Existen dos casos que se deben de considerar al remover componentes. Uno es cuando la serie no tiene componte estacional y el otro cuando sí. A continuación se describirán ambos casos.

Caso 1: Supongamos que el proceso $\{X_t\}$ que se está analizando no tiene componente cíclica s_t, es decir, que satisface el modelo

$$X_t = m_t + Y_t, \quad t = 1, 2, \ldots, n, \tag{2.13}$$

donde $E(Y_t) = 0$.

Metodo T1. *Estimación de la tendencia.* En una serie de tiempo que cumple con el modelo

de la ecuación (2.13) existen diferentes métodos para estimar la tendencia, a continuación mencionaremos algunos.

(a) *Suavizamiento.* El suavizamiento de series de tiempo es un método no paramétrico que se utiliza regularmente para obtener tendendencias y en algunos casos periodos, ya que este método elimina las fluctuaciones de las series de tiempo. Existen diferentes formas de suavizar una serie de tiempo, una de las más usadas es mediante *filtros lineales*, es decir, convirtiendo la serie de tiempo X_t a $\hat{m}_t$ mediante el filtro lineal

$$\hat{m}_t = \sum_{j=-q}^{q} w_j X_{t-j}. \tag{2.14}$$

Un filtro lineal es un *filtro de promedios móviles* si $\sum_{j-m}^{m} w_j = 1$, por ejemplo, en la ecuación (2.1) se definió un filtro de promedios móviles con $w_j = 1/(2q+1)$ y $q = 1$.

Si la serie de tiempo $\{X_t\}$ cumple con el modelo de la ecuación (2.13) se justifica el uso de filtros de medias móviles para estimar tendencias, pues

$$\sum_{j=-q}^{q} w_j X_{t-j} = \sum_{j=-q}^{q} w_j m_{t-j} + \sum_{j=-q}^{q} w_j Y_{t-j} \sim m_t,$$

ya que $\sum_{j=-q}^{q} w_j Y_{t-j} \sim 0$ pues $E(Y_t) = 0$ y $\sum_{j=-q}^{q} w_j m_{t-j} \sim m_t$, suponiendo que m_t se comporta linealmente en el intervalo $[t-q, t+q]$. Este método puede ser ineficiente si se escoge q demasiado grande o demasiado chica.

(b) *Ajuste polinomial.* Este método propone que la tendencia sea modelada por un polinomio $m_t = \sum_{i=0}^{k} \beta_i t^i$ y para ajustar el polinomio a la serie de tiempo X_t se use el método de mínimos cuadrados visto en la sección 1.9 con el cual se estiman los parámetros $(\beta_0, \beta_1, \ldots, \beta_k)$. Un ejemplo de Ajuste polinomial se muestra en la Figura 2.10, donde se puede ver la serie de tiempo de las ganancias cuatrimenstrales de la empresa Johnson & Johnson en color negro y en azul el ajuste de un polinomio de segundo grado.

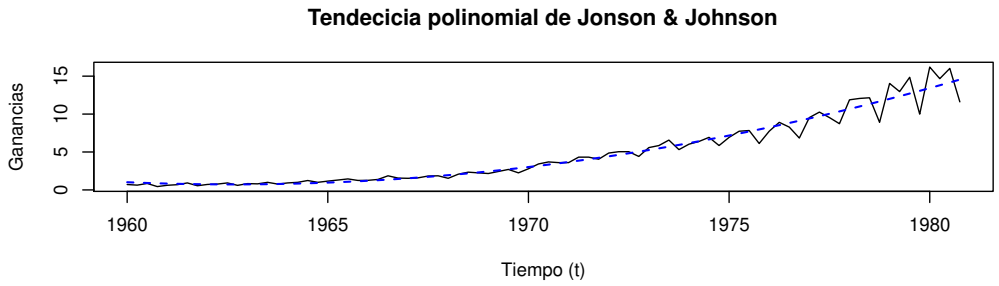

Figura 2.10: En esta figura se muestra en color negro la gráfica de la serie de tiempo de las ganancias cuatrimenstrales de la empresa Johnson & Johnson del año 1960 hata 1980 y en color azul, la tendencia de la serie de tiempo ajustando un polinomio de segundo grado.

Metodo T2. *Eliminar la tendencia por diferencias.* Para hacer uso de este método se definirá el *operador diferencia* ∇ como

$$\nabla(X_t) = X_t - X_{t-1} = (1 - \mathbf{B})X_t,$$

donde $\mathbf{B}$ se le conoce como el *operador Backshift* que se define como $\mathbf{B}X_t = X_{t-1}$. De la misma manera se definen $\mathbf{B}^i(X_t) = X_{t-i}$ y $\nabla^i(X_t) = \nabla^{i-1}(\nabla(X_t))$, para $i \geq 1$ y $\nabla^0(X_t) = X_t$. *Los polinomios de estos operadores pueden ser manipulados igual que los polinomios de números reales* según [11].

Teorema 2.7.1. *Sea $m_t^k = c_0 + c_1 t + \ldots + c_k t^k$ un polinomio de grado k, entonces $\nabla^k(m_t^k)$ es una constante.*

Demostración. Se procederá a demostrar por inducción.

Vemos que para $m_t^1 = c_0 + c_1 t$ un polinomio de grado $k = 1$, se tiene que $\nabla(m_t^1) = m_t^1 - m_{t-1}^1 = (c_0 + c_1(t) - (c_0 + c_1(t-1) = c_1$, por lo tanto se cumple la base de inducción.

Como hipótesis de inducción supongamos que para $m_t^k = c_0 + c_1 t + \ldots + c_k t^k$, un polinomio de grado k, se cumple que $\nabla^k(m_t^k)$ es una constante. Por demostrar que $\nabla^{k+1}(m_t^{k+1}) = \nabla^{k+1}(c_0 + c_1 t + \ldots + c_k t^k + c_{k+1} t^{k+1})$ es una constante.

Veamos que $\nabla(m_t^{k+1})$ es un polinomio de grado k.

$$\begin{aligned} \nabla(m_t^{k+1}) &= m_t^{k+1} - m_{t-1}^{k+1} && (2.15)\\ &= c_0 + \ldots + c_{k+1} t^{k+1} - (c_0 + \ldots + c_{k+1}(t-1)^{k+1}) && (2.16)\\ &= c_1 + c_2(t^2 - (t-1)^2) + \ldots + c_{k+1}(t^{k+1} - (t-1)^{k+1}) && (2.17) \end{aligned}$$

Notemos que sólo el último sumando de la ecuación (2.17) podría tener términos de orden $k+1$, usando el binomio de Newton tenemos que

$$(t-1)^{k+1} = t^{k+1} - (k+1)t^k + \ldots + (-1)^k(k+1)t + (-1)^{k+1}, \tag{2.18}$$

entonces sustituyendo (2.18) en el último sumando de la ecuación (2.17) tenemos que

$$c_{k+1}(t^{k+1} - (t-1)^{k+1}) = c_{k+1}(t^{k+1} - t^{k+1} + (k+1)t^k + \ldots + (-1)^{k+2}) \tag{2.19}$$

$$= c_{k+1}((k+1)t^k + \ldots + (-1)^{k+2}), \tag{2.20}$$

es un polinomio de grado k, por lo tanto, $\nabla(m_t^{k+1})$ es de grado k y el coeficiente de t^k es $c_{k+1}(k+1)$ pues al igual que en (2.20), el término de orden mayor de cada sumando de (2.17) se elimina.

Ahora, dado que $\nabla^{k+1}(m_t^{k+1}) = \nabla^k(\nabla(m_t^{k+1}))$ y se ha mostrado que $\nabla(m_t^{k+1})$ es de orden k, entonces por hipótesis de inducción, $\nabla^{k+1}(m_t^{k+1})$ es una constante. □

Supongamos que se tiene una serie de tiempo $X_t = m_t^k + Y_t$ donde la tendencia de X_t está dada por el polinomio m_t^k de grado k y la serie Y_t estacionaria con media cero, entonces

$$\nabla^k(X_k) = c_k k! + \nabla^k(Y_t),$$

es un proceso estacionario con media $c_k k!$. Este método sugiere la posibilidad de que dada una serie de tiempo $\{x_t\}$ al aplicar repetidas veces el operador ∇ el resultado se pueda modelar con un proceso estacionario.

Caso 2: Ahora supongamos que el proceso $\{X_t\}$ se rige por el modelo de descomposición clásica, es decir,

$$X_t = m_t + s_t + Y_t, \quad t = 1, 2, \ldots, n, \tag{2.21}$$

donde $E(Y_t) = 0$ y s_t tiene periodo d o bien, que $s_t = s_{t+d}$ y $\sum_{j=1}^{d} s_t = 0$.

Método S1 *Estimación de periodos.* Para el caso en el que la serie temporal cumple con el modelo de la ecuación (2.21) explicaremos dos métodos para estimar la componente estacional.

(a) *Suavizamiento.* Al igual que en el Método T1 vamos suavizar la serie $\{X_t\}$ usando promedios móviles. Este método es usado por algunos algoritmos de descomposición. Dado que la serie cumple con el modelo de la ecuación (2.21), entonces $\sum_{j=1}^{d} s_t = 0$, por lo que si se elige el filtro de promedios móviles

$$\hat{m}_t = (0{,}5x_{t-q} + x_{t-q+1} \ldots + x_{t+q-1} + 0{,}5x_{t+q})/d, \quad q < t < n-q, \tag{2.22}$$

cuando el periodo $d = 2q$ es par o los promédios móviles como en (2.14) con $w_j = 1/(2q+1)$ si el periodo d es impar, entonces m_t será una series sin periodos.

La serie resultante $X_t - \hat{m}_t$ es entonces una serie que contiene la componente estacional. Para estimar la componente periódica se van a calcular los promedios w_k de las desviaciones $\{X_{k+jd} - \hat{m}_{k+jd}, q < k + jd < n - q\}$. Dado que la suma de estos promedios no necesariamente es cero, entonces la estimación de la componente periídica s_k estará dada como

$$\hat{s}_k = w_k - d^{-1} \sum_{j=1}^{d} w_j, \quad k = 1, 2, \ldots, d. \tag{2.23}$$

De esta forma se cumple que $\hat{s}_t = \hat{s}_{t+k}$ y $\sum_{i=1}^{d} \hat{s}_i = 0$.

Una vez estimada la componente cíclica podemos estimar la componente tendencial con alguno de los métodos anteriores sobre la serie sin ciclos $x_t - \hat{s}_t$ y finalmente la estimación de la componente irregular estaría dada por

$$\hat{Y}_t = x_t - \hat{m}_t - \hat{s}_t, \quad t = 1, 2, \ldots, n.$$

(b) *Regresión harmónica.* Cuando una serie de tiempo presenta un comportamiento periódico pero sin tendencia, una opción para modelar este comportamiento es usando la regresión harmónica

$$s_t = a_0 + \sum_{j=1}^{k} \Big(a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t) \Big), \tag{2.24}$$

donde $a_1, \ldots, a_k$ y $b_1, \ldots, b_k$ son parámetros desconocidos y $\lambda_1, \ldots, \lambda_k$ son frecuencias fijas. Por ejemplo con el modelo de la ecuación (2.5) se generaron tres series de tiempo que están dadas en la Figura 2.6, este modelo se puede expresar como regresión harmónica de la siguiente forma

$$A \cos(2\pi\omega t + \phi) = \beta_1 \cos(2\pi\omega t) + \beta_2 \sin(2\pi\omega t), \tag{2.25}$$

donde $A = 2$, $\omega = 1/50$, $\phi = 0{,}6\pi$, $\beta_1 = A\cos(\phi)$ y $\beta_2 = -A\sin(\phi)$. Usanndo el conjunto de datos generados con $\sigma^2 = 9$ y asumiendo que $\omega = 1/50$ se hizo el ajuste mediante mínimos cuadrados el cual se muestra en la Figura 2.11.

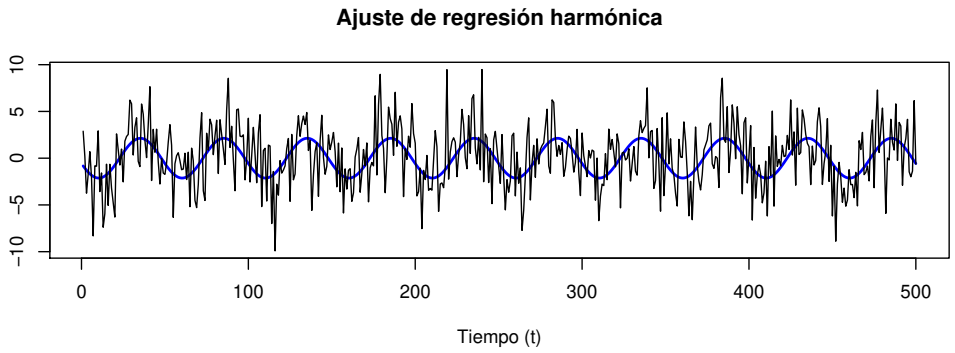

Figura 2.11: En esta gráfica se muestra la gráfica de la serie de tiempo generada con el modelo de la ecuación (2.5) con varianza $\sigma^2 = 9$ en negro y el ajuste usando regresión la harmónica dada en (2.25) en color azul.

Método S2 *Eliminación de periodos con diferencias.* Para este método vamos a definir el operador $\nabla_d(X_t) = X_t - X_{t-d} = (1 - \mathbf{B}^d)X_t$. Si $\{X_t\}$ cumple con el modelo de la ecuación (2.21) entonces

$$\nabla_d(X_t) = m_t - m_{t-d} + Y_t - Y_{t-d},$$

por lo tanto ahora tenemos una serie sin ciclos, con tendencia $m_t - m_{t-d}$ y componente estacionaria $Y_t - Y_{t-d}$, la tendencia en este caso puede ser eliminada con ayuda del operador ∇ como se mencionó antes.

A parte de los métodos mencionados en este capítulo, existen gran variedad de métodos para la obtención de componentes de series de tiempo que no serán mencionados en este trabajo pero que si son del interés del lector pueden ser consultados en [11] y [1].

En el siguiente capítulo estudiaremos más a fondo las propiedades de las series estacionarias y cómo obtener buenas predicciones de éstas.

Capítulo 3

Modelos estacionarios y no estacionarios

En el capítulo anterior se mencionaron ya algunos modelos estacionarios como fue el modelo de *ruido blanco*, el modelo *autoregresivo* de la ecuación (2.2) y el modelo de *promedios móviles* de la ecuación (2.1), también se mencionó un modelo que resultó ser no estacionario, este fue el modelo de *caminata aleatoria con desvío* de la ecuació (2.3). Todos estos modelos son casos particulares de una gama de modelos conocidos como los modelos ARIMA(p, d, q) que serán estudiados en este capítulo.

Los modelos ARIMA son un grupo de modelos lineales de mucha importancia para serries de tiempo estacionarias y no estacionarias. En este capítulo estudiaremos bajo qué condiciones estos modelan series estacionarias, también mencionaremos algunos predictores para este tipo de modelos.

3.1. Procesos Lineales

Definición 3.1.1. *Una serie de tiempo $\{X_t\}$ es un proceso lineal si tiene la representación*

$$X_t = \sum_{i=-\infty}^{\infty} \psi_i W_{t-i}, \tag{3.1}$$

para toda t, donde $\{W_t\} \sim WN(0, \sigma^2)$ y $\{\psi_i\}$ es una secuencia de constantes tal que $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$.

Proposición 3.1.1. *Sea $\{Y_t\}$ una serie estacionaria con media cero y función de covarianza γ_Y, si $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$, entonces la serie de tiempo*

$$X_t = \sum_{i=-\infty}^{\infty} \psi_i Y_{t-i} \tag{3.2}$$

es una serie estacionaria con $E(X_t) = 0$ para toda t y función de autocovarianza

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h-k+j). \tag{3.3}$$

En el caso especial cuando $\{X_t\}$ *es un proceso lineal entonces*

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \sigma^2. \tag{3.4}$$

La Proposición 3.1.1 la podemos encontrar en [11] como Proposición 2.2.1 con su demostración.

3.2. Proceso AR(p)

Los modelos regresivos se basan en la idea de que el valor actual X_t de una serie de tiempo $\{X_t\}$ está en función de los p valores que le preceden $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$, es decir, p es el número de valores hacia atrás que se ocupan para predecir el valor de X_t. El modelo AR(p) es un caso particular de un modelo regregresivo, donde la función que describe a X_t en términos de sus p valores anteriores es lineal, como se describe en la siguiente definición.

Definición 3.2.1. *Un* ***modelo autoregresivo de orden*** p ***o proceso AR(***p***)*** *es de la forma:*

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + W_t, \tag{3.5}$$

donde $\{X_t\}$ *es una serie de tiempo estacionaria,* $W_t \sim WN(0, \sigma^2)$ *y* $\phi_1, \phi_2, \ldots, \phi_p$ *son constantes y* $\phi_q \neq 0$*. Si* $\mu = E(X_t) \neq 0$ *sustituimo* X_t *por* $X_t - \mu$ *en la ecuación (3.5) obteniendo*

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \ldots + \phi_p(X_{t-p} - \mu) + W_t,$$

o bien

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + W_t,$$

donde $\alpha = \mu(1 - \phi_1 - \phi_2 - \ldots - \phi_p)$.

También puede ser descrito este modelo con el operador backshift como

$$(1 + \phi_1 \mathbf{B} + \ldots + \phi_p \mathbf{B}^p) x_t = w_t$$

o $\phi(B) = w_t$, donde $\phi(\mathbf{B})$ denota el *operador autorregresivo* dado por

$$\phi(\mathbf{B}) = 1 + \phi_1 \mathbf{B} + \ldots + \phi_p \mathbf{B}^p.$$

Ejemplo 3.2.1. *Comenzaremos nuestra investigación con el modelo AR(1) que está dado por $x_t = \phi x_{t-1} + w_t$ e iterando hacia atrás k veces tenemos que*

$$\begin{aligned} x_t =& \phi x_{t-1} + w_t = \phi(\phi x_{t-2} + w_{t-1}) + w_t \\ &= \phi^2 x_{t-2} + \phi w_{t-1} + w_t \\ &\vdots \\ &= \phi^k x_{t-k} + \sum_{i=0}^{k-1} \phi^i w_{t-i}. \end{aligned} \tag{3.6}$$

Cuando k tiende a infinito la ecuación (3.6) converge solo si $|\phi| < 1$, por lo que el modelo AR(1) queda representado por el proceso lineal

$$x_t = \sum_{i=0}^{\infty} \phi^i w_{t-i},$$

el cual es un proceso estacionario con media $E(x_t) = 0$, $\gamma(h) = \sigma^2\phi^h/(1-\phi^2)$, $\rho(h) = \phi^h$ que satisface el modelo AR(1) y se puede demostrar que es único, los detalles se pueden revisar en el Ejercicio 3.1 de [1]. Nótese que la función de covarianza cumple la ecuación recursiva

$$\rho(h) = \phi\rho(h-1),$$

a este tipo de ecuaciones se les conoce como ecuaciones homogeneas de diferencias de grado 1, las soluciones de este tipo de de ecuaciones son de utilidad para calcular las correlaciones con propiedades recursivas.

En el Ejemplo 2.3.3 se puede ver que el modelo de caminata aleatoria dado por la ecuación $x_t = x_{t-1} + w_t$ no es un proceso estacionario. Sin embargo existen procesos estacionarios que satisfacen el modelo AR(1) con $|\phi| > 1$, a este tipo de procesos se les conoce como *explosivos* pues $\phi^j \to \infty$ cuando $j \to \infty$ y $\sum_{i=0}^{\infty} \phi^i w_{t-i}$ no converge en media. De cualquier forma se puede obtener una serie estacionaria escribiendo el modelo AR(1) como $x_{t+1} = \phi x_t + w_{t+1}$, de esta forma $x_t = \phi^{-1}x_{t+1} - w_{t+1}$, por lo tanto

$$x_t = \sum_{i=0}^{\infty} \phi^{-i} w_{t+i},$$

es un proceso estacionario con $E(x_t) = 0$, $\gamma(h) = \sigma^2\phi^{-2}\phi^{-h}/(1-\phi^{-2})$ y $\rho(h) = \phi^{-h}$. Los procesos AR(1) con $|\phi| > 1$ aunque son estacionarios no resultan de mucha utilidad para hacer predicciones pues dependen de los valores futuros. A los procesos que no dependen del futuro, en este caso los AR(1) con $|\phi| < 1$ se les conoce como *causales*. Un modelo AR(1) causal digamos $x_t = \phi x_{t-1} + w_t$ y uno explosivo $y_t = \phi^{-1}y_{t-1} + v_t$ tienen las mismas propiedades estocásticas (ver Ejemplo 3.4 de [1]) sin embargo el proceso de nuestro interés

es el causal.

En la sección 3.4 se hablará de la forma de generalizar las propiedades del modelo AR(1) para modelos AR(p).

3.3. Procesos MA(q)

Definición 3.3.1. *Un proceso estacionario $\{X_t\}$ es q-**correlacionado** si $\gamma(h) = 0$ para toda h tal que $|h| > q$ pero $\gamma(h) \neq 0$ si $|h| = q$.*

Definición 3.3.2. *$\{X_t\}$ es un **proceso de promedios móviles de orden q** o **proceso MA(q)** si*

$$X_t = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_q W_{t-q} \tag{3.7}$$

donde $\{W_t\} \sim WN(0, \sigma^2)$ y $\theta_1, \theta_2, \ldots, \theta_q$ son constantes.

Este modelo también puede ser descrito como $x_t = \theta(\mathbf{B})w_t$, donde $\theta(\mathbf{B})$ define el *operador de promedios móviles* como

$$\theta(\mathbf{B}) = 1 + \theta_1 \mathbf{B} + \ldots + \theta_q \mathbf{B}^q.$$

Sea $\theta_0 = 1$ y $\{X_t\}$ un proceso MA(q) dado como $X_t = \sum_{i=0}^{q} \theta_i W_{t-i}$, entonces el valor esperado de X_t es cero, como lo muestra la siguiente ecuación:

$$E(X_t) = E\left(\sum_{i=1}^{q} \theta_i W_{t-i}\right) = \sum_{i=1}^{q} \theta_i E(W_{t-i}) = 0.$$

Además la covarianza de cualesquiera dos valores que están a distancia h se puede calcular como a continuación:

$$\gamma(h) = Cov(X_t, X_{t+h}) = Cov\left(\sum_{i=0}^{q} \theta_i W_{t-i}, \sum_{j=0}^{q} \theta_j W_{t+h-j}\right) \tag{3.8}$$

$$= \sum_{i=0}^{q} \sum_{j=0}^{q} \theta_i \theta_j Cov(W_{t-i}, W_{t+h-j}) \tag{3.9}$$

$$= \begin{cases} 0 & \text{si } |h| > q \\ \sum_{i=0}^{q-h} \theta_i \theta_{i+h} \sigma^2 & \text{si } |h| \leq q. \end{cases} \tag{3.10}$$

Para pasar de la ecuación (3.8) usamos la Propiedad A.10.3 y para pasar de la ecuación (3.9) a (3.10) usamos el hecho de que como $\{W_t\} \sim WN(0, \sigma^2)$ entonce

$$Cov(W_{t-i}, W_{t+h-j}) = \begin{cases} 0 & \text{si } |h| > q \\ \sum_{i=0}^{q-h} \theta_i \theta_{i+h} \sigma^2 & \text{si } |h| \leq q. \end{cases},$$

dado que $i, j = 0, 1, \ldots, q$, para que se cumpla que $h = j - i$ entonces $|h| \leq q$. La función de corrrelación para los modelos de promedios móviles está dada entonces por

$$\rho(h) = \begin{cases} 0 & \text{si } |h| > q \\ \frac{\sum_{i=0}^{q-h} \theta_i \theta_{i+h}}{\theta_0^2 + \theta_1^2 + \ldots + \theta_q^2} & \text{si } |h| \leq q. \end{cases},$$

A diferencia del modelo AR(p), podemos afirmar que si $\{X_t\}$ es un proceso MA(q) entonces $\{X_t\}$ es un proceso estacionario independientemente del valor de sus parámetos, pues $E(X_t)$ es constante y $\gamma(h)$ no depende del tiempo sino del valor de h.

La importancia de los modelos MA(q) recaé en que de acuerdo con la Definición 3.3.1 un proceso MA(q) es q-correlacionado y según la Proposición 3.3.1, también el converso es cierto. La demostración de la Proposición 3.3.1 se encuentra en la Sección 3.2 de [11].

Proposición 3.3.1. *Si $\{X_t\}$ es una serie de tiempo estacionaria q-correlacionada con media cero, entonces puede ser representada como un proceso MA(q) definido por la ecuación (3.7).*

Al igual que en los modelos AR(p), no hay unicidad en los modelos MA(q), es decir, puede haber varios con las mismas propiedades estocásticas, veamos el siguiente ejemplo.

Ejemplo 3.3.1. *Sean los modelos MA(1) $x_t = w_t + 5w_{t-1}$ con $\{w_t\} \sim N(0, 1)$ y $y_t = v_t + 1/5v_{t-1}$ con $\{v_t\} \sim N(0, 25)$, podemos verificar que ambos tienen media cero y la misma función de covarianza*

$$\rho(h) = \begin{cases} 26 & si\ h = 0 \\ 5 & si\ h = 1 \\ 0 & si\ h > 1 \end{cases}.$$

En estos casos nos interesan los modelos que sean *invertibles*, o bien que puedan ser escritos con una representación infinita AR. Por ejemplo, un modelo MA(1) $x_t = w_t + \theta w_{t-1}$ puede ser escrito como $w_t = -\theta w_{t-1} + x_t$, entonces si $|\theta| < 1$ se tiene que $w_t = \sum_{i=0}^{\infty} (-\theta)^i x_{t-i}$, usando el Ejemplo 3.3.1 el modelo elegido sería el que tiene $\theta = 1/5$ y $\sigma^2 = 25$, pues este sería invertible.

La propiedad de invertibilidad será definida en la sección siguiente y está dependerá totalmente del operador $\theta(\mathbf{B})$.

3.4. Procesos ARMA(p, q)

Ahora describiremos un modelo más general, este modelo es una combinación de los modelos AR(p) y MA(q). El modelo ARMA(p, q) es un modelo para series de tiempo estacionarias.

Definición 3.4.1. *Una serie de tiempo $\{X_t\}$ cumple con el modelo **autoregresivo de promedios móviles** o es **ARMA(p, q)** si es estacionaria y*

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_q W_{t-q}, \qquad (3.11)$$

*donde $\phi_p \neq 0$ $\theta_q \neq 0$ y $\sigma_W^2 > 0$, los parámetros p y q son el **orden autorregresivo** y el **orden de promedios móviles** respectivamente. Si la media μ de X_t es distinta de cero, hacemos $\alpha = (1 - \phi_1 - \phi_2, \ldots, \phi_p)$ y escribimos el modelo ARMA(p, q) como*

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_q W_{t-q}$$

donde $W_t \sim WN(0, \sigma_W^2)$. Notemos que si se hace $q = 0$ y $p \neq 0$ tenemos e modelo AR(p) mientras que si se hace $p = 0$ y $q \neq 0$ tenemos el modelo MA(q).

El modelo también puede ser descrito por los operadores $\phi(\mathbf{B})$ y $\theta(\mathbf{B})$ como

$$\phi(\mathbf{B})x_t = \theta(\mathbf{B})w_t.$$

Este modelo en ocaciones puede ser complicado multiplicando ambos lados de la ecuación aterior por un operador, por ejemplo,

$$\eta(\mathbf{B})\phi(\mathbf{B})x_t = \eta(\mathbf{B})\theta(\mathbf{B})w_t.$$

En estos casos la dinámica del modelo no cambia, es decir, sigue cumpliendo con el modelo ARMA(p, q), veamos un ejemplo a continuación.

Ejemplo 3.4.1. *Considere el proceso de ruido blanco $x_t = w_t$ donde $\{w_t\} \sim WN(0, \sigma^2)$, si multiplicamos ambos lado por $\eta(\boldsymbol{B}) = 1 - ,5\boldsymbol{B}$, entonces el modelo se vuelve $(1 - ,5\boldsymbol{B})x_t = w_t(1 - ,5\boldsymbol{B})$ o bien*

$$x_t = ,5x_{t-1} - ,5w_{t-1} + w_t,$$

que aparentemente es un proceso ARMA(1,1), pero enrealidas sigue siendo una proceso de ruido blanco.

Definición 3.4.2. *Los polinomios **AR(p)** y **MA(q)** son definidos como*

$$\phi(z) = 1 + \phi_1 z + \ldots + \phi_p z^p, \quad \phi_p \neq 0,$$

$$\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q, \quad \theta_q \neq 0,$$

respectivamente, para $z \in \mathbb{C}$.

Al fenómeno descrito en el Ejemplo 3.4.1 se le llama redundancia de parámetros y para que no ocurra, los parámetros de los polinomios AR(p) y MA(q) deben ser distintos. A

continuación definiremos formalmente los conceptos de causalidad e invertibilidad para modelos ARMA(p, q).

Definición 3.4.3. *Un modelo ARMA(p, q) es **causal** si la serie de tiempo $\{x_t\}$ puede ser escrita como un proceso lineal de un lado, es decir,*

$$x_t = \sum_{i=0}^{\infty} \psi_i w_{t-j} = \psi(\boldsymbol{B}) w_t, \tag{3.12}$$

donde $\psi(\boldsymbol{B}) = \sum_{i=0}^{\infty} \psi_j \boldsymbol{B}^j$, $\sum_{i=0}^{\infty} |\psi_j| < \infty$ y $\psi_0 = 1$.

Propiedad 3.4.1. *Un proceso ARMA(p, q) es causal si y sólo si $\phi(z) \neq 0$ para $|z| \leq 1$. Los coeficientes del proceso lineal de la ecuación (3.12) pueden ser determinados resolviendo*

$$\sum_{i=0}^{\infty} \psi_i z^i = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1. \tag{3.13}$$

La ecuación (3.13) implica que la sucesión de coeficientes $\{\psi_i\}$ puede ser calculada numéricamente haciendo

$$\psi_j = \sum_{k=1}^{p} \phi_k \psi_{j-k} + \theta_j, \quad j = 0, 1, \ldots \tag{3.14}$$

donde $\theta_0 := 1$, $\theta_j := 0$ para $j > q$ y $\psi_j :=$ para $j < 0$.

Definición 3.4.4. *Se dice que un modelo ARMA(p, q) es **invertible** si la serie de tiempo $\{X_t\}$ puede ser escrita como*

$$\pi(\boldsymbol{B}) x_t = \sum_{i=0}^{\infty} \pi_i x_{t-i} = w_t, \tag{3.15}$$

donde $\pi(\boldsymbol{B}) = \sum_{i=0}^{\infty} \pi_j \boldsymbol{B}^j$, $\sum_{i=0}^{\infty} |\pi_j| < \infty$ y $\pi_0 = 1$.

Propiedad 3.4.2. *Un proceso ARMA(p, q) es si y solo si $\theta(z) \neq 0$ para $|z| \leq 1$. Los coeficientes π_i de la ecuación (3.15) pueden ser determinados resolviendo*

$$\sum_{i=0}^{\infty} \pi_i z^i = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1. \tag{3.16}$$

La ecuación (3.16) implica que la sucesión de coeficiente $\{\pi_i\}$ pueda ser calculados numéricamente como

$$\pi_j = -\sum_{k=1}^{q} \theta_k \pi_{j-k} - \theta_j, \quad j = 0, 1, \ldots,$$

donde $\theta_0 := -1$, $\theta_j := 0$ para $j > p$ y $\pi_j := 0$ para $j < 0$.

La demostración a la Propiedad 3.4.1 es similar a la demostración de la Propiedad 3.4.2 la cual se puede encontrar en la sección B.2 de [1].

Para un modelo ARMA(p,q) $\phi(\mathbf{B})x_t = \theta(\mathbf{B})w_t$ causal, podemos escribir

$$x_t = \sum_{i=0}^{\infty} \psi_i w_{t-i}.$$

Inmediatamente tenemos que $E(x_t) = 0$ y la función de autocovarianza está dada por

$$\begin{aligned}\gamma(h) = Cov(x_{t+h}, x_t) &= Cov\left(\sum_{i=1}^{p} \phi_i x_{t+h-i} + \sum_{i=1}^{q} \theta_i w_{t+h-i}, x_t\right) \\ &= \sum_{i=1}^{p} \phi_i \gamma(h-i) + \sum_{i=1}^{q} \theta_i Cov(w_{t+h-i}, \sum_{j=0}^{\infty} \psi_j w_{t-j}) \\ &= \sum_{i=1}^{p} \phi_i \gamma(h-i) + \sigma_w^2 \sum_{i=h}^{q} \theta_i \psi_{h-i}, \quad h \geq 0.\end{aligned}$$

La función de correlación para un modelo ARMA(p,q) satisface una ecuación recursiva, en la sección 3.2 de [1] se puede encontrar un método para solucionar este tipo de ecuaciones, este mismos método puede usarse para encontrar la solución de manera explícita de las sucesión de coeficientes $\{\psi_j\}$ y $\{\pi_j\}$.

3.5. Función de autocorrelación parcial (PACF)

Para el caso en el que se trabaja con un modelo MA(q) la función de autocorrelación es distinta de cero para $h = \pm q$ y $\rho(h) = 0$ cuando $|h| > q$, esto nos da información sobre el grado de dependencia cuando el modelo es MA, sin embargo para el caso de los modeloas AR y ARMA la función de correlación nos da poca información sobre el grado de dependencia. Es por esto que vale la pena introducir una función que se comporte como la ACF para los modelos MA, esta función es llamada *función de correlación parrcial* (PACF).

La idea de la PACF es calcular la correlación de dos variariables aleatorias, digamos X y Y sin la dependencia lineal de otra variable aleatoria, digamos Z. Para el caso de un proceso estocástico estacionario con media cero $\{x_t\}$ se busca quitar la dependencia lineal que se encuatra entre x_t y x_{t+h}, es decir, quitar

$$\hat{x}_{t+h} = \beta_1 x_{t+h-1} + \beta_2 x_{t+h-2} + \ldots + \beta_{h-1} x_{t+1},$$

o bien, que es lo mismo

$$\hat{x}_t = \beta_1 x_{t+1} + \beta_2 x_{t+2} + \ldots + \beta_{h-1} x_{t+h-1},$$

	AR(p)	MA(q)	ARMA(p, q)
ACF	Las colas decrecen	Se corta después de $h = q$	Las colas decrecen
PACF	Se corta después de $h = p$	Las colas decrecen	Las colas decrecen

Cuadro 3.1: Comportamiento de ACF y PACF para modelos ARMA.

donde los coeficientes $\beta_1, \ldots, \beta_{h-1}$ son calculados minimizando el MSE definido en la ecuación (1.2).

Definición 3.5.1. *La **función parcial de autocorrelación (PACF)** de un proceso estacionario $\{x_t\}$ se denota como ϕ_{hh}, para $h = 1, 2, \ldots$, y se define como*

$$\phi_{11} = \rho(x_t, x_{t+1}) = \rho(1),$$

y

$$\phi_{hh} = \rho(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t), \quad h \geq 2.$$

Aunque aquí no será demostrado la $\phi_{hh} = 0$ para un modelo AR(p) con $h > p$. En el Cuadro 3.1 se muestra el comportamiento de la ACF y la PACF para los modelos ARMA.

3.6. Procesos ARIMA(p, d, q)

En la sección 2.6 vimos como una serie de tiempo puede ser descompuesta en tres series o componentes: la componente tendencial, la estacional y la estacionaria. También se estudiaron algunos métodos para eliminar y estimar tendencia y ciclos, un ejemplo es el método de diferencias. Los modelos mencionados para la estimación de tendencia y ciclos fueron modelos determinísticos. Las componentes también pueden ser no determinísticas. Por ejemplo, se puede proponer un modelo para una serie de tiempo como

$$x_t = \mu_t + y_t,$$

donde $\mu_t = \mu_{t-1} + v_t$ y $\{y_t\}$, $\{v_t\}$ son estacionarias. De este modo $\nabla x_t = v_t + \nabla y_t$ resulta una serie estacionaria. Es por esto que se introduce el modelo ARIMA el cual será descrito a continuación.

Definición 3.6.1. *Se dice que un proceso $\{x_t\}$ es **ARIAM(p, d, q)** si*

$$\nabla^d x_t = (1 - B)^d x_t$$

es ARMA(p, q). En general escribimos el modelo como

$$\phi(B)\nabla^d x_t = \theta(B) w_t.$$

Si $E(\nabla^d x_t) = \mu$*, escribimos el modelo como*

$$\phi(B)\nabla^d x_t = \delta + \theta(B)w_t,$$

donde $\delta = \mu(1 + \phi_1 + \ldots + \phi_p)$.

Aunque este es un modelo para series no estacionarias, en general se busca convertirla a un modelo ARMA(p, q), es decir, retirar la no estacionariedad y volverla estacionaria. Por lo tanto los méodos con los que se estarán trabajando serán los de la sección 2.7.

3.7. Estimación de parámetros en modelos ARMA(p, q)

Suponiendo que se tienen n observaciones $\{x_1, x_2, \ldots, x_n\}$ y los órdenes p y q del modelo ARMA que representa al conjunto de observaciones han sido fijados, es necesario estimar los parámetros $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ y σ_w^1. Algunos de los método que se pueden utilizar para la estimación de estos parámetros son los métodos de mínimos cuadrados ordinarios, método de los momentos y método de máxima verosimilitud que fueros descritos en el Capítulo 1. Los detalles de la obtención de estos estimadores con cada método en particular pueden ser revisados en la sección 3.5 de [1]. Para la estimación de parámetros de nuestros modelos ARMA estaremos usando los métodos de R.

3.8. Predicciones

Definición 3.8.1. *Sea X una variable aleatoria con media* μ *y varianza* σ^2*, una* ***pedicción*** *de X es una estimación del valor de X antes de ser observada.*

En series de tiempo es de interés hacer predicciónes del valor futuro de x_{n+m}, $m = 1, 2, \ldots$ teniendo en cuenta la colección de datos $x_{1:n} = \{x_n, x_{n-1}, \ldots, x_1\}$. Los predictores de interés son aquellos que minimicen el MSE. Por ejemplo, dado $x_{1:n}$, la función de $x_{1:n}$ que minimiza el MSE con x_{n+m} o bien el *predictor de mínimo error cuadrático medio* de x_{n+m} dado $x_{1:n}$ es

$$x_{n+m}^n = E(x_{n+m}|x_{1:n}). \tag{3.17}$$

Consideremos los predictores lineales,

$$x_{n+m}^n = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i, \tag{3.18}$$

donde α_i, $i = 1, \ldots, n$ son coeficientes reales. Los predictores lineales que minimizan el MSE se les conoce como los *mejores predictores lineales (BLP)* por sus siglas en inglés.

En el Teorema B.3 de [1] se muestra que si un proceso es gaussiano entonces el mejor predictor lineal y el predictor de mínimo error cuadrático medio son el mismo y además solo dependen de los momentos de segundo orden.

Para minimizar el MSE de un predictor lineal, es decir, $Q = E\left(x_{n+m} - \sum_{i=0}^{n} \alpha_i x_i\right)^2$, donde $x_0 = 1$, buscamos que $\frac{\partial Q}{\partial \alpha_i} = 0$, lo cual genera las *ecuaciones de predicción*

$$E([x_{n+m} - x_{n+m}^n]x_k) = 0, \quad k = 1, 2, \ldots, n. \tag{3.19}$$

La solución del sistema de ecuaciones (3.19) genera el mejor predictor lineal de x_{n+m} y nos da los coeficientes $\alpha_0, \alpha_1, \ldots, \alpha_n$.

Cuando $\{x_t\}$ es una serie estacionaria, $E(x_t) = \mu$, entonces la ecuación (3.19) cuando $k = 0$ implica que

$$E(x_{n+m}) = E(x_{n+m}^n) = \mu.$$

Tomando la esperanza de (3.18) tenemos que

$$\mu = \alpha_0 + \sum_{i=1}^{n} \alpha_i \mu \quad o \quad \alpha_0 = \mu(1 - \sum_{i=0}^{n} \alpha_i).$$

Entonces el BLP está dado como

$$x_{n+m}^n = \mu + \sum_{i=1}^{n} \alpha_i (x_i - \mu),$$

por lo tanto, si $\mu = 0$ entonces $\alpha_0 = 0$.

Consideremos el *predictor de un paso hacia adelante*, es decir, dado el conjunto $\{x_1, x_2, \ldots, x_n\}$ deseamos predecir el siguiente valor de la serie de tiempo x_{n+1}, entonces el predictor lineal está dado como

$$x_{n+1}^n = \phi_{n1} x_n + \phi_{n2} x_{n-2} + \ldots + \phi_{nn} x_1$$

y debe satisfacer las ecuaciones de predicción dadas en la ecuación (3.19) o bien

$$\sum_{i=1}^{n} \phi_{ni} \gamma(k - i) = \gamma(k), \quad k = 1, \ldots, n$$

que en su forma matricial se ven como

$$\Gamma_n \phi_n = \gamma_n, \tag{3.20}$$

donde $\Gamma_n = \{\gamma(k - i)\}_{j,k=1}^{n}$ es una matriz de $n \times n$, $\phi_n = (\phi_{n1}, \ldots, \phi_{nn})'$ un vector de $n \times 1$ y $\gamma_n = (\gamma(1), \ldots, \gamma(n))'$ un vector de $n \times 1$.

La matriz Γ_n es no negativa definida por las propiedades descritas en el capítulo anterior.

Si Γ_n es no singular entonces podemos escribir

$$\phi_n = \Gamma_n^{-1}\gamma_n \tag{3.21}$$

Para los modelos ARMA, como $\sigma_w^2 > 0$ y $\gamma(h) \to 0$, cuando $h \to \infty$ es suficiente para asegurar que Γ_n es positiva definida y es conveniente escribir la predicción como

$$x_{n+1}^n = \phi_n x,$$

donde $x = (x_1, x_2, \ldots, x_n)$.

La media del error cuadrático de predicción para este modelo es

$$P_{n+1}^n = E(x_{n+1} - x_{n+1}^n)^2 = \gamma(0) - \gamma_n'\Gamma_n^{-1}\gamma_n. \tag{3.22}$$

La ecuación (3.21) no siempre puede ser utilizada cuando n es muy grande, pues como se tiene que calcular una inversa, existe un problema numérico, sin embargo existen algoritmos iterativos con los que se abordan este tipo de problemas.

Para el caso en que se tiene una serie de tiempo $\{x_t\}$ que cumple con el modelo ARMA(p, q) causal e invertible es más fácil calcular el predictor x_{n+m}^n basados en *el pasado infinito*, es decir, es más fácil calcular

$$\tilde{x}_{n+m} = E(x_{n+m}|x_n, x_{n-1}, \ldots x_1, x_0, x_{-1}, \ldots)$$

en lugar de x_{n+m}^n dado por la ecuación (3.17). En general x_{n+m}^m y $\tilde{x}_{n+m}$ no son igual, pero la idea es que $\tilde{x}_{n+m}$ sea una buena aproximación para x_{n+m}^n cuando n es grande.

El predictor $\tilde{x}_{n+m}$ puede ser calculado recursivamente con la ecuación (3.23) comenzando con $m = 1$ y después $m = 2, 3, \ldots$.

$$\tilde{x}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j}, \tag{3.23}$$

Además la media del error de predicción puede ser escrita como

$$P_{n+m}^n = E(x_{n+m} - \tilde{x}_{n+m}) = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2. \tag{3.24}$$

Dado que solo se cuenta con una cantidad finita de muestras $\{x_1, \ldots, x_n\}$, este predictor se trunca, de modo que obtenemos

$$\tilde{x}_{n+m}^n = -\sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j}^n - \sum_{j=m}^{n+m-1} \pi_j x_{n+m-j}, \tag{3.25}$$

el cual es calculado recursivamente, con $m = 1, 2, \ldots$, el error medio de la predicción es

aproximado usando (3.24).

El análisis de series de tiempo es un campo bastante amplio, donde se estudian diferentes modelos y métodos para hacer el análisis de las serie de tiempo en los que se podría indagar y profundizar para poder hacer un mejor estudio de éstas, sin embargo para estudiar modelos no lineales preferimos abarcar los algunos modelos de inteligencia artificial, para de este modo comparar resultados con los modelos dados en este capítulo.

Capítulo 4

Redes neuronales artificiales

Las redes neuronales artificiales (RNAs) proveen una amplia gama de nuevas técnicas para la resolución de problemas en diferentes áreas de la ciencia. Por ejemplo, entre las cosas por las cuales se han dado a conocer es por su alto desempeño en reconocimiento de patrones, análisis de datos, control, optimización, aproximación de funciones, predicciones, combinatoria, entre otros, ver [12].

El avance del cómputo científico ha ayudado en gran medida al desarrollo y desempeño de esta herramienta. Algunas de las características que presentan las redes neuronales y que hacen de ellas una buena herramienta son: paralelismo masivo, habilidad de aprendizaje, adaptabilidad, habilidad de generalización, bajo consumo de energía y tolerancia a fallos, ver [12].

Existen dos grandes grupos de RNAs, el grupo de RNAs de alimentación hacia adelante y el grupo de RNAs recurrentes, revisar [5] y [12]. Dentro del grupo de redes neuronales de alimentación hacia adelante encontramos un tipo de RNAs que son conocidas como *perceptrón multicapa*, este tipo de redes neuronales son las que se estudiarán durante este capítulo.

Las RNAs *perceptrón multicapa* utilizan un algoritmo de aprendizaje supervisado, el cual se basa en realizar un proceso conocido como error-corrección. Este proceso se hace mediante la optimización de una función que mide el error empírico que generan las aproximaciones del modelo de la RNA. El objetivo del algoritmo de aprendizaje es minimizar una función que mide el error empírico. La minimización de la función de error se realiza mediante algún método de optimización numérica como *descenso de gradiente*. El algoritmo del proceso de aprendizaje realizado para este tipo de redes neuronales se le conoce como *back propagation*.

Las RNAs son un modelo matemático basado en la forma en la que aprende el cerebro humano. A continuación se describirá la conexión y semejanza que tiene el modelo de las RNAs con las redes neuronales biológicas (RNBs), además se dará una descripción del modelo *perceptrón multicapa*, y el algoritmo de aprendizaje o entrenamiento *back propagation*, también se explicará el criterio de paro que estaremos utilizando en este trabajo

para entrenar una RNA.

4.1. Redes neuronal biológicas

Las RNAs fueron inspiradas en las mismas RNBs. Las RNAs intentan simular la forma en la que el cerebro aprende y procesa la información. Se dice que el cerebro cuenta con un aproximado de 10^{11} neuronas, las cuales están interconectadas y por medio de impulsos eléctricos se efectúa la transmisión de información.

Una neurona biológica está conformada principalmente por las dendritas, el núcleo o soma, el axón y las sinapsis, ver Figura 4.1. Las dendritas es la parte de la neurona que le permite recibir información de otras neuronas para después ser procesada por la neurona. El núcleo o soma de la neurona hace una recopilación de todos los impulsos recibidos de otras neuronas por las dendritas. Una vez hecha esta suma, la información es procesada a través del axón para finalmente ser transmitida a otras neuronas por medio de las sinapsis.

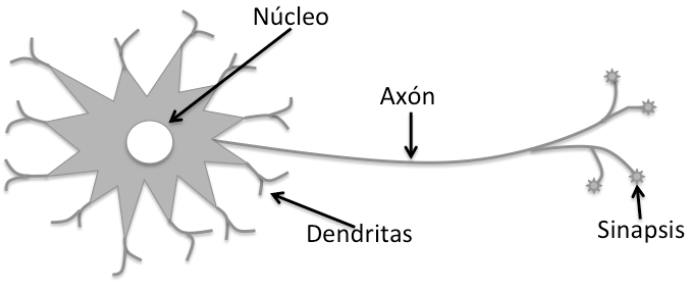

Figura 4.1: En esta figura se muestra la representación de una neurona biológica indicando sus partes principales.

En una RNA las neuronas son valores numéricos. La semejanza de una RNB y una RNA recae en la manera en la que varias neuronas transmiten su información a una neurona en específico y después esta la transmite a otras neuronas.

El proceso de transmisión de información puede ser modelado y representado. Por ejemplo, en la Figura 4.2 se muestra la representación de una RNA muy simple, la cual muestra un conjunto de neuronas artificiales (valores numéricos) denotadas por $x_1, x_2, \ldots, x_l$, que pueden ser representados por un un vector de la forma $\mathbf{x} = (x_1, x_2, \ldots, x_l)$. El conjunto de neuronas $\mathbf{x}$ está conectado con otras neuronas, en particular denotemos como Σ a una de esas neuronas con la cual se conectarán las neuronas $\mathbf{x}$ por medio de un conjunto de coeficientes que son conocidos como pesos y pueden ser representados por un vector $\mathbf{w} = (w_1, w_2, \ldots, w_l)$.

La forma en la que hacen la conección las neurona $\mathbf{x}$ con la neurona Σ mediante los pesos o bien la recopilación de información, se lleva a cabo mediate la asignación $\sigma = \mathbf{x} \cdot \mathbf{w}$, y por último, para transmitir la información procesada a otras neuronas, la neurona artificial

evalúa en el valor de σ una función $f : \mathbb{R} \to \mathbb{R}$ conocida como *función de activación*, entonces $f(\sigma)$ será la salida o valor de la neurona Σ e información de entrada para otras neuronas.

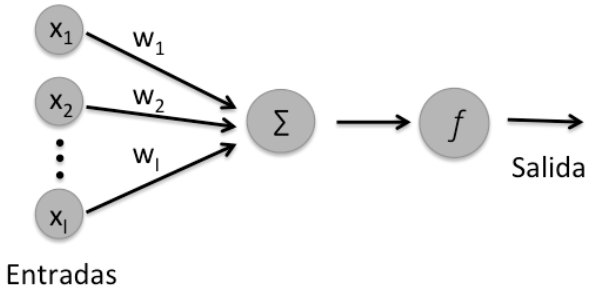

Figura 4.2: En esta figura se muestra la representación gráfica del modelo de una red neuronal artificial en la que un conjunto de neuronas transmiten información a otra neurona y mediante un proceso numérico esta arroja una salida o valor numérico.

La función de activación puede ser una función altamente no lineal. Existen diferentes funciones que pueden ser usadas como función de activación para una red neuronal artificial, por ejemplo, en la Figura 4.3 podemos ver las gráficas de algunas de las funciones de activación más conocidas, la elección de una función de activación depende del tipo de problemas que se quiera resolver. Las funciones de activación son una pieza clave dentro del modelo de las RNA pues dependiendo de la función de activación elegida será lo que modele la RNA.

$$f(x) = \begin{cases} 0 & \text{si } x < 1, \\ 2 & \text{si } x \geq 1. \end{cases} \tag{4.1}$$

$$f(x) = \begin{cases} -2 & \text{si } x < -2, \\ x & \text{si } 2 \leq x \leq 2, \\ 2 & \text{si } x \geq 2. \end{cases} \tag{4.2}$$

$$f(x) = \frac{1}{1 - \mathbf{e}^{-x}}. \tag{4.3}$$

$$f(x) = C\mathbf{e}^{-Kx^2} \text{ para } C, K \in \mathbb{R}. \tag{4.4}$$

En esta sección dimos una explicación del funcionamiento en general y la similitud de este de una RNA con una RNB, sin embargo, como mencionamos en la introducción, existen diferentes modelos o estructuras de RNA, un ejemplo es el modelo *perceptrón multicapa* que será descrito en la siguiente sección.

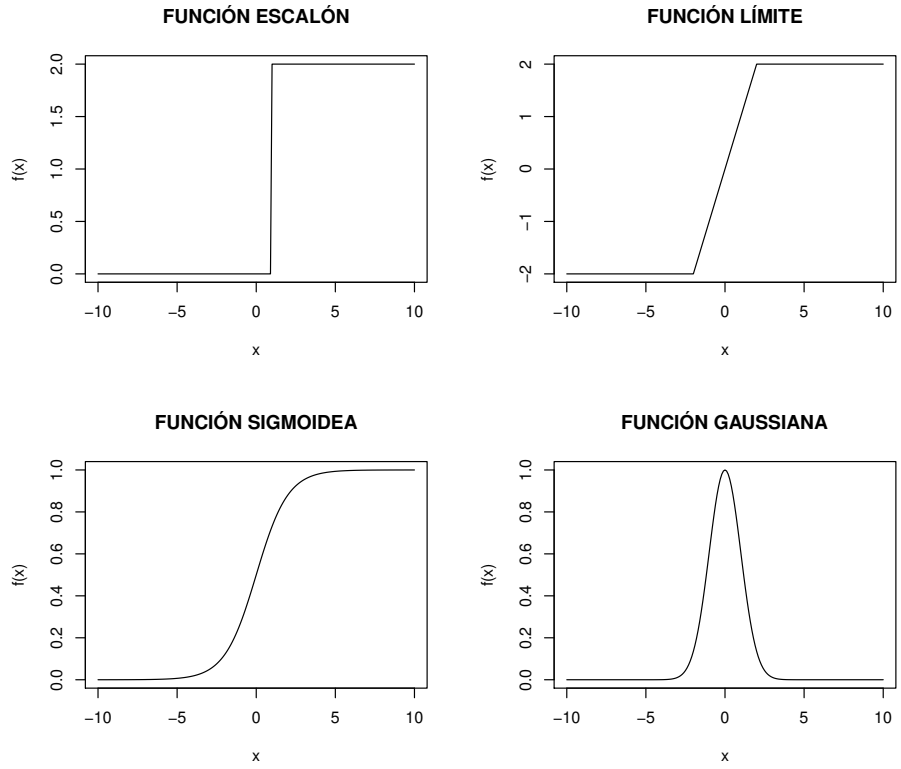

Figura 4.3: En la parte superior derecha vemos la gráfica de la función de activación dada por la ecuación (4.1), esta es conocida como función escalón. En la esquina superior derecha se muestra función de activación límite, definida por la ecuación (4.2). En la parte inferior vemos la función sigmoide, una de las más usadas, ésta está dada por la ecuación (4.3), y por último, en la parte inferior derecha se muestra las gráfica de la función Gaussiana que también puede utilizada como función de activación y está dada por ecuación (4.4), donde C y K son constantes.

4.2. Estructura perceptrón multicapa

Una RNA *perceptrón multicapa* está compuesta por capas ordenadas, cada capa es un conjunto de neuronas y cada neurona de una capa está conectada con cada neurona de la capa consecutiva por medio de pesos. Las RNAs *perceptrón multicapa* cuentan con una capa de entrada en la que el usuario introduce los datos que serán procesados, una capa de salida en donde la información procesada es recibida y capas ocultas que son las capas que están entre la capa de entrada y la capa de salida procesando información.

Cada capa puede ser representada por un vector y los pesos que relacionan dos capas consecutivas pueden ser expresados por medio de una matriz con un número de filas igual a

la longitud del la capa que aparece primero y una cantidad de columnas igual a la longitud de la capa que aparece después menos uno, esto es porque cada vector que representa una capa (a excepción de la capa de salida) tiene una entrada que tendrá siempre el valor 1, a esta neurona o entrada de cada capa se le llama *bias*. Las funciones de activación que conectan capas consecutivas no tienen que ser la misma para cada conexión entre capas.

En la Figura 4.4 se muestra la estructura de una RNA *perceptrón multicapa*, la cual está compuesta por una capa de entrada indicada por el vector **x** , una capa de salida indicada por el vector **y** y solamente una capa oculta indicada por el vextor **z**. Cada flecha hace referencia a un peso o coeficiente, los pesos están indicadas por las matrices $\tilde{W}_{J\times K-1}$ y $W_{K\times L-1}$ donde *J, K, L* indican la longitud de cada capa desde la capa de entrada hasta la capa de salida respectivamente.

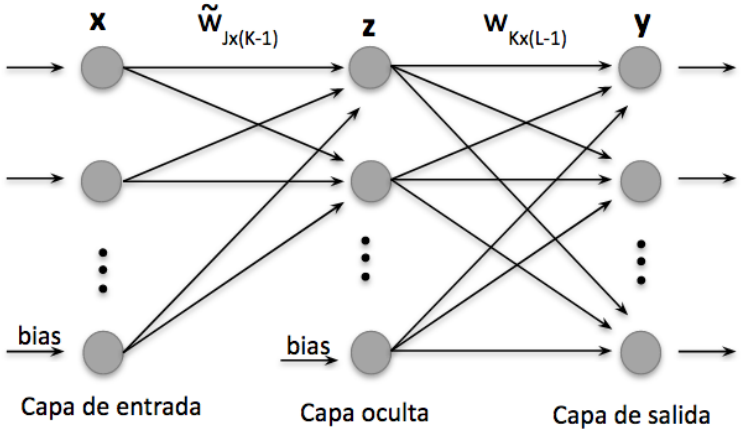

Figura 4.4: En esta imagen se observa la estructura de una red neuronal de alimentación hacia adelante multicapas. En ella se indican las matrices de pesos, las cuales corresponden a las flechas, los vectores de entrada, de salida y la capa oculta. También se puede ver que a excepción de la capa de salida, cada capa tiene una estrada que es el *bias* la cual tiene el valor 1.

Para la red neuronal descrita en la Figura 4.4, las neuronas de la capa oculta $\mathbf{z} = (z_1, z_2, \ldots, z_{K-1})$ están relacionadas con las neuronas de la capa de entrada conforme a la siguiente ecuación

$$z_k = f_1\left(\sum_{j=0}^{J} x_j \tilde{w}_{j,k}\right), \tag{4.5}$$

donde los coeficientes $\tilde{w}_{i,k}$ son entradas de la matriz de pesos $\tilde{W}_{J\times K-1}$ para $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K-1$ y f_1 es una función de activación. Por otro lado, las neuronas de la capa de salida se conectan con las neuronas de la capa oculta y con las neuronas de la capa de

entrada como lo dicta la ecuación

$$y_l = f_2\left(\sum_{k=0}^{K} z_k w_{k,l}\right) = f_2\left(\sum_{k=0}^{K} f_1\left(\sum_{j=0}^{J} x_j \tilde{w}_{j,k}\right) w_{k,l}\right), \tag{4.6}$$

donde los coeficientes $w_{k,l}$ son entradas de la matriz de pesos $W_{K\times L-1}$ para $k = 1, 2, \ldots, K$, $l = 1, 2, \ldots, L-1$ y f_2 una función de activación.

4.3. Entrenamiento

Una RNA es una máquina de aprendizaje y los parámetros ajustables son los pesos. Entonces, los pesos adecuados para la red neuronal depende de los datos de entrenamiento. Dado un conjunto de datos de entrenamiento $\{(x_i, y_i)\}_{i=1}^{p}$, los pesos adecuados para una red neuronal son los parámetros que minimizan el error empírico, como se mencionó en la sección 1.10, en el caso de la RNA perceptrón de tres capas el error empírico está dado por la ecuación

$$EE = \frac{1}{p}\sum_{i=1}^{p} E_i, \tag{4.7}$$

donde cada E_i está dado por

$$E_i = \frac{1}{2}\sum_{l=1}^{L}\left(y_l^i - f_2\left(\sum_{k=0}^{K} z_k^i w_{k,l}\right)_l\right)^2 = \frac{1}{2}\sum_{l=1}^{L}\left(y_l^i - f_2\left(\sum_{k=0}^{K} f_1\left(\sum_{j=0}^{J} x_j^i \tilde{w}_{j,k}\right)_k w_{k,l}\right)_l\right)^2, \tag{4.8}$$

x_j^i indica la entrada j-ésima del vector i-ésimo y de la misma manera y_l^i indica la entrada m-ésima del vector de salida i-ésimo.

Para optimizar esta función se usan métodos de optimización numérica, en este caso lo desarrollaremos con el método de descenso de gradiente, dado por el Algoritmo 1. El método de descenso de gradiente, es un método iterativo que hace uso del gradiente de la función objetivo (ver sección B.1 del Apéndice B), el gradiente es una dirección de descenso y α es el tamaño de paso lo suficientemente pequeño, en este caso es un número fijo. El algoritmo se detiene cuando el gradiente de la función es cercano a cero, para esto se define una tolerancia, en el Algoritmo 1 se expresa como *tol*.

Para hacer uso del algoritmo de descenso de gradiente se quiere calcular la derivara de la ecuación (4.7) con respecto a cada peso, esta es la suma de las derivadas de la ecuación (4.8) con respecto a cada peso. Para realizar el cálculo de derivadas con respecto a los pesos vamos a dividir el problema en dos partes, primero calcularemos las derivadas parciales con respecto a las entradas de la matriz de pesos $W_{K\times L-1}$ y después las derivadas con respecto a

Algoritmo 1 Descenso de gradiente

Entrada: f, x_0, α, tol,i=0
Salida: aproximación de $x^* = arg\,\text{mín}\, f(x)$

1: **mientras** $\|\nabla f(x_i)\| < tol$ **hacer**
2: $\quad x_{i+1} = x_i - \alpha \nabla f(x_i)$
3: $\quad i = i + 1$
4: **fin mientras**
5: **devolver** x_i

la matriz de pesos $\tilde{W}_{J\times K-1}$. Entonces se tiene que:

$$\frac{\partial E_i}{\partial w_{k,l}} = -f_2'\left(\sum_{k=0}^{K} z_k^i w_{k,l}\right)_l \left(y_l^i - f_2\left(\sum_{k=0}^{K} z_k^i w_{k,l}\right)_l\right) z_k^i,$$

$$\frac{\partial E_i}{\partial \tilde{w}_{j,k}} = \frac{\partial E_i}{\partial z_k^i}\frac{\partial z_k^i}{\partial \tilde{w}_{j,k}} = f_1'\left(\sum_{j=0}^{J} x_j^i \tilde{w}_{j,k}\right) x_j^i \sum_{l=1}^{L} -f_2'\left(\sum_{k=0}^{K} z_k^i w_{k,l}\right)\left(y_l^i - f_2\left(\sum_{k=0}^{K} z_k^i w_{k,l}\right)_l\right) w_{k,l}.$$

De este modo teneniendo las derivadas parciales de la función de error con respecto a cada uno de los pesos se podrá minimizar utilizando el método de gradiente descendente para entrenar la red neuronal. Existen otros métodos de optimización numérica que podrían ser aplicados, por ejemplo el método de Newton o algún método cuasi-Newton pero estos requieren que se calcule o aproxime el Hessiano, es decir, ocupan derivadas de segundo orden y no serán implementados en este trabajo. Este tipo de métodos de optimización por lo regular alcanzan solo óptimos locales dependiendo de los valores con los que se inicialice el algoritmo.

El entrenamiento de una RNA, dados los datos $\{(x_i, y_i)\}_{i=1}^{p}$, no se realiza con todos los datos disponibles. Para poder evaluar la capacidad predictiva de nuestra red neuronal los datos se dividen en tres conjuntos:

- El *conjunto de entrenamiento*, este sirve para entrenar la red neuronal, con la mayor cantidad de datos, digamos un 70 % de los datos. El algoritmo de entrenamiento de una RNA en cada iteración intenta minimizar el error empírico con este conjunto de datos.

- El *conjunto de validación* está compuesto por otra parte del conjunto de datos. Este conjunto de datos no tiene intersección con el conjunto de entrenamiento. Digamos que el conjunto de validación está compuesto por un 15 % de los datos y este conjunto sirve para calcular el error de la RNA en cada iteración conjuntamente con el de entrenamiento pero con los datos del conjunto de validación.

- El *conjunto de prueba o predicción* está compuesto por el conjunto de datos sobrantes, digamos el otro 15 %. Una vez entrenada la red neuronal se calcula el porcentaje de aciertos que tuvo la red neuronal (en el caso de ser un clasificador) con este conjunto y ese porcentaje se le conoce como *capacidad de predicción* (CDP) de la RNA.

Para el caso de series de tiempo, la división de los conjuntos de entrenamiento, validación y prueba se hace mediante ventanas de tiempo como se muestra en la Figura 4.5, sin embargo, para otro tipo de datos la selección para los conjuntos de entrenamiento, validación y prueba puede ser aleatoria como es el caso del problema de clasificación de imagenes de números que veremos en la sección 6.1. La selección de los conjuntos dependerá del problema que se esté tratando.

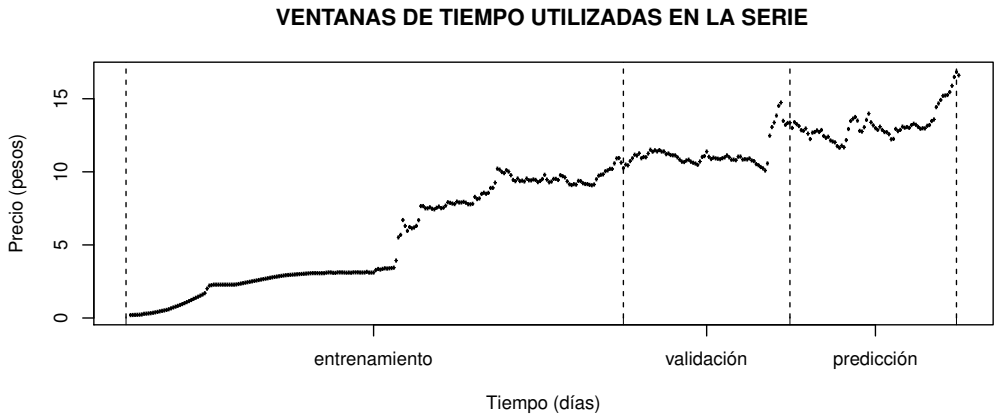

Figura 4.5: En esta imagen se muestra el conjunto de datos de la serie de tiempo de la cotización del dolar. Las lineas punteadas verticales delimitan las ventanas de tiempo de una serie, o bien, la divición de los datos en un conjunto de entrenamiento, un conjunto de validación y un conjunto de prueba.

El *error de validación* es el error empírico de la RNA evaluada con los datos del conjunto de validación. El *error de predicción* es el error empírico de la RNA evaluada con los datos del conjunto de prueba. El interés de calcular el *error de validación* y el *error de predicción* es para ver qué capacidad de generalización tiene la red neuronal, es dicir, dado un conjunto nuevo de datos con el que no fue entrenada nuestra RNA qué tan buena es acertando en el valor correcto.

Por lo regular el algoritmo de entrenamiento no es dejado hasta la convergencia, es decir, hasta que el gradiente de la función de error es identicamente cero, pues esto podría no suceder o tardar demasiado tiempo. El *error de entrenamiento* o error empírico de la RNA evaluada en el conjunto de entrenamiento, siempre disminuye en cada iteración, sin embargo, esto no quiere decir que se mejore la capacidad de predicción para el resto de los

datos, al contrario, se puede dar el caso en el que la RNA se *sobreentrene*, es decir, que tenga un error de entrenamiento muy bajo, pero un error de validación muy alto y también muy poca CDP.

Un criterio de paro en una red neuronal es analizando el error de validación y cuando este deje de disminuir dejar de actualizar los pesos del modelo. En general la CDP y el EE dependen del conjunto en que se evalue, en los problemas del capítulo de resultados se probó que este es un buen criterio, pues en todos los casos a excepción de la serie del covid, se obtuvo que la CDP (en los casos bien entrenados) con los pesos entrenados hasta que el error de validación fue mínimo, fue mejor o igual más del 50 % de las veces que la CDP con los pesos entrenados las 200 iteraciones. Durante este trabajo estaremos tomando la CDP y el EE en el momento que el error de validación alcanzó su mínimo. Decimos que una RNA *se entrenó bien* si alcanzó un error de validación mínimo antes del número total de iteraciones.

En la Figura 4.6 se muestra la gráfica del error de entrenamiento y el error de validación en cada iteración, en este caso se usó la red neuronal con 95 neuronas en la capa oculta que se implementó para resolver el problema de clasificación de imagenes de números de la sección 6.1 del capítulo de resultados. Este tipo de gráficos nos sirven como referencia para saber cuántas iteraciones es suficiente dejar entrenando una RNA para un problema específico.

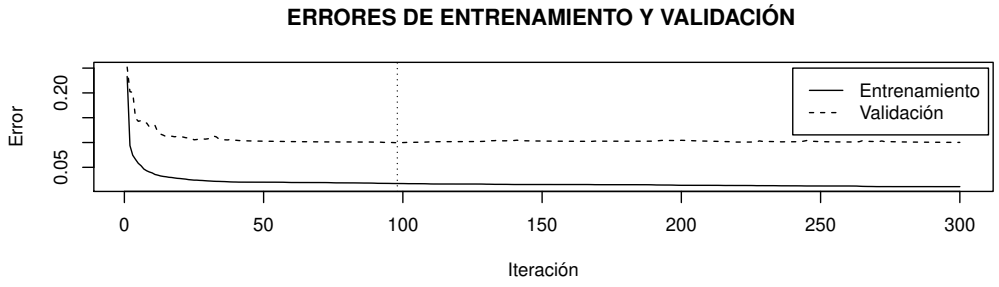

Figura 4.6: En esta gráfica podemos ver como el error de entrenamiento siempre es menor que el de validación y aunque el error de entrenamiento disminuye en cada iteración, el error de validación deja de disminur en la iteración 98.

En el siguiente capítulo se estudiará otro modelo de inteligencia artificial que es principalmente utilizado como clasificador. Este modelo es el de las MSV. La ventaja que tiene dicho modelo sobre las RNAs, es que este modelo alcanza su óptimo global, por ser un problema de programación convexa.

Capítulo 5

Máquinas de soporte vectorial

Las *máquinas de soporte vectorial* (MSV) son un conjunto de algoritmos de aprendizaje supervisado, las cuales tienen origen en trabajos sobre el aprendizaje estadístico. En principio este algoritmo fue pensado para hacer clasificación binaria, aunque en la actualidad son usadas para resolver problemas de regresión, agrupamiento y multiclasificación. Existen varios campos donde han sido utilizadas con éxito, tales como visión artificial, reconocimiento de caracteres, clasificación de proteinas, análisis de series temporales y otros. Alguna de las aplicaciones más comunes que podemos encontrar de las MSV es el reconocimiento de patrones, que se ha aplicado a reconocimiento de dígitos manuscritos, reconocimiento de objetos, clasificación de texto, identificación de locutores y detección de rostros en imágenes, ver [6, 13].

A diferencia de otros métodos de aprendizaje por ejemplo las RNA o las regresiones lineales, las MSV no se basan en minimizar los errores cometidos por el modelo generados a partir de los ejemplos de entrenamiento o *error empírico*. Podemos considerar a las MSV como clasificadores lineales, ya que su tarea en el caso de clasificador binario es encontrar un hiperplano $h(x)$ que separe las clases con las que se está trabajando de tal forma que equidiste de los ejemplos más cercanos de cada clase y de esta forma conseguir un margen máximo de cada lado del hiperplano el cuál es la distancia del ejemplo más cercano de la clase al hiperplano, estos ejemplos son conocidos como vectores soporte y basta conocer quienes son para construir el hiperplano, esto se logra optimizando la normal del hiperplano. Al proceso de maximizar el margen se le llama minimización del *riesgo estructural* y es en lo que se basan las MSV.

Desde el punto algorítmico el problema de minimizar la normal representa un problema de optimización cuadrática con restricciones de desigualdad lineales, que pueden ser resueltas mediante técnicas de programación cuadrática (revisar Apéndice B). La propiedad de convexidad garantiza una solución única a diferencia de las RNA en las que se pueden estancar en puntos óptimos locales.

Durante el desarrollo de este capítulo estudiaremos las MSV para clasificación binaria y para regresiones. Estudiaremos diferentes clases de MSV empezando por las MSV

lineales para el caso de conjuntos de datos separables, en la sección 5.1.1. En la sección 5.1.2 podremos ver el caso de las MSV lineales para el caso de conjuntos de datos cuasi-separables. En la sección 5.2 haremos la generalización de las MSV para el caso de datos cuasi-separables no linealmente, introduciendo las funciones denominadas kerneles. Finalmente en la sección 5.3 estudiaremos la forma de aplicar las MSV para hacer regresiones. Se recomienda revisar el Apéndice B para ver una breve introducción sobre optimización, la cual es necesaria para entender como se resuelve el problema que plantean las MSV.

5.1. Maquinas de soporte vectorial lineales

5.1.1. Caso separable

Para el caso de las MSV de clasificación binaria se quiere clasificar un conjunto de puntos $\{x_i, y_i\}_{i=1}^n$ donde n es el número de datos de entrenamiento, $x_i \in \mathbb{R}^d$ y $y_i \in \{-1, 1\}$. Se dice que el conjunto $\{x_i, y_i\}_{i=1}^n$ es linealmente separable si exite un hiperplano $h(x) = \mathbf{w}\cdot x + b$ tal que

$$\mathbf{w}\cdot x_i + b \geq 0 \text{ si } y_i = +1, \tag{5.1}$$

$$\mathbf{w}\cdot x_i + b \leq 0 \text{ si } y_i = -1, \tag{5.2}$$

$h(x)$ es conocido como *hiperplano de separación*, donde $\mathbf{w} \in \mathbb{R}^n$ es la normal al hiperplano y $b \in \mathbb{R}$ es la distancia más corta del origen a $h(x)$. Si el conjunto es separable entonces existe una infinidad de hiperplanos que cumplen las ecuaciones (5.1) y (5.2), a nosotros nos interesa un hiperplano que sea óptimo, para esto se introduce el concepto de margen.

Sea $\{x \mid \mathbf{w}\cdot x + b = 0, x, \mathbf{w} \in \mathbb{R}^n\}$ un hiperplano de separación, entonces la distancia más corta del origen al hiperplano de separación es $\frac{|b|}{\|\mathbf{w}\|}$, donde $\|\mathbf{w}\|$ es la norma euclideana de $\mathbf{w}$. Llamemos d_+ la distancia más corta entre el hiperplano de separación y el punto más cercano de la clase con valor $y_i = 1$ y d_- la distancia más corta entre el hiperplano y el punto más cercano de la clase con $y_i = -1$. Al conjunto de datos del conjunto de entrenamiento que está a una distancia d_+ o d_- del hiperplano de separación se les conoce como *vectores de soporte*. Definamos H_1 y H_2 a los hiperplanos que son paralelos al hiperplano de separación y que están a distancia de éste d_+ y d_- respectivamente, éstos son conosidos como *hiperplanos de soporte*. La separación entre los hiperplanos H_1 y H_2 le llamaremos *margen* y es equivalente a $d_+ + d_-$, nótese que entre los hiperplanos H_1 y H_2 no existen datos de entrenamiento ya que son linealmente separables. En la Figura 5.2 podemos visualizar un ejemplo de un conjunto de puntos separables en $\mathbb{R}^2$, éste nos ayudará a ver de manera grafica las definiciones descritas en este párrafo.

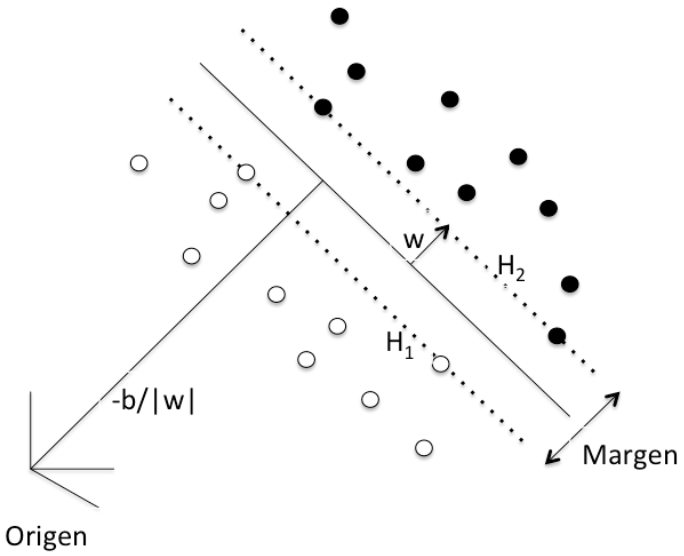

Figura 5.1: En esta figura podemos ver un conjunto separable en $\mathbb{R}^2$ donde los puntos en negro indican la clase con $y_i = 1$ y los puntos en blanco muestra la clase con $y_i = -1$, éstos se encuentran separados por un hiperplano o bien una linea recta, para este caso, los vectores de soporte se muestran sobre los hiperplanos H_1 y H_2.

La tarea de una MSV es encontrar un hiperplano de separación que maximice el margen entre los hiperplanos H_1 y H_2, pueden existir una infinidad de hiperplanos que cumplen con esta propiedad, sin embargo para el caso lineal separable es sufisiente con suponer al hiperplano H_1 como $\{x \mid \mathbf{w}\cdot x + b = 1, x, \mathbf{w} \in \mathbb{R}^n\}$ y al hiperplano H_2 como $\{x \mid \mathbf{w}\cdot x + b = -1, x, \mathbf{w} \in \mathbb{R}^n\}$, este resultado se puede consultar en [13]. Decimos entonces que en un conjunto separable linealmente los datos de entrenamiento sarisfacen las ecuaciones (5.3) y (5.4) o bien la ecuación (5.5).

$$x_i\cdot \mathbf{w} + b \geq +1 \;\text{ si }\; y_i = +1 \tag{5.3}$$
$$x_i\cdot \mathbf{w} + b \leq -1 \;\text{ si }\; y_i = -1 \tag{5.4}$$

$$y_i(x_i\cdot \mathbf{w} + b) - 1 \leq 0 \;\text{ para }\; i \in \{1, 2, \ldots, n\} \tag{5.5}$$

Como ya mencionamos previamente, el margen está definido por los hiperplanos H_1 y H_2, ambos hiperplanos son paralelos al hiperplano de separación y su normal es justamente $\mathbf{w}$, por lo tanto la distancia del origen a H_1 está dada por $|1 - b|/\|\mathbf{w}\|$ y el hiperplano H_2 tiene una distancia al origen de $|-1-b|/\|\mathbf{w}\|$, de donde obtenemos que $d_+ = d_- = 1/\|\mathbf{w}\|$, la prueba de que $d_+ = d_-$ se puede consultar en [13]. El margen es simplemente $\frac{2}{\|\mathbf{w}\|}$, de aquí

se puede ver que el margen depende solo de la normal del hiperplano de separación.

El problema de una MSV se reduce entonces a un problema de optimización donde se quiere maximizar el margen $\frac{2}{\|\mathbf{w}\|}$ o bien minimizar la función $\|\mathbf{w}\|$, la cual será sustituida por la función $\|\mathbf{w}\|^2$ por convenencia, sujeta a las restricciones descritas en la ecuación (5.5). El problema de optimización a resolver finalmente es el siguiente:

$$\begin{aligned} &\text{minimizar } \|\mathbf{w}\|^2 \\ &\quad \text{sujeta a } y_i(x_i \cdot \mathbf{w} + b) - 1 \leq 0 \text{ para } i \in \{1, 2, \ldots, n\}. \end{aligned} \tag{5.6}$$

El problema de la ecuación (5.6) es un problema de optimización convexo con restricciónes lineales, el cuál es un problema que puede ser atacado y resuelto con teoría Lagrangiana, se puede revisar el Apéndice B para más detalles sobre teoría de optimización.

La teoría de Lagrange nos dice que el problema de la ecuación (5.6) se reduce a solucionar el problema de la ecuación (5.7), éste es conocido como *problema primal*, donde $\alpha_i \geq 0$ para $i \in \{1, 2, ..., n\}$ son los multiplicadores de Lagrange que corresponden a las restricciones de la ecuación (5.5) y L_P es conocida como la *función Lagrangiana primal*. Este método nos permitirá poder hacer una generalización de forma sencilla al caso no lineal.

$$\text{minimizar } L_P(\mathbf{w}, \alpha, b) \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(x_1 \cdot \mathbf{w} + b) - 1] \tag{5.7}$$

El problema (5.7) es un problema de optimización sin restricciones y puede ser resuelto aplicando las condiciones de Karush-Kuhn-Tuker (KKT) a la función Lagrangiana $L_P(\mathbf{w}, \alpha, b)$. De aplicar la primera condición de KKT obtenemos el resultado de la ecuación (5.8) y de la ecuación (5.9), además la ecuación (5.10) es el resultado de aplicar la conocida condición complementaria de KKT (revisar el Apéndice B).

$$\frac{\partial L_P(\mathbf{w}, \alpha, b)}{\partial w} \rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{5.8}$$

$$\frac{\partial L_P(\mathbf{w}, \alpha, b)}{\partial b} \rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{5.9}$$

$$\alpha_i(1 - y_i(\mathbf{w} \cdot x_i + b)) = 0 \text{ para } i \in \{1, 2, \ldots, n\} \tag{5.10}$$

Sustituyendo los resultados de la ecuación (5.8) y la ecuación (5.9) en la función Lagrangiana L_P obtenemos la función Lagrangiana dual $L_D(\alpha)$ dada por la ecuación (5.11) y el problema dual equivalente dado en la ecuación 5.12. Se dice que computacionalmente es más fácil de resolver el problema dual, ya que este depende solamente de el vector α de dimensión n, en lugar del vector w de dimensión d, de este modo, si $n << d$, sería mucho más rápido solucionar el problema dual.

$$L_D(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \tag{5.11}$$

$$\begin{aligned} \text{maximizar } \; & L_D(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ \text{sujeta a } \; & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \;\text{ para }\; i \in \{1, 2, \ldots, n\} \end{aligned} \tag{5.12}$$

Una vez conociendo el vector α del problema dual, se sustituye en la ecuación (5.8) y se obtiene la normal del plano, la ecuación del plano estaría dada entonces por la ecuación (5.13). Para encontrar el parámetro b basta despejar la ecuación (5.10).

$$h(x) = \sum_{i=1}^{n} \alpha_i y_i (x \cdot x_i) + b \tag{5.13}$$

5.1.2. Caso cuasi-separable

Según la sección anterior un conjunto de datos $\{x_i, y_i\}_{i=1}^n$ donde n es el número de datos de entrenamiento, $x_i \in \mathbb{R}^d$ y $y_i \in \{-1, 1\}$ es no separable si no se puede encontrar un hiperplano que satisfaga las ecuaciones (5.1) y (5.2), por lo tanto, es muy dificil que en la práctica se encuentren con conjuntos de datos perfectamente separables, ya que los problemas reales se caracterizan por tener ejemplos con ruido. Para resolver este problema se propone permitir tener errores de clasificación en el conjunto de datos de entrenamiento, sin abandonar el objetivo de obtener un margen máximo para el conjunto de datos que sí están bien clasificados.

Para permitir que puedan existir datos de entrenamiento con error en su clasificación la idea es relajar las ecuaciones de (5.5), para esto se introducen las conocidas *variables de holgura* $\xi_i \geq 0$ para $i = 1, \ldots, n$, que permiten cuantificar el error de clasificación en los datos de entrenamiento mediante la suma de todas éstas $(\sum_{i=1}^n \xi_i)$. De relajar las ecuaciones de (5.5), obtenemos las restricciones de las ecuaciones (5.14) y (5.15) o bien las de la ecuación (5.16).

$$x_i \cdot \mathbf{w} + b \geq +1 - \xi_i \qquad \text{si } \; y_i = +1 \tag{5.14}$$

$$x_i \cdot \mathbf{w} + b \leq -1 + \xi_i \qquad \text{si } \; y_i = -1 \tag{5.15}$$

$$y_i(x_i \cdot \mathbf{w} + b) + 1 - \xi_i \geq 0 \text{ para } \ i \in \{1, 2, \ldots, n\} \tag{5.16}$$

De acuerdo con las ecuaciones, cuando $\xi_i = 0$ quiere decir que el punto x_i se encuentra bien clasificado. Si $0 < \xi_i < 1$, entonces x_i tiene error de clasificación pero no está mal clasificado, esto quiere decir que se encuentra entre el hiperplano de separación y el hiperplano que define el margen de su clase. Un punto x_i se dice que está mal clasificado si se encuetra del lado opuesto al hiperplano de separación de su clase, esto quiere decir que $\xi_i > 1$. Ver la Figuara 5.2 para ver un ejemplo de un conjunto de datos mal clasificados.

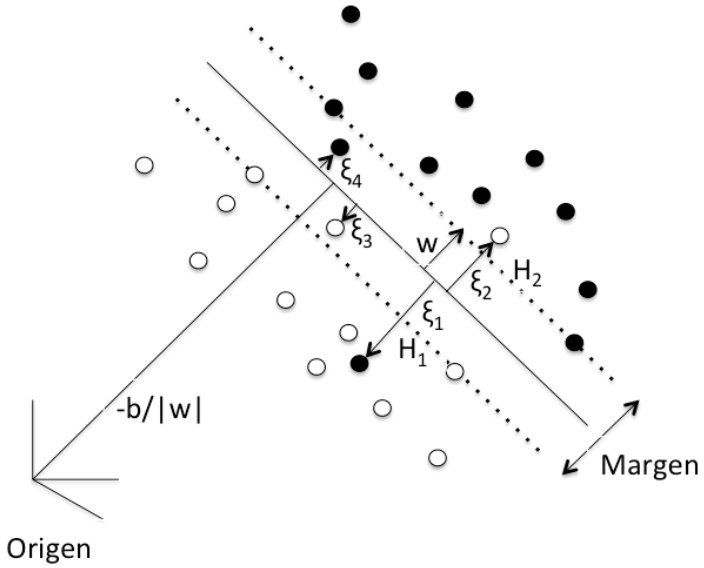

Figura 5.2: En esta figura podemos ver un conjunto cuasi-separable en $\mathbb{R}^2$, donde los puntos en negro indican la clase con $y_i = 1$ y los puntos en blanco muestra la clase con $y_i = -1$, éstos están separados por un hiperplano o bien una linea recta (en este caso). Los vectores de soporte se muestran sobre los hiperplanos H_1 y H_2, los puntos con ξ_1 y ξ_2 son ejemplos de puntos mal clasificados, mientras que los puntos con ξ_3 y ξ_4 son ejemplos de puntos con error de clasificación pero bien clasificados.

Nuestro interés ahora será no solo maximizar el margen en los datos de entrenamiento bien clasificados, sino minimizar el error de clasificación de los datos de entrenamiento, por lo tanto, la función que nos interesa minimizar en este caso es

$$f(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i,$$

donde C representa un parámetro a fijar por el usuario, éste permitirá determinar qué tanto error de clasificación se estará permitiendo. Si C es muy grande quiere decir que se permitirá poco error de clasificación, incluso si se hace demasiado grande estaríamos pensando que el problema es el del caso separable. Cuando C es muy pequeña se permite que exista más error de clasificación. El problema de optimización convexo con restricciónes lineales a resolver para el caso cuasi-separable sería el siguiente:

$$\begin{aligned} \text{minimizar} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{sujeta a} \quad & y_i(x_i\cdot\mathbf{w}+b)+1-\xi_i \geq 0 \ \text{ para } \ i\in\{1,2,\ldots,n\} \\ & \xi_i \geq 0 \ \text{ para } \ i\in\{1,2,\ldots,n\}. \end{aligned} \tag{5.17}$$

Este problema se resuelve encontrando el mínimo de su función Lagrangiana, es decir, se reduce a resolver el problema primal (5.18), donde $L_P(\mathbf{w},b,\xi,\alpha,\beta)$ es la función Lagrangiana del problema (5.17) y $\alpha_i \geq 0, \beta_i \geq 0$ son los multiplicadores de Lagrange.

$$\text{minimizar } \ L_P(\mathbf{w},b,\xi,\alpha,\beta) \equiv \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(x_1\cdot\mathbf{w}+b)-1+\xi_i] - \sum_{i=1}^{n}\beta_i\xi_i \tag{5.18}$$

Al igual que en el caso separable, el problema primal (5.18) puede ser resuelto aplicando las condiciones de KKT. Las ecuaciones (5.19), (5.20) y (5.21) son de aplicar la primera condición de KKT y las ecuaciones (5.22) y (5.23) son las resultantes de aplicar la condición complementaria de KKT.

$$\frac{\partial L_P(\mathbf{w},b,\xi,\alpha,\beta)}{\partial w} \rightarrow \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i x_i \tag{5.19}$$

$$\frac{\partial L_P(\mathbf{w},b,\xi,\alpha,\beta)}{\partial b} \rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{5.20}$$

$$\frac{\partial L_P(\mathbf{w},b,\xi,\alpha,\beta)}{\partial \xi_i} \rightarrow C = \alpha_i + \beta_i \ \text{ para } \ i\in\{1,2,\ldots,n\} \tag{5.21}$$

$$\alpha_i(1-y_i(\mathbf{w}\cdot x_i+b)-\xi_i) = 0 \ \text{ para } i\in\{1,2,\ldots,n\} \tag{5.22}$$

$$\beta_i\cdot\xi_i = 0 \ \text{ para } i\in\{1,2,\ldots,n\} \tag{5.23}$$

Susituyendo la ecuaciones (5.19) y (5.20) en la función Lagrangiana $L_P(\mathbf{w},b,\xi,\alpha,\beta)$ obtenemos el la función Lagrangiana dual, dada por la ecuación (5.24). Podemos ver que

esta función no cambia con respecto a la ecuación (5.11) del problema separable, sin embargo, el problema dual, dado por la ecuación (5.25), acota los multiplicadores de Lagrange α_i por la constante C, este resultado se obtiene analizando las ecuaciones (5.22) y (5.23).

$$L_D(\alpha) = \sum_{i}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{5.24}$$

$$\begin{aligned} \text{maximizar} \;\; & L_D(\alpha) = \sum_{i}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ \text{sujeta a} \;\; & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \;\; \text{para} \;\; i \in \{1, 2, \ldots, n\} \end{aligned} \tag{5.25}$$

Al igual que en el caso anterior, la solución del problema dual nos permitirá expresar el hiperplano de separación en términos de α y también encontrar el término b despejando la ecuación (5.22). La ecuación del hiperplano de separacieon en este caso está dada por la siguiente ecuación:

$$h(x) = \sum_{i=1}^{n} \alpha_i y_i (x \cdot x_i) + b. \tag{5.26}$$

5.2. Máquinas de soporte vectorial no lineales

Cuando los datos no pueden ser clasificados linealmente, se propone mapear los datos a un espacio de de dimensión mayor, conocido como *espacio de característica*, donde los datos mapeados pueden ser clasificados linealmente, mediante una función $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$, donde m es la dimensión del espacio de característica, véase la Figura 5.3. El problema a resolver en este caso es el mismo que en las sección anterior, solamente que en un espacio diferente, es decir, ahora sustituiremos los valores de x_i por $\Phi(x_i)$ para $i \in \{1, 2, \ldots, n\}$, por ejemplo, ahora la función de decisión en el espacio de característica estaría dada por la siguiente ecuación:

$$h(x) = \sum_{i=1}^{n} \alpha_i y_i < \Phi(x), \Phi(x_i) > .$$

Si definimos la función kernel como $K(x, y) = \Phi(x) \cdot \Phi(y)$ entonces el problema a resolver para el caso de datos no linealmente separables estaría dado por la ecuación (5.27). Notemos que en este caso se quiere resolver solo el problema dual y no el primal, ya que el primal tendría una dimensión mayor en el espacio característica y esto lo haría muy difícil

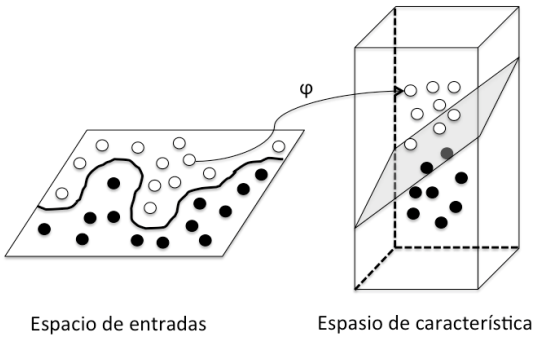

Figura 5.3: Datos linealmente no separables en el espacio de entrada pero separables en el espacio de característica.

de resolver numéricamente hablando.

$$\begin{aligned} \text{maximizar } & L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{sujeto a } & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \;\; i = 1, \ldots, n \end{aligned} \tag{5.27}$$

Muchas veces la función $\Phi(x)$ no es conocida explícitamente pero la función $K(x, y)$ sí, ésta es otra de las ventajas de trabajar con el problema dual y no con el primal. Existen algunos criterios para definir las funciones kernel, sin embargo, no serán mencionadas en este trabajo. Algunas de las funciones kernel más conocidas son las siguientes:

$$\text{Lineal } \; K_L(x, y) = (x \cdot y), \tag{5.28}$$

$$\text{Polinómico } \; K_P(x, y) = (x \cdot y + 1)^p, \tag{5.29}$$

$$\text{Gausiano } \; K_G(x, y) = e^{\frac{-\|x-y\|^2}{2\sigma^2}}, \tag{5.30}$$

$$\text{Sigmoidal } \; K_S(x, y) = tanh(a(x \cdot y) + b). \tag{5.31}$$

5.3. MSV para regresiones

Las máquinas de soporte vectorial pueden ser adaptadas también para resolver problemas de regresión y es común designarles SVR (del inglés Support Vector Regression). Aho-

ra queremos clasificar el conjunto $\{x_i, y_i\}_{i=1}^n$, donde n es el número de elementos, $x_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$. En este caso nos gustaría que todos los elementos del conjunto $\{x_i, y_i\}_{i=1}^n$ se pudieran ajustar o cuasi-ajustar a un hiperplano $h(x) = \mathbf{w}\cdot x + b$, donde $\mathbf{w}$ es la normal al hiperplano y b la distancia del hiperlano al origen. Con ajustar nos referimos a que se cumpla la ecuación

$$h(x_i) = \mathbf{w}\cdot x_i + b = y_i \text{ para } i \in \{1, 2, \ldots, n\}.$$

Dado que en la práctica es muy dificil que los datos de entrenamiento se ajusten al modelo lineal con un error de predicción igual a cero, se recurre al concepto de margen blando. Para introducir el concepto de margen blando y relajar las condiciones de error entre el valor obtenido por el modelo y el valor real se introduce la *función de pérdida ε-insensible* L_ε definida por la ecuación

$$L_\varepsilon = \begin{cases} 0 & \text{si } |y - h(x)| \leq \varepsilon, \\ |y - h(x)| - \varepsilon & \text{en otro caso.} \end{cases} \tag{5.32}$$

También se introducen las variables ξ_i^+ y ξ_i^-, las cuales cuantificarán el error fuera de la región tubular 2ε, donde $\xi_i^+ > 0$ si $h(x_i) - y_i > \varepsilon$ y cero en el otro caso. De igual manera $\xi_i^- > 0$ si $y_i - h(x_i) > 0$ y cero en el otro caso. Como ξ_i^+ y ξ_i^- no pueden ser mayor que cero simultáneamente, entonces $\xi_i^+ \cdot \xi_i^- = 0$.

El problema primal en este caso es muy parecido al problema primal del caso cuasi-parable, pero en este caso se introducen dos variables de holgura, dando el siguiente problema:

$$\begin{aligned} \text{minimizar} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) \\ \text{sujeta a} \quad & (x_i\cdot \mathbf{w} + b) - y_i - \varepsilon - \xi_i^+ \leq 0 \text{ para } i \in \{1, 2, \ldots, n\}, \\ & y_i - (x_i\cdot \mathbf{w} + b) - \varepsilon - \xi_i^- \leq 0 \text{ para } i \in \{1, 2, \ldots, n\}, \\ & \xi_i^+, \xi_i^- \geq 0 \text{ para } i \in \{1, 2, \ldots, n\}. \end{aligned} \tag{5.33}$$

Al igual que en los problemas anteriores, se siguen los mismos pasos para resolver el problema primal. La función Lagrangiana correspondiente al problema primal (5.33), está dada por la ecuación (5.34), entonces queremos encontrar el mínimo de la función Lagrangiana, para resolver el problema aplicamos las condiciones de KKT.

$$
\begin{aligned}
L_P(\mathbf{w}, b, \xi^+, \xi^-, \alpha^+, \alpha^-, \beta^+, \beta^-) \equiv & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i,j=1}^{n}(\xi_i^+ + \xi_j^-) + \\
& \sum_{i=1}^{n}\alpha_i^+[(x_i \cdot \mathbf{w} + b) - y_i - \varepsilon - \xi_i^+] + \\
& \sum_{i=1}^{n}\alpha_i^-[y_i - (x_i \cdot \mathbf{w} + b) - \varepsilon - \xi_i^-] - \\
& \sum_{i=1}^{n}\beta_i^+\xi_i^+ - \sum_{i=1}^{n}\beta_i^-\xi_i^-
\end{aligned}
\tag{5.34}
$$

De aplicar la primera condición de KKT a la función lagrangiana de la ecuación (5.34) obtenemos los resultados de las ecuaciones (5.35), (5.36), (5.37) y (5.38), mientras que si se aplica la condición complementaria de KKT obtenemos las ecuaciones (5.39), (5.40), (5.42) y (5.42).

$$
\frac{\partial L_P(\mathbf{w}, b, \xi^+, \xi^-, \alpha^+, \alpha^-, \beta^+, \beta^-)}{\partial w} \rightarrow \mathbf{w} = \sum_{i=1}^{n}(\alpha_i^- - \alpha_i^+)x_i \tag{5.35}
$$

$$
\frac{\partial L_P(\mathbf{w}, b, \xi^+, \xi^-, \alpha^+, \alpha^-, \beta^+, \beta^-)}{\partial b} \rightarrow \sum_{i=1}^{n}(\alpha_i^- - \alpha_i^+) = 0 \tag{5.36}
$$

$$
\frac{\partial L_P(\mathbf{w}, b, \xi^+, \xi^-, \alpha^+, \alpha^-, \beta^+, \beta^-)}{\partial \xi_i^+} \rightarrow \beta_i^+ = C - \alpha_i^+ \text{ para } i \in \{1, 2, \ldots, n\} \tag{5.37}
$$

$$
\frac{\partial L_P(\mathbf{w}, b, \xi^+, \xi^-, \alpha^+, \alpha^-, \beta^+, \beta^-)}{\partial \xi_i^-} \rightarrow \beta_i^- = C - \alpha_i^- \text{ para } i \in \{1, 2, \ldots, n\} \tag{5.38}
$$

$$
\alpha_i^+[(x_i \cdot \mathbf{w} + b) - y_i - \varepsilon - \xi_i^+] = 0 \text{ para } i \in \{1, 2, \ldots, n\} \tag{5.39}
$$

$$
\alpha_i^-[y_i - (x_i \cdot \mathbf{w} + b) - \varepsilon - \xi_i^-] = 0 \text{ para } i \in \{1, 2, \ldots, n\} \tag{5.40}
$$

$$
\beta_i^+ \cdot \xi_i^+ = 0 \text{ para } i \in \{1, 2, \ldots, n\} \tag{5.41}
$$

$$
\beta_i^- \cdot \xi_i^- = 0 \text{ para } i \in \{1, 2, \ldots, n\} \tag{5.42}
$$

Sustituyendo los resultados de las ecuaciones (5.35) y (5.36) en la función Lagrangiana primal, se obtiene la función Lagrangiana dual que está dada por la ecuación (5.43) y el problema dual dado por la ecuación (5.44).

$$\begin{aligned} L(\alpha^+, \alpha^-) = &\sum_{i=1}^{n}(\alpha_i^- - \alpha_i^+)y_i - \varepsilon \sum_{i=1}^{n}(\alpha_i^- + \alpha_i^+) - \\ &\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^- - \alpha_i^+)(\alpha_j^- - \alpha_j^+)(x_i \cdot x_j) \end{aligned} \tag{5.43}$$

$$\begin{aligned} \text{maximizar } &\sum_{i=1}^{n}(\alpha_i^- - \alpha_i^+)y_i - \varepsilon \sum_{i=1}^{n}(\alpha_i^- + \alpha_i^+) - \\ &\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^- - \alpha_i^+)(\alpha_j^- - \alpha_j^+)(x_i \cdot x_j) \\ \text{sujeto a } &\sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-) = 0 \\ &0 \leq \alpha_i^+, \alpha_i^- \leq C \text{ para } \in \{1, 2, \ldots, n\} \end{aligned} \tag{5.44}$$

Al igual que en los casos anteriores, una vez resolviendo el problema dual se puede expresar el hiperplano en terminos de α^+ y α^-. La ecuación (5.45) indica la ecuación del hiperplano que mejor aproxima a los datos de entrenamiento y el parámetro b puede ser encontrado usando las ecuaciones (5.46) y (5.47).

$$h(x) = \sum_{i=1}^{n}(\alpha_i^- - \alpha_i^+)(\mathbf{x} \cdot x_i) + b \tag{5.45}$$

$$b = y_i + \varepsilon - (\mathbf{w} \cdot x_i) \text{ si } 0 < \alpha_i^+ < C \tag{5.46}$$

$$b = y_i - \varepsilon - (\mathbf{w} \cdot x_i) \text{ si } 0 < \alpha_i^- < C \tag{5.47}$$

Podemos aplicar el mismo truco de mapear los datos a un espacio donde los datos puedan ser aproximados linealmente y se genere menos error de predicción, para esto se vuelve a usar la función kernel para mapear los datos, entonces el hiperplano de aproximación estaría dado por la ecuación (5.48), el cual está dado en terminos de α^+ y α^-, los cuales se obtienen de resolver el problema dual de la ecuación (5.49).

$$h(x) = \sum_{i=1}^{n} (\alpha_i^- - \alpha_i^+) K(\mathbf{x} \cdot x_i) \tag{5.48}$$

$$\begin{aligned} \text{maximizar} \sum_{i=1}^{n} (\alpha_i^- - \alpha_i^+) y_i - \varepsilon \sum_{i=1}^{n} (\alpha_i^- + \alpha_i^+) - \\ \frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i^- - \alpha_i^+)(\alpha_j^- - \alpha_j^+) K(x_i \cdot x_j) \\ \text{sujeto a} \quad \sum_{i=1}^{n} (\alpha_i^+ - \alpha_i^-) = 0 \\ 0 \leq \alpha_i^+, \alpha_i^- \leq C \quad \text{para} \quad \in \{1, 2, \ldots, n\} \end{aligned} \tag{5.49}$$

Existen varios métodos numéricos que pueden ser usados para atacar este tipo de problemas de optimización con restricciones, por ejemplo, el método del gradiente descendente que puede ser aplicado para resolver el problema primal, el método del gradiente descendente proyectado que se puede usar para resolver el problema dual. Existen paquetes de softwere pensado para resolver problemas de clasificación y regresión mediante el uso de MSV, nosotros estaremos usando el paquete LIBSVM, ver [14].

Capítulo 6

Resultados

Durante este capítulo se presentarán varias aplicaciones de los métodos mencionados a lo largo de este trabajo. En primer lugar veremos un ejemplo de clasificación de imágenes, donde se utilizan RNAs y MSV. Posteriormente se hará el análisis y predicción de un conjunto de series de tiempo en economía, donde se hará uso de los diferentes métodos vistos para analizar series temporales y se realizarán predicciones tanto con los modelos ARIMA como con los modelos de inteligencia artificial: RNAs y MSV. Finalmente, al igual que en las series de economía, se hará el mismo análisis para la serie de tiempo del número de infectados por covid-19 en Michoacán.

Para el caso del problema de clasificación, la CDP es medida calculando el porcentaje de aciertos que tuvo clasificando la RNA o la MSV en el conjunto de prueba. En el caso de las series de tiempo nos interesa poder predecir la dirección del movimiento de estas, es decir, se quiere predecir si la serie de tiempo incrementará o disminuirá su valor con respecto al actual, por lo tanto, mediremos la CDP calculando el porcentaje de aciertos de las predicciones hechas para la dirección del movimiento en el conjunto de prueba.

Para el caso de los modelos ARIMA y de RNAs, una vez fijado el modelo y estimados los parámetros, se realizará la predicción de la serie de tiempo en el conjunto de prueba. Ya hecha la predicción, se realizará una diferenciación tanto en la serie de tiempo original como en la que se predijo para comparar el signo del movimiento de la serie, este será positivo si incrementa, negativo si decrece y cero si es igual. En el caso de las MSV se diferenciará la serie previamente y se intentará clasificar el movimiento, asignando +1 si crece, -1 si decrece y 0 si es igual. Para hacer uso de las MSV se empleará la librería LBSVM con un kernel Gaussiano.

Para abordar los problemas de este trabajo se implementó en Fortran 90 una red neuronal de alimentación hacia adelante perceptrón de tres capas, donde el número de neuronas en la capa oculta se estará probando con 5, 20, 35, 50, 65, 80, 95 y 110 neuronas para ver qué modelo da un mejor desempeño y tanto en la capa de entrada como en la capa de salida el número va a variar dependiendo del problema que se esté resolviendo. El tamaño de paso que se usará para el descenso de gradiente será de 0.5 y los pesos se inicializarán

aleatoriamente con una distribción uniforme en [-1,1]. La función de activación para pasar de capa a capa será la sigmoide dada en la ecuación (4.3), como esta función tiene una rango en el intervalo (0,1) en los problemas de series de tiempo será necesario trasladar y normalizar los datos. Finalmente, todos los entrenamientos de RNAs serán de 200 iteraciones para poder realizar comparaciones. Entre las características a evaluar de los resultados de cada modelo serán: la CDP, el TDE y en caso de los modelos ARIMA y RNAs el EE.

6.1. Clasificación de números

Un ejemplo de problema de clasificación en imagenes es el siguiente. Se tiene un conjunto de imagenes en escala de grises, cada imagen contiene un dígito del 0 al 9 que fue escrito por una persona y se quiere clasificar a cada imagen en un grupo según el número que fue escrito en la imagen. Para este ejemplo se tienen 5000 imágenes de 20×20 pixeles que pueden ser encontradas en la base de datos MNIST en http://yann.lecun.com/exdb/mnist/. Cada imagen se puede transformar en un vector de 400 entradas en $\mathbb{R}$. En la Figura 6.1 podemos ver algunos ejemplos de las imágenes con las que esteremos trabajando, en la parte inferior de cada imagen se muestra la etiqueta en la que se clasifica o bien el número que la persona quiso dibujar.

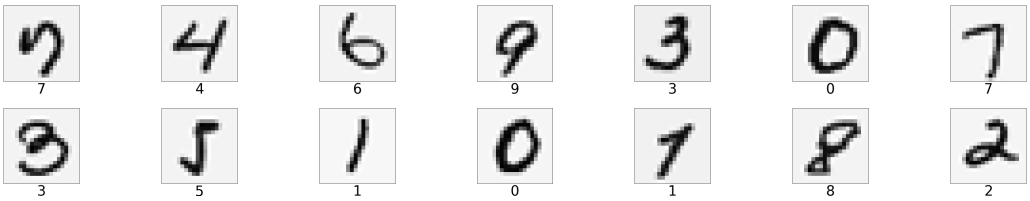

Figura 6.1: En esta figura vemos un conjunto de los números con los que fue entrenada la RNA y la MSV para el problema de clasificación de imagenes de números. En la parte inferior de cada imagen se muestra el número que representa cada imagen o bien la clase a la que corresponde. Cabe mencionar que para mostrar estas imágenes, el color fue invertido de las originales.

En este problema la capa de entrada recibe un vector de $\mathbb{R}^{400}$ que corresponde a una imagen, dicho de otra forma, esta RNA tiene 400 neuronas en la capa de entrada y 10 neuronas en la capa de salida, donde se espera que si el valor de la n-ésima neurona es cercano a uno y en el resto de las neuronas el valor es cercano a cero, entonces la imagen corresponde al dígito n (en el caso de la 10-ésima neurona, el dígito que le corresponde es el cero). Para entrenar las redes neuronales se usaron 3000 (60 %) de los datos del conjunto total como conjunto de entrenamiento, 1000 datos (20 %) como conjunto de validación y 1000 datos (20 %) para calcular la CDP de la RNA. Cada modelo de RNA se ejecutó 9 veces para poder obtener promedios y distribuciones los resultados, recordando que los modelos

de RNAs los inicializamos aleatoriamente y no siempre se obtiene el mismo resultado para el mismo modelo, pero sí se distribuyen alrededor de algunos valores.

En la Figura 6.2 podemos ver las distribuciones de la frecuencia de la CDP para cada modelo, variando la cantidad de neuronas en la capa oculta, de aquí podemos ver que el modelo con mejor desempeño en cuanto a la CDP fue el modelo con 95 neuronas en la capa oculta. Este modelo tuvo en CDP, un mínimo del 93.2 %, un máximo del 94.2 % y una media del 93.7 %.

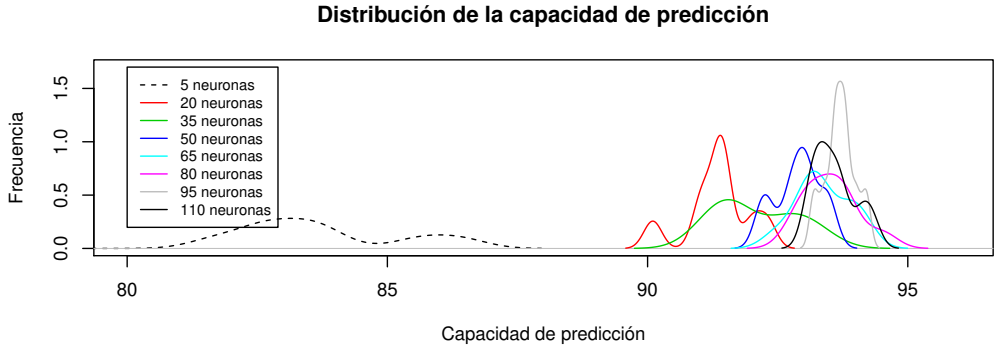

Figura 6.2: En esta gráfica podemos ver las distribuciones de frecuencia de la CDP que arrojaron los modelos de RNAs para el problema de clasificación de imágenes de números, variando el número de neuronas en la capa oculta. En el recuadro de la la esquina superior izquierda se muestra el número de neuronas del modelo con el que se generó cada distribución.

En la Figura 6.3 se puede ver como el tiempo de entrenamiento crece linealmente al incrementar el número de neuronas, mientras que el error de entrenamiento se mantiene en apariencia constante para un número de neuronas lo suficientemente grande en la capa oculta, como se ve en la Figura 6.4. Otra cosa que podemos observar de la resolución de este problema es que el número de iteraciones en validación mínima incrementa conforme se aumenta el número de neuronas en la capa oculta (ver Figura 6.5).

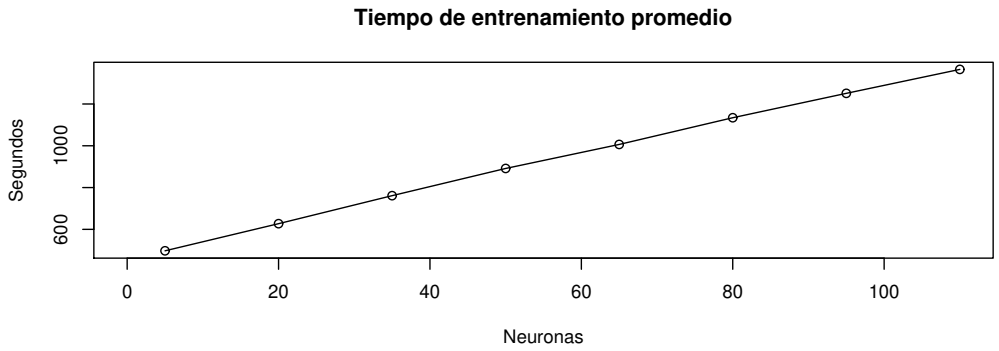

Figura 6.3: Esta gráfica muestra el tiempo de entrenamiento en segundos durante 200 iteraciones para RNAs variando el número de neuronas en la capa ocuta de cada RNA.

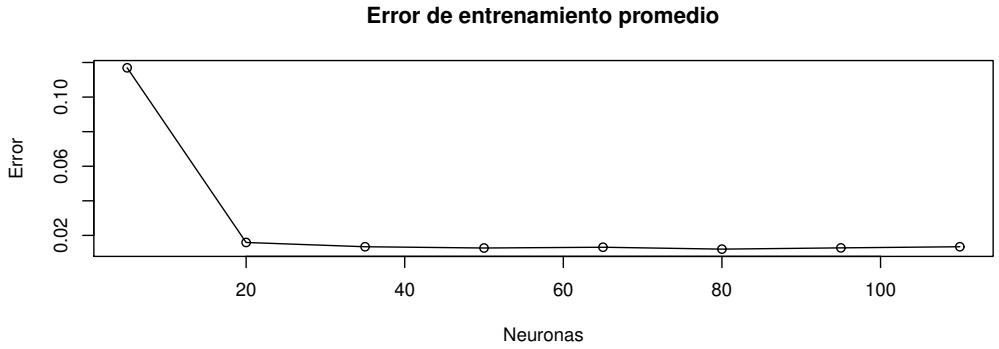

Figura 6.4: En este gráfico podemos ver como el error de entrenamiento es mayor cuando se tienen muy pocas neuronas, en este caso se puede ver que con 5 neuronas se tiene mucho error y a partir de 20 neuronas el error menor.

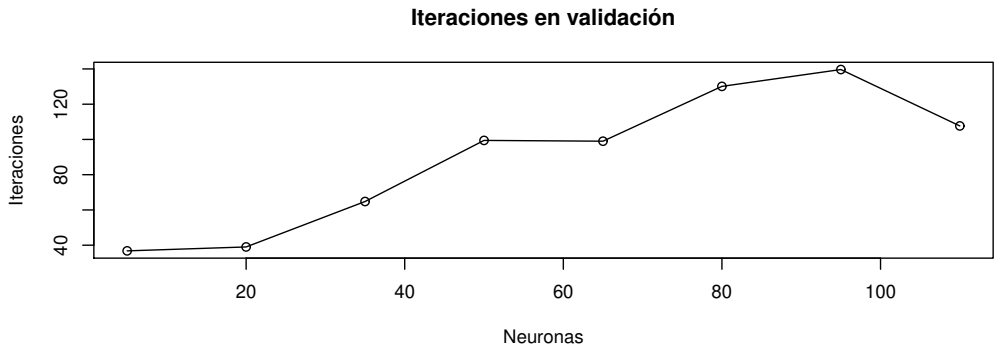

Figura 6.5: Aquí podemos ver la gráfica del número de iteraciones promedio que se realizaron para cada modelo cuando el error de validación fue mínimo.

En el Cuadro 6.1 se muestran los resultados obtenidos con el algoritmo de MSV y los promedios de la RNA con 95 neuronas en la capa oculta, donde TDE es el tiempo de entrenamiento en las 200 iteraciones para el caso de la RNA.

Modelo	CDP	TDE
MSV	95.0 %	1 hora y 19 minutos
RNA	93.7 %	21 minutos

Cuadro 6.1: En esta tabla se muestran los resultados del modelo de la MSV y el de la RNA con 95 neuronas en la capa oculta.

6.2. Series en economía

Para el caso de las predicciones de series de tiempo, las RNAs solo tendrán una neurona de salida, en este caso, como mencionamos antes, se trasladará la serie de tiempo restando el valor mínimo de ésta y se normalizará dividiendo entre el valor máximo de la serie, de este modo se tendrá una serie normalizada con rango en [0,1]. Además el número de neuronas de la capa de entrada dependerá de la serie que se esté analizando y de sus valores mayormente correlacionados. Dado que los pesos fueron inicializados aleatoriaente, cada modelo se entrenó 30 veces para poder obtener una distribución de la CDP de los modelos de RNAs y de esta forma poder seleccionar el modelo que genere mejores resultados. Una vez seleccionado el modelo, en la tabla de resultados se pondrá lo obtenido en un entrenamiento en particular.

En esta sección aplicaremos los conocimientos estudiados a lo largo de los capítulos a un conjunto de series de tiempo de economía, estas series son la de la **Tasa de Cotización**

(TC), la del **índice nacional de precios del considor (INPC)**, la del **índice de precios del considor (CPI)**, la de la **inflación mexicana** y la de la **inflación de USA**, las cuales fueron documentadas del año 1986 al 2014 cada día primero del mes, con un total de 348 valores cada una. Estas series pueden ser encontradas en https://www.banxico.org.mx/. Para realizar el análisis y hacer el ajuste del modelo estudiaremos solo el 70 % de los datos y con el 30 % restante definiremos el criterio de paro (15 %) y calcularemos la CDP del modelo (15 %).

6.2.1. Serie de tiempo de la TC

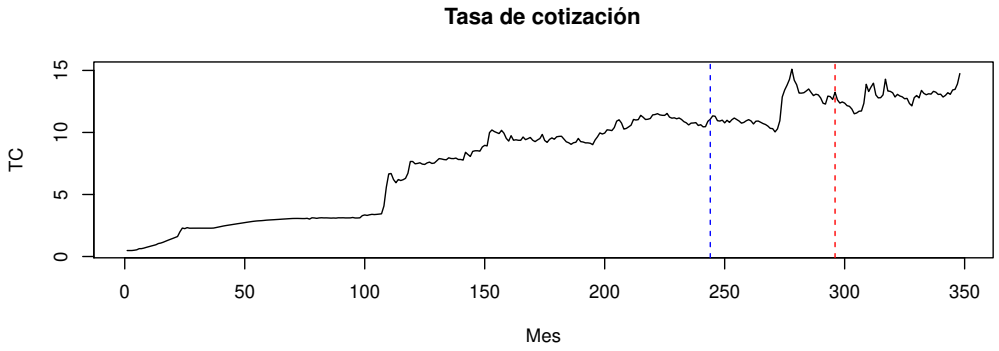

Figura 6.6: Gráfica de la serie de tiempo TC, las lineas verticales azul y roja dividen la serie en tres conjuntos, el de entrenamiento, el de validación y prueba de izquierda a derecha.

Análisis de la serie

En la Figura 6.6 se puede ver la serie de tiempo TC, donde se observa a simple vista que la serie TC tiene una tendencia o crecimento y no presenta ciclos aparentes. Revisando la ACF y la PACF que se muestran en la parte izquierda de la Figura 6.7, podemos pensar que la tendencia podría ser bien aproximada linealente, pues una ACF que decrece lentamente con el tiempo indica una alta dependencia lineal en los datos, además una PACF que se hace casi cero después del Lag 1 indica un modelo AR(1).

Dado que ni la ACF de TC ni su gráfica muestran periodos en la serie de tiempo TC, pero sí una tendencia, podemos hacer uso de alguno de los métodos del Caso 1 expuestos en la sección 2.7 para remover tendencias y después ajustar algún modelo ARMA(p,q). El método que utilizaremos para volver la serie TC estacionaria es el de **diferenciar** la serie. Una vez diferenciada la serie TC, su ACF y su PACF se puden ver en la parte derecha de la Figura 6.7, estas indican que la serie de tiempo TC diferenciada puede satisfacer un modelo MA(9).

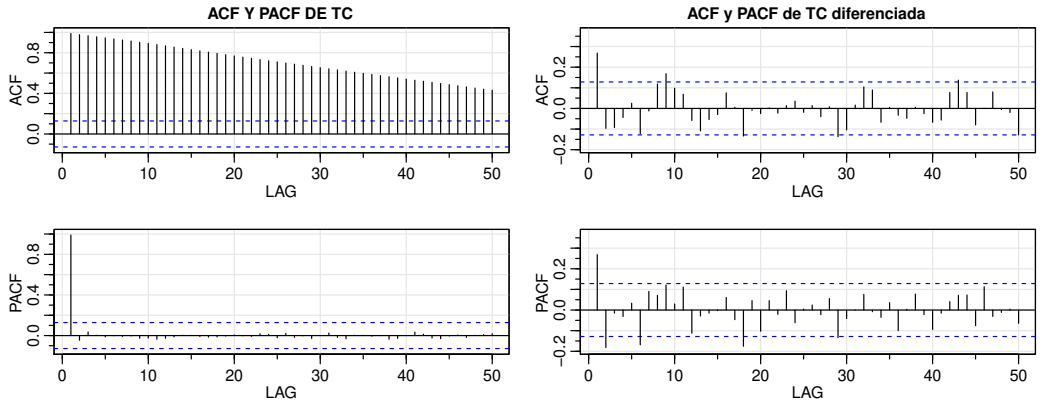

Figura 6.7: En la parte izquierda de esta figura se muestra la ACF y la PACF de TC, éstas indican una alta dependencia lineal en los datos. En la parte derecha se puede ver la ACF y la PACF de la serie TC diferenciada a un dato o Lag de distancia.

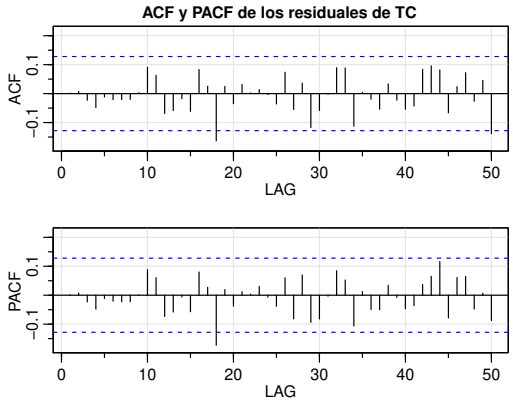

Figura 6.8: ACF y PACF de los residuales de TC después de ajustar un modelo ARIMA(0,1,9)

Entrenamiento de la RNA

Para encontrar el modelo de RNA adecuado de la serie de tiempo TC, se hicieron pruebas entrenando la RNAs con los datos que se encontraban a las distancias más correlacionados del valor que se quería predecir. En la Figura 6.9 podemos observar las distribuciones de frecuencia de la CDP, se puede ver claramente que las RNAs que arrojaron las mejores CDP fueron las que se entrenaron con los datos a una distancia hacia atrás 3, 8 y 9.

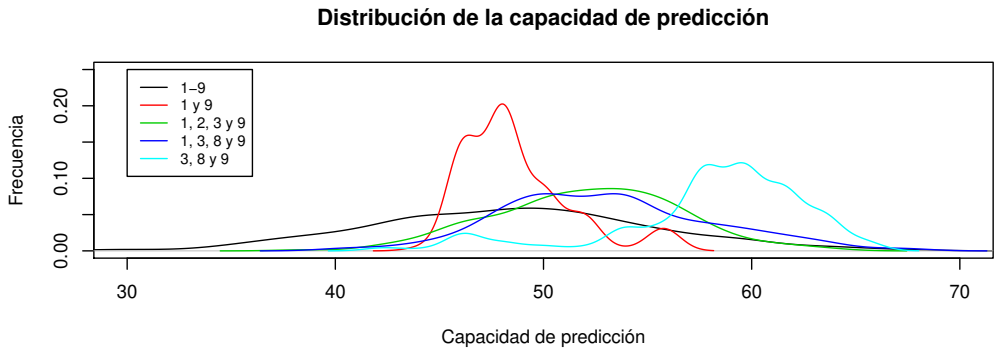

Figura 6.9: En esta figura se muestran las distribuciones de la frecuencia de la CDP. En el recuadro de de la esquina superior izquierda se muestran los datos con los que fueron entrenadas las RNAs.

En la Figura 6.10 se pueden revisar las distribuciones de frecuencia de la CDP para las RNAs entrenadas con los valores a distancias 3, 8 y 9 en la capa de entrada, en este caso los mejores resultados de CDP los logró la RNA con 5 neuronas en la capa oculta, obteniendo un mínimo del 57.69 %, una media del 61.15 % y un máximo del 65.38 % de CDP.

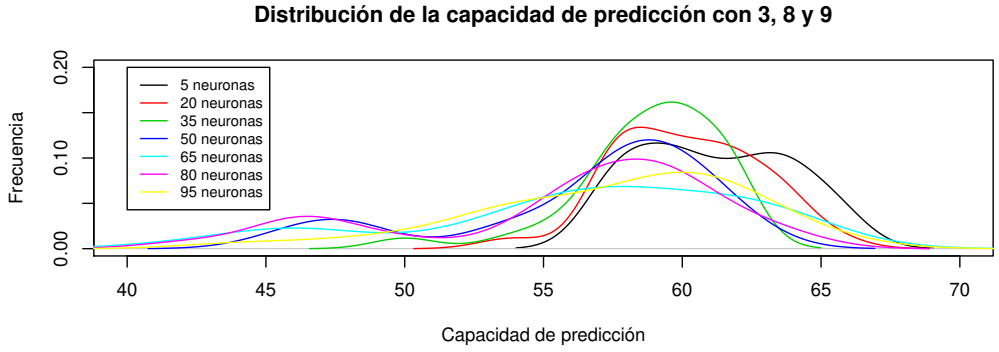

Figura 6.10: Una vez fijados los datos de entrada 3, 8 y 9, en esta gráfica vemos cómo se comportan las distribiciones de la frecuencia de la CDP variando el número de neuronas en la capa oculta.

Entrenamiento de la MSV

Se hicieron pruebas entrenando MSV con la misma estructura de los datos de entrada que las RNAs, en este caso el mejor resultado la obtuvo la MSV entrenada con los valores hacia atrás 1-9 y 18, los resultados se pueden consultar en el Cuadro 6.2.

Resultados

En el Cuadro 6.2 podemos ver los resultados agrupados de los modelos ARIMA(0,1,9), RNA usando los valores hacia atrás 3, 8 y 9 con 5 neuronas en la capa oculta y la MSV entrenada con los valores hacia atrás 1-9 y 18. En la Figura 6.11 se pueden observar la serie TC y sus predicciones en la región de prueba.

Modelo	CDP	TDE	EE
ARIMA(0,1,9)	53.84 %	1.275 s	2.17
RNA	61.53 %	1.78 s	2.28
MSV	55.1 %	15.25 s	-

Cuadro 6.2: Aquí se muestran los resultados de los mejores modelos seleccionados para la serie TC.

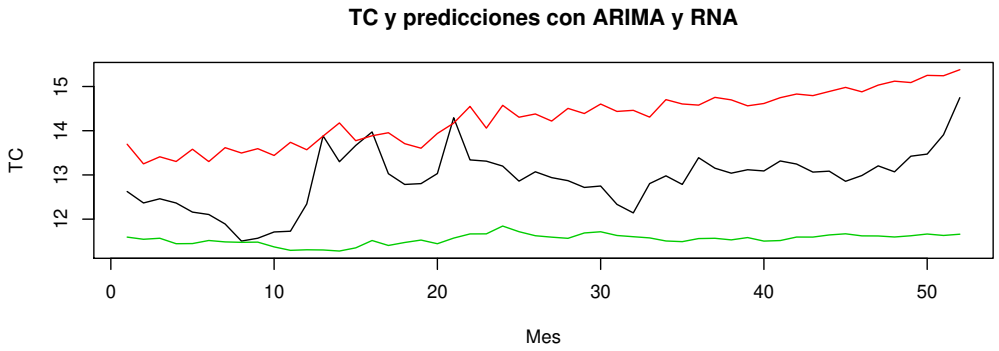

Figura 6.11: En esta gráfica se pueden observar el conjunto de prueba de la serie de tiempo TC en color negro, mientras que en color verde podemos ver la predicción del modelo de la RNA que usa los valores 3, 8 y 9 con 5 neuronas en la capa oculta y la predicción con el modelo de ARIMA(0,1,9) en color rojo.

6.2.2. Serie de tiempo del INPC

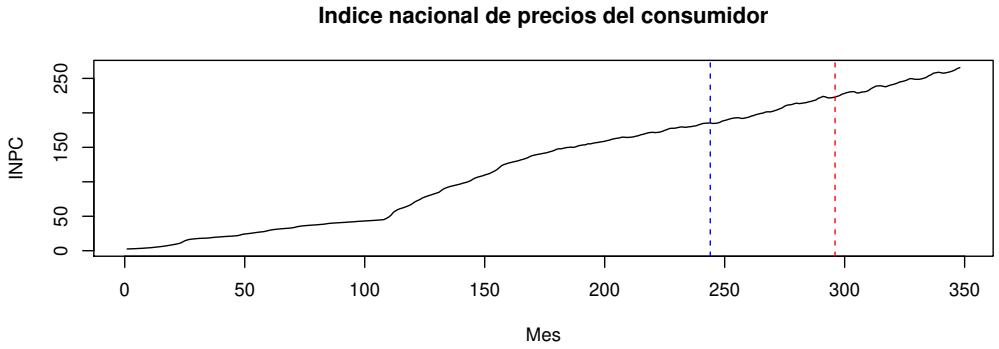

Figura 6.12: Esta gráfica muestra la serie de tiempo del INPC, dividida por las lineas verticales en el conjunto de entrenamiento, el conjuto de validación y el de prueba de izquierda a derecha respectivamente.

Análisis de la serie

En la Figura 6.12 tenemos la serie de tiempo del INPC, ésta muestra claramente una tendencia creciente lineal que podemos confirmar revisando la ACF y la PACF que se encuentran en la parte izquierda de la Figura 6.13. Para retirar la tendencia lineal existente en la serie de tiempo del INPC se realizó una diferenciación. Una vez sin tendencia la ACF dejó ver un comportamiento periódico de 12 meses, esto se puede revisar en la parte superior derecha de la Figura 6.13.

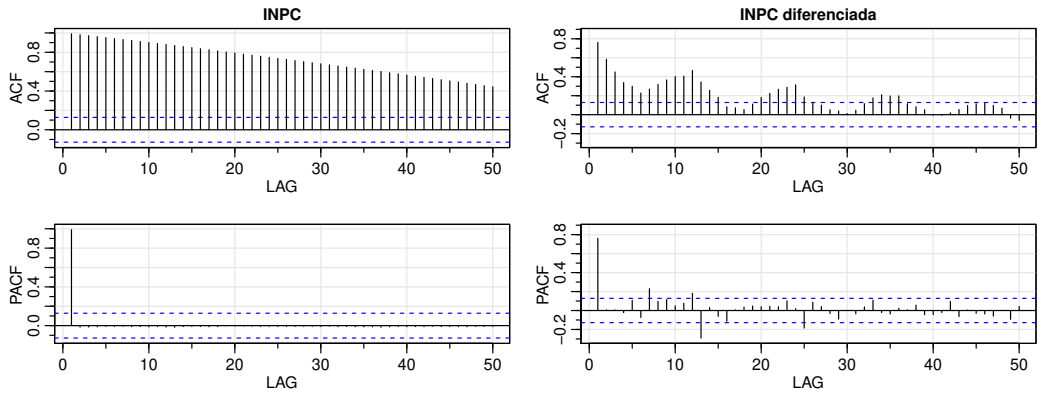

Figura 6.13: En la parte izquiereda de la figura podemos ver las graficas de la ACF y la PACF de la serie INPC, las cuales muestran claramente una dependencia lineal en los datos. En la parte derecha se muestran la ACF y la PACF de la serie INPC diferenciada a un dato de distancia, estas muestran la presencia de ciclos en la serie.

Para eliminar los periodos de la serie INPC sin tendencia, diferenciamos nuevamente, pero a una distancia de 12, este es el *Método S2* que mencionamos en el Capítulo 2. Una vez retirada la periodicidad, revisamos nuevamente la ACF y la PACF que se ilustran en la parte izquierda de la Figura 6.14, las cuales sugieren un comportamiento de los datos según un modelo AR(1), la ACF y la PACF de los residuos de este modelo los podemos ver en la parte derecha de la Figura 6.14.

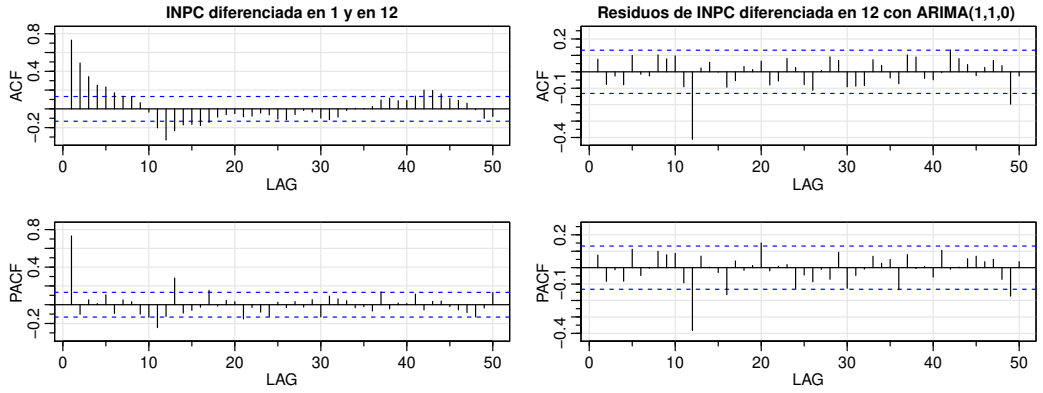

Figura 6.14: En la parte izquierda se muestran la ACF y la PACF de la serie INCP sin tendencia ni periodos. En la parte derecha se puede observar la ACF y la PACF de los residuos de ajustar un modelos ARIMA(1,1,0) de periodo 12 a la serie INPC.

Como en la ACF de la derecha de la Figura 6.14 se ve una alta correlación de los datos a una distancia 12, se sugiere que el modelo de la serie INPC sea un ARIMA(1,1,12) con periodo 12.

Entrenamiento de la RNA

Considerando el hecho de que la serie de tiempo INPC tiene periodo 12 y tendencia lineal, podríamos pensar que para hacer predicciones sobre esta serie sería suficiente con introducir en la capa de entrada los 12 valores que anteceden al valor que queremos predecir, sin embargo, este modelo, dado el número tan alto de neuronas en la capa de entrada, no genera buenos resultados de predicción, por lo tanto se entrenaron otros modelos con menos cantidad de neronas.

En la Figura 6.15 podemos ver las distribuciones de la frecuencia de la CDP, dependiendo de los valores que fueron introducidos en la capa de entrada. Se experimentó con 7 modelos diferentes para la capa de entrada, en uno usa los 12 valores hacia atrás del valor que se quiere predecir (12 neuronas), en los otros solo se introducen algunos de los datos que se encuentran a una distancia h=12,11,2 o 1 con 2,3 o 4 neronas, se pueden revisar mejor los resultados en la Figura 6.15.

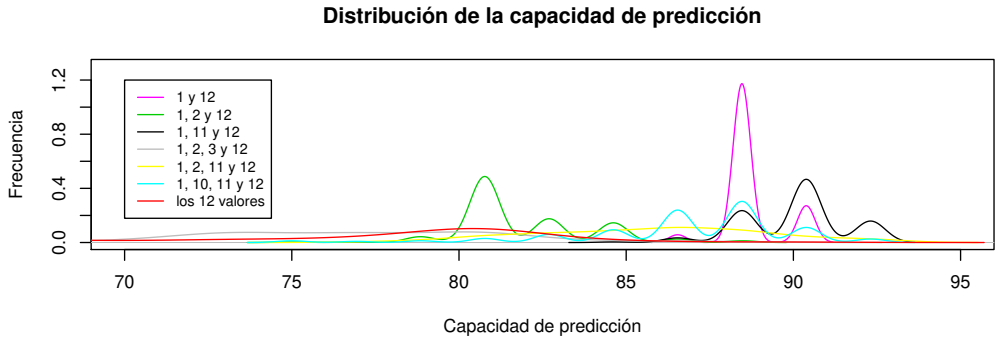

Figura 6.15: Aquí se pueden obsevar las distribuciones de cada modelo variando los valores de la capa de entrada. Los números del recuadro indican los valores hacia atrás del valor que se desea predecir y que fueron considerados para entrenar la RNA.

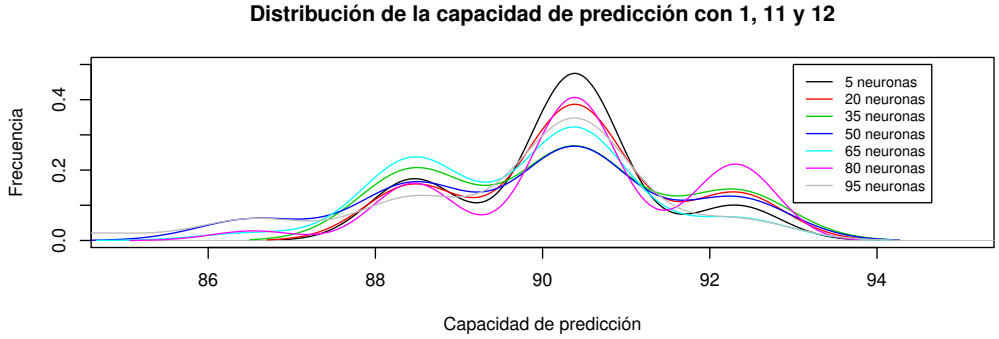

Figura 6.16: Aquí se pueden visualizar las gráficas de las distribuciones de la frecuencia de la CDP de las RNAs entrenada con los valores hacia atras 1, 11 y 12, variando el número de neuronas en la capa oculta.

Como pudimos ver en la Figura 6.15, el modelo con 12 neuronas fue de los que peores resultados de predicción dio, mientras que el modelo de 3 neuronas que consideró los valores de distancia hacia atras h=1,11 y 12 fue el que mejor desempeño tuvo. En la Figura 6.16 se muestran las distribuciones de la CDP para el modelos de 3 neronas 1, 11 y 12, pero variando las neuronas de la capa oculta. El mejor modelo obtenido fue el de 80 neuronas en la capa oculta, logrando un mínimo del 86.53 %, una media del 90.44 % y un máximo del 92.30 % de CDP.

Entrenamiento de la MSV

Como en el caso de la serie de tiempo INPC el 80.39 % de los datos eran crecientes respecto al valor anterior, entonces no hubo suficientes datos para hacer una clasificación

correcta con una MSV, ésta fue entrenada con la misma estrctura de los datos que las RNA's y en cada caso dio resultados menores o iguales al 80.39 %, es decir, en el mejor de los casos clasificó a todos los valores de predicción como crecientes.

Resultados

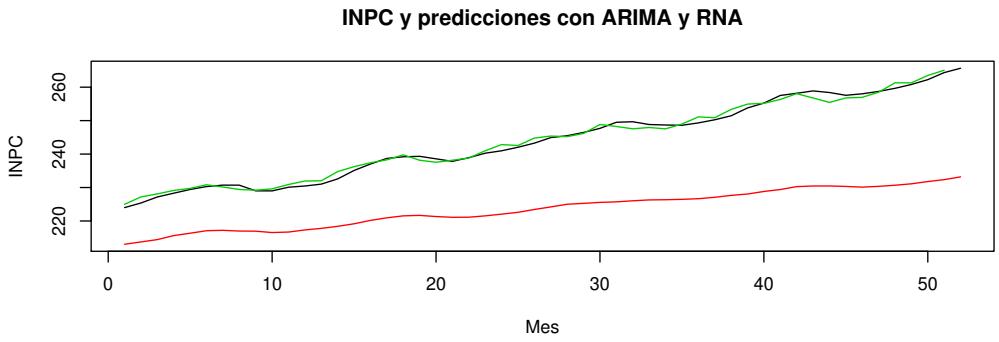

Figura 6.17: Aquí se pueden visualizar las gráficas de las predicciones de la serie de tiempo INPC en la región de prueba, en color negro se muestra la serie INPC. de rojo se ven las predicciones hechas con el modelo de RNA con 1, 11 y 12 neuronas en la capa de entrada y 80 en la capa oculta. De color verde se pueden ver las predicciones del modelo ARIMA(1,1,12).

Modelo	CDP	TDE	EE
ARIMA(1,1,12)	88.46 %	5.382 s	0.893
RNA	90.38 %	4.297 s	457.1
MSV	80.39 %	10.85 s	-

Cuadro 6.3: Aquí se muestran los resultados de los mejores modelos seleccionados para la serie INPC

Revisando los resultados del Cuadro 6.3 y la Figura 6.17 de la comparación de la serie INPC con las predicciones hechas en la región de prueba, podemos darnos cuenta que aunque la RNA resulto mejor que el modelo ARIMA en cuanto a la CDP en aproximadamente un 2 %, el error empírico de la RNA se elevó en gran medida, es decir, aunque la red aprendió bien el movimiento de la serie, no aprendió bien la tendencia.

6.2.3. Serie de tiepo del CPI

Análisis de la serie

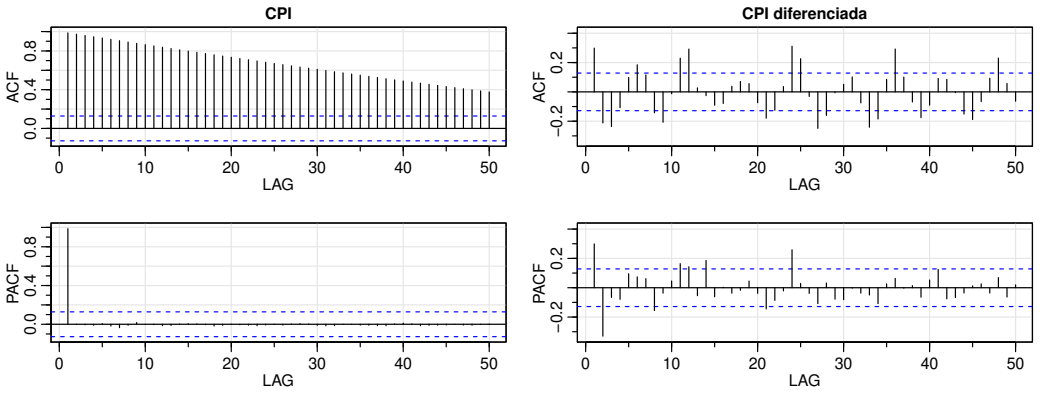

Figura 6.19: En la parte izquierda de la figura podemos ver las graficas de la ACF y la PACF de la serie CPI, estas muestran una dependencia lineal en los datos. En la parte derecha se muestran la ACF y la PACF de la serie CPI diferenciada a un dato de distancia, estas muestran la presencia de ciclos con periodo 12 en la serie.

En la Figura 6.18 se muestra la gráfica de la serie de tiempo CPI, esta aunque parece ser una serie sencilla a simple vista, es decir, con una tendencia claramente lineal, sin ciclos aparentes y sin tanto ruido, veremos que al analizarla presenta algunas complicaciones. En la Figura 6.19 podemos ver a la izquierda la ACF y la PACF de la serie CPI, las cuales indican una clara dependencia lineal de los datos, una vez retirando la tendencia por el método de diferenciación, vemos que la ACF de la parte derecha de la Figura 6.19 indica la presencia de ciclos de preriodo 12.

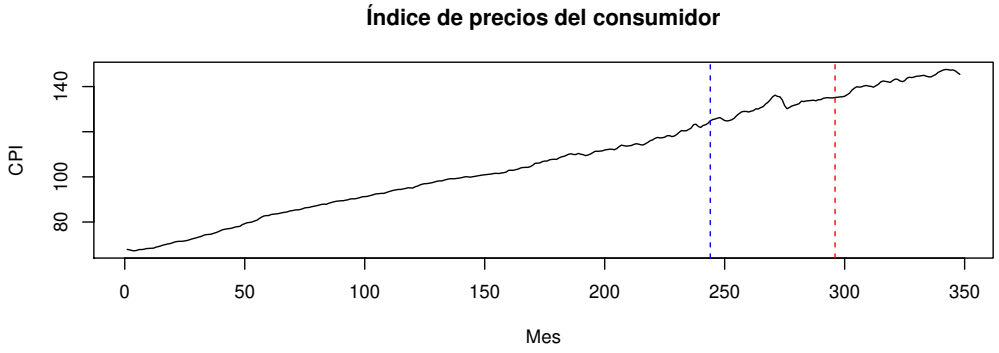

Figura 6.18: Esta gráfica muestra la serie de tiempo del CPI, dividida por las lineas verticales punteadas de izquierda a derecha en el conjunto de entrenamiento, el conjuto de validación y el de prueba.

Una vez eliminada la periodicidad y la tendencia de la serie CPI, podemos ver su ACF y su PACF en la parte izquierda de la Figura 6.20, la PACF indica un coportamiento AR(12).

En parte derecha la Figura 6.20 se pueden ver la ACF y la PACF de los residuos de la serie CPI después de ajustar un modelo ARIMA(12,1,0) con periodo 12, estos residuos se comportan como ruido blanco.

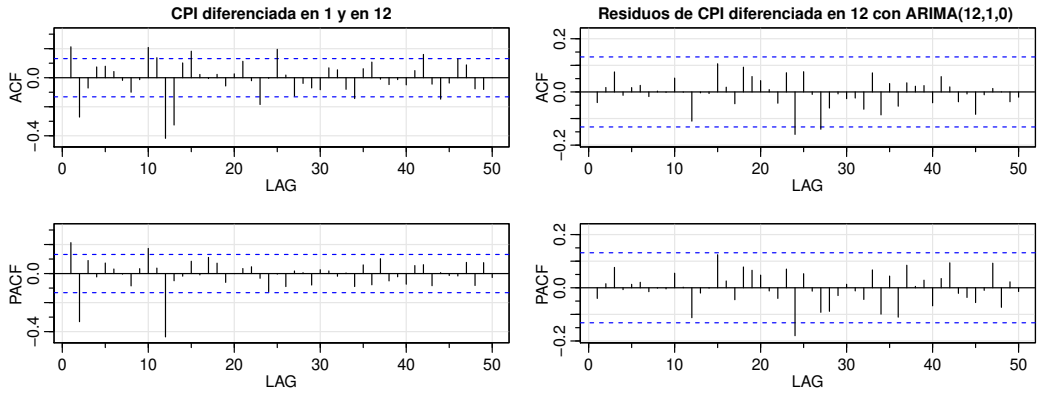

Figura 6.20: En la parte izquierda de la figura podemos ver las graficas de la ACF y la PACF de la serie CPI sin tendencia ni ciclos, estas sugieren un comportamiento de los datos según un modelo AR(12). En la parte derecha se muestran la ACF y la PACF de los residuos de la serie CPI después de ajustar un modelo ARIMA(12,1,0) con periodo 12.

Dado que la tendencia no fue bien modelada por el modelo anterior (ver Figura 6.24), se probó estimando la tendencia de CPI con una función lineal y retirando los ciclos con el método de diferenciación, en la parte izquierda de la Figura 6.21 podemos ver la ACF y la PACF de la serie CPI depués de retirar las componentes de esta forma, la ACF y la PACF sugieren un comportamiento AR(13). Después de ajustar el modelo AR(13) a la serie CPI sin tendencia lineal y sin periodos, la ACF y la PACF se pueden ver a la derecha de la Figura 6.21.

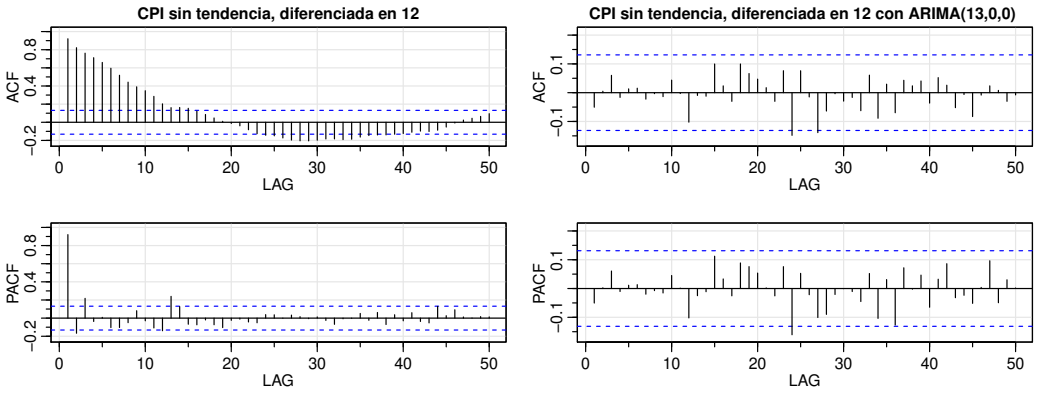

Figura 6.21: En la parte izquierda de esta figura se pueden ver la ACF y la PACF de la serie CPI sin tendencia lineal y sin periodos de longitud 12, mientras que en la parte derecha podemos ver la ACF y la PACF de los residuos de la serie CPI al ajustar un modelo AR(13) con tendencia lineal y periodo 12, estos residuos se comportan como ruido blanco.

Entrenamiento de la RNA

Al igual que en el caso de la serie de tiempo del INPC, la serie del CPI es una serie de tiempo con componente estacional de periodo 12 y tendencia lineal, por lo que se debería de considerar al menos el valor con distancia 12 y 1 hacia atrás del valor que se quiere predecir. Como podemos ver en la Figura 6.22 que muestra las distribuciones de la frecuencia de la CDP de algunos modelos de RNA's variando los datos de entrada, se puede ver que el modelo que usó los 12 valores hacia atrás fué el peor, mientras que el que tuvo mejor comportamiento fue el que usó los valores de distancia hacia atrás 1, 11 y 12. Cabe mencionar que ninguno de los modelos entrenados llegó al error mínimo de validación, por lo que podemos asegurar que fue sumamente difícil entrenar las RNA's para la serie CPI.

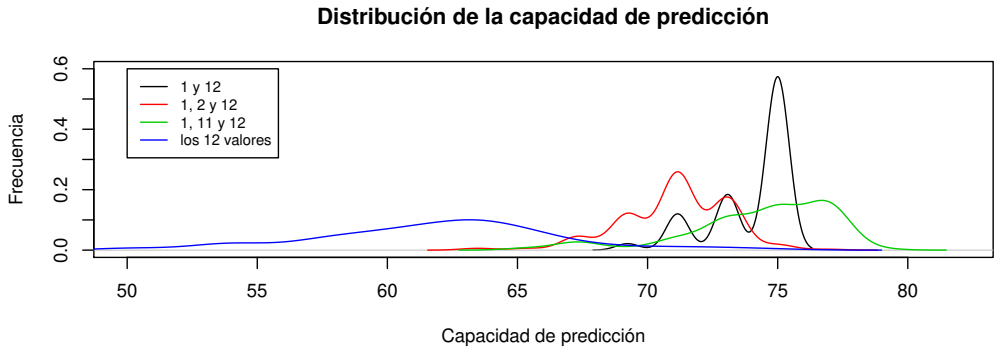

Figura 6.22: Aquí se pueden observar las distribciones de freccencia del CPI de los modelos de RNA's variando los datos de la capa de entrada.

En la Figura 6.23 podemos ver el comportamiento de las distribuciones de la frecuencia de la CDP de las RNA's con 3 neuronas en la capa de entrada que usan los valores de distancia hacia atrás 1, 11 y 12. Podemos ver que el comportamiento no varía en gran medida dependiendo del número de neuronas de la capa oculta, pero la mejor está vez podríamos decir que fue la de 20 neuronas en la capa oculta, logrando un mínimo del 67.30 %, una media del 74.81 % y un máximo del 76.92 % de CDP a pesar de haber estado mal entrenadas las RNAs.

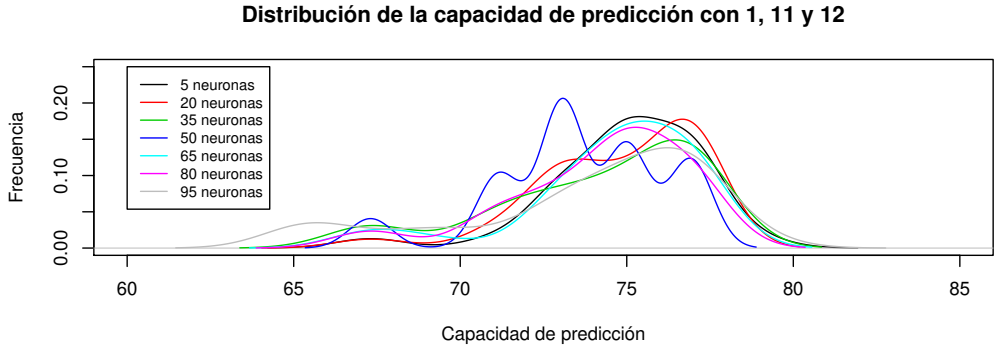

Figura 6.23: Aquí podemos observar las distribuciones de freccencia del CPI de los modelos de RNA's con tres neuronas en la capa oculta, usando los datos hacia atrás 1, 11 y 12, y variando las neuronas de la capa oculta.

Entrenamiento de la MSV

Para entrenar la MSV se utilizó en mismo arreglo de datos que en las RNAs sin obtener una buena clasificación, pues en el mejor de los casos todo los datos de predicción los clasificó como incrementos.

Resultados

En la Figura 6.24 se muestran las diferentes predicciones para la serie CPI con todos los modelos que se han mencionado y en el Cuadro 6.4 se muestran los resltados de los modelos, tomando en cuenta la CDP, el TDE y el EE.

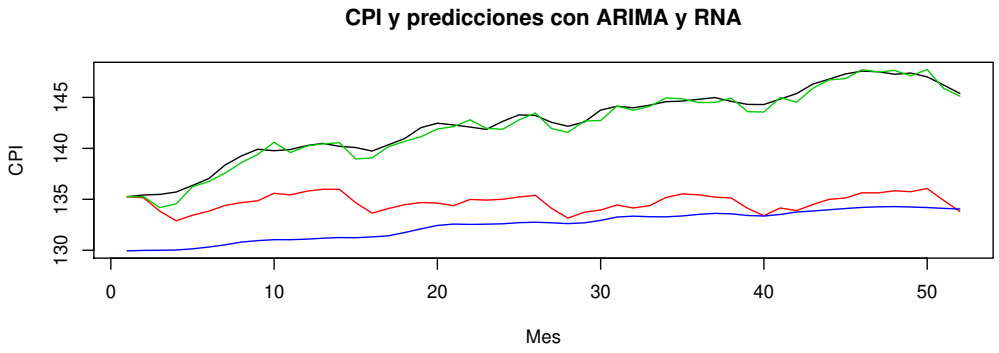

Figura 6.24: En esta gráfica se pueden visalizar las predicciones para la serie de tiempo CPI que se muestra de color negro. En color verde está la predicción del modelos ARIMA(13,0,0) con tendencia lineal y periodo 12. De color rojo está la predicción del modelo ARIMA(12,1,0) y periodo 12. Finalmente en color azul se puede ver la predicción de la RNA con los valores de entrada 1, 11 y 12 y 20 neuronas en la capa oculta.

Modelo	CDP	TDE	EE
ARIMA(12,1,0)	80.39 %	1.946 s	73.71
ARIMA(13,0,0)	69.23 %	3.250 s	0.295
RNA	75.00 %	2.569 s	104.86
MSV	62.74 %	72.07 s	-

Cuadro 6.4: Resultados de los diferente modelos ajustados a la serie CPI.

6.2.4. Serie de tiempo IMEX

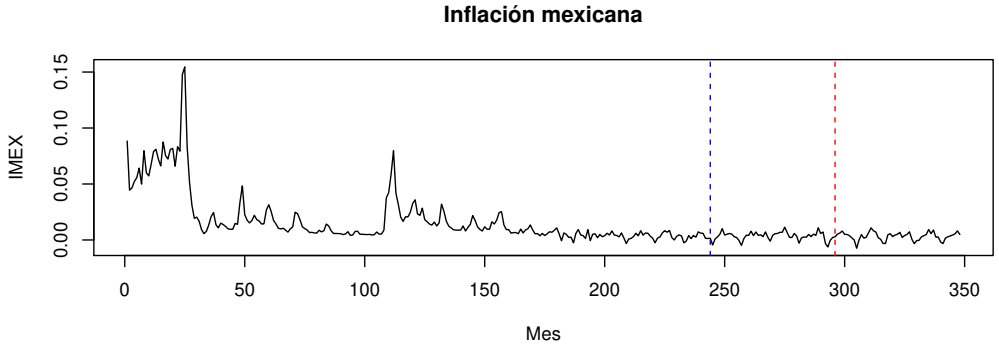

Figura 6.25: Esta gráfica muestra la serie de tiempo de la IMEX, dividida por las lineas punteadas de izquierda a derecha en el conjunto de entrenaiento, el conjuto de validación y el de prueba.

Análisis de la serie

En la Figura 6.25 se muestra la gráfica de la serie de tiempo IMEX, esta muestra una tendencia decreciente y comportamiento periódico a simple vista, revisando su ACF y PACF en la parte izquierda de la Figura 6.26 podemos verificar que esta tiene una dependencia lineal en los datos y además deja ver la existencia de periodos de longitud 12. Una vez retirada la tendencia por medio de diferenciación podemos ver más claramente el comportamiento periódico revisando la ACF y la PACF en la parte derecha de la Figura 6.26.

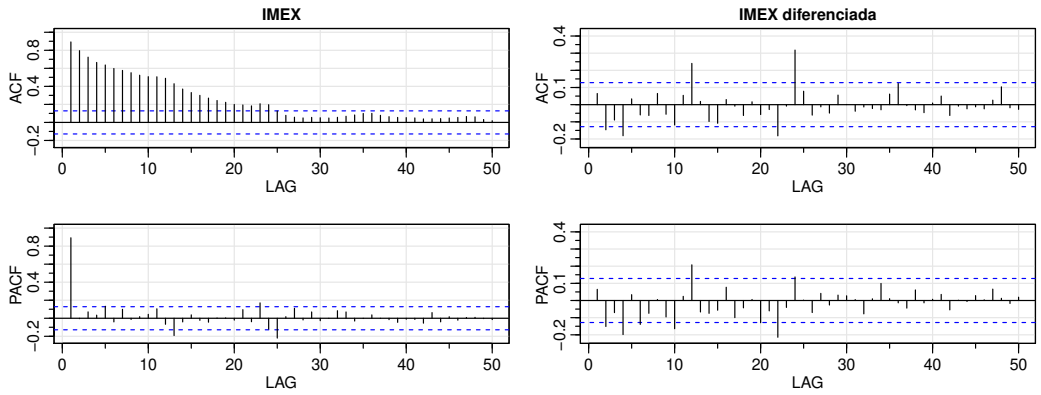

Figura 6.26: En la parte izquierda de la figura podemos ver las graficas de la ACF y la PACF de la serie IMEX, estas muestran una dependencia lineal en los datos y comportamiento periódico. En la parte derecha se muestran la ACF y la PACF de la serie IMEX diferenciada a un dato de distancia, éstas muestran la presencia de ciclos de periodo 12.

Para eliminar la periodicidad de la serie IMEX usamos el método de diferenciación a una distancia 12, una vez eliminada la periodicidad y la tendencia de la serie IMEX podemos ver su ACF y su PACF en la parte izquierda de la Figura 6.27, la PACF indica un comportamiento AR(12). En la parte izquierda de la Figura 6.27 podemos revisar la ACF y la PACF de los residuos de IMEX al ajustar un modelos ARIMA(12,1,0) de periodo 12 con los métodos que acaban de ser mencionados.

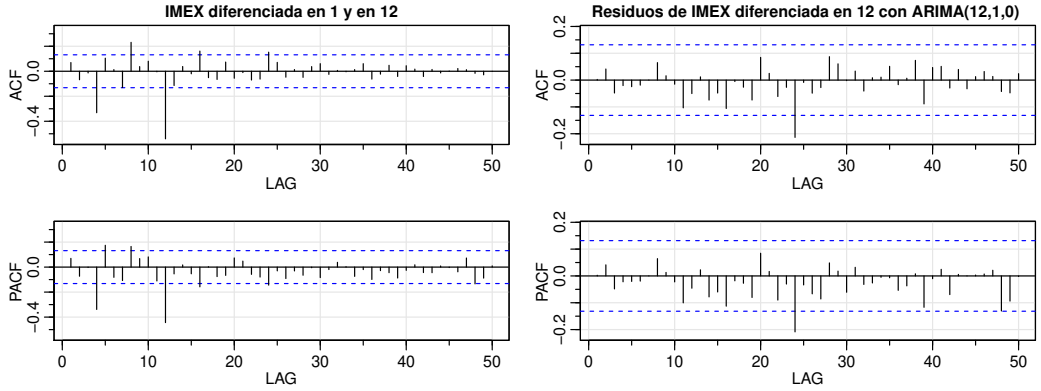

Figura 6.27: En la parte izquierda de la figura podemos ver las graficas de la ACF y la PACF de la serie IMEX sin tendencia ni ciclos, estas sugieren un comportamiento de los datos según un modelo AR(12). En la parte derecha se muestran la ACF y la PACF de los residuos de la serie IMEX después de ajustar un modelo ARIMA(12,1,0) con periodo 12, dado el número de correlaciones fuera de la región de confianza podemos decir que el comportamiento es de ruido blanco segun el criterio que hemos utilizado.

En la Figura 6.31 se muestra en color negro la serie IMEX en la región de predicción y en rojo las predicciones del modelos ARIMA(12,1,0) con periodo 12, en el que se usaron métodos de diferenciación, como podemos ver la tendencia de la serie no fue modelada del todo bien por lo que se trató de modelar la tendencia y los ciclos con otros métodos, aunque estos no resultaron más apropiados, mostraremos los resultados que se obtuvieron de quitar los ciclos por el método de suavizamiento y la tendencia por medio de diferenciación.

En la parte izquierda de la Figura 6.28 se muestran la ACF y la PACF de la serie IMEX después de ser diferenciada y eliminar los periodos con el método de suavizamiento, estas gráficas nos sugieren un comportamiento de los datos según un modelo MA(24) o AR(22). En la parte derecha de la Figura 6.28 podemos ver la ACF y la PACF de los residuos de ajustar un modelos ARIMA(0,1,24) con periodo 12 usando método de suavizamiento para la serie IMEX, estas gráficas muestran un comportamiento de ruido blanco.

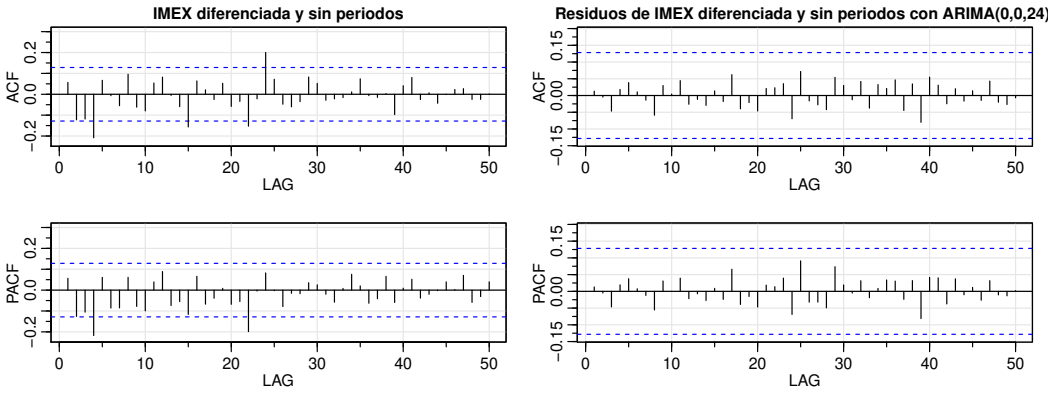

Figura 6.28: En la parte izquierda de la figura podemos ver las graficas de la ACF y la PACF de los residuos de la serie IMEX después de diferenciar y quitar los periodos con suavizamiento éstas indican un comportamiento MA(24) o AR(22). En la parte derecha se pueden ver la ACF y la PACF de los residuos de la serie IMEX al ajustar un modelo ARIMA(0,1,24) y periodo 12.

Entrenamiento de la RNA

Para el entrenamiento de las RNAs se hicieron experimentos utilizando los valores con mayor correlación que se observaron durante el análisis, como fueron los datos a una distancia 1, 8, 12, 22 y 24, también se probó usando los valores vecinos de estos datos como los datos a distancia 7, 11 y 23. En la Figura 6.29 se muestran las distribuciones de la frecuencia de la CDP de las RNAs variando los datos de entrada, como se puede ver, los mejores resultados los dieron las RNAs en las que la capa de entrada recibía los valores a distancia 11, 12, 22, 23 y 24.

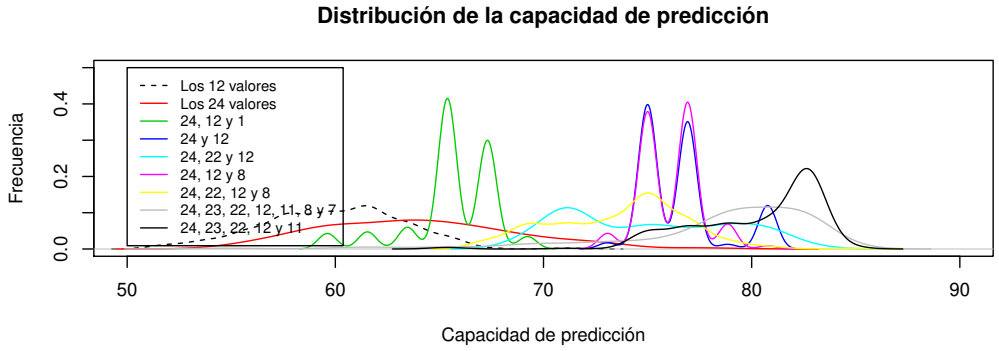

Figura 6.29: Aquí podemos observar las distribuciones de la freccencia de la CDP de la serie IMEX de los modelos de RNAs variando los datos de la capa de entrada. En el recuadro se indican las distancias de los valores con respecto al valor que se está prediciendo y que se utilizaron en la capa de entrada.

En la Figura 6.30 podemos ver el comportamiento de las distribuciones de la frecuencia de las RNAs con 5 neuronas en la capa de entrada que usan los valores de distancia hacia atrás 11, 12, 22, 23 y 24. Podemos ver que el comportamiento varía en gran medida para las redes con un número de neuronas en la capa oculta menor a 50. En este caso la RNA con mejor desempeño en la CDP fue la red con 80 neuronas en la capa oculta, logrando un mínimo del 78.8 %, una media del 82 % y un máximo del 82.6 % de CDP.

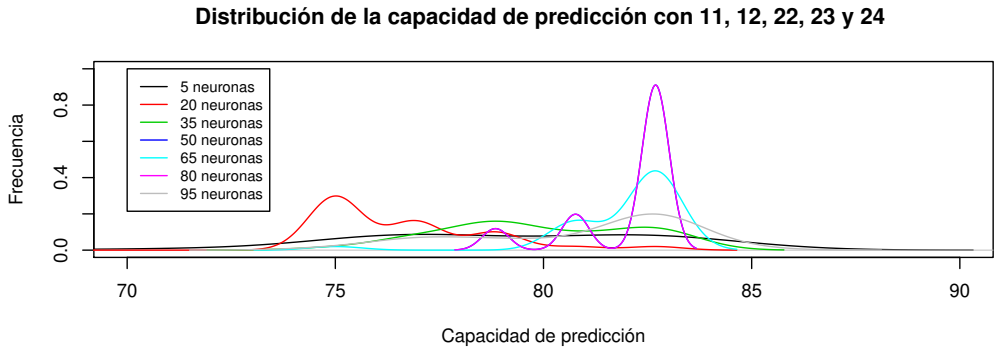

Figura 6.30: En esta figura se pueden ver las distribuciones de frecuencia de la CDP de las RNAs entrenadas con los valores 11, 12, 22, 23 y 24 y variando las neuronas de la capa oculta.

Entrenamiento de la MSV

Para entrenar la MSV se utilizó en mismo arreglo de datos que en las RNAs, pero a diferencia de las RNAs en esta se obtuvieron mejores resultados utilizando los 24 datos hacia atrás del valor que se quiere predecir y no solamente los de mayor correlación.

Resultados

En la Figura 6.31 podemos ver la serie IMEX en la región de prueba y las predicciones de los diferentes modelos con mejor CDP obtenida. En el Cuadro 6.5 se pueden ver los resultados de los mejores modelos implementados para la serie IMEX, en esta ocasión el modelo de MSV fue el que mejor desempeño logró en cuanto a CDP, esto se puede deber a que cerca de la mitad de los datos son crecientes y la otra mitad son decrecientes por lo que se pudo lograr una buena clasificación.

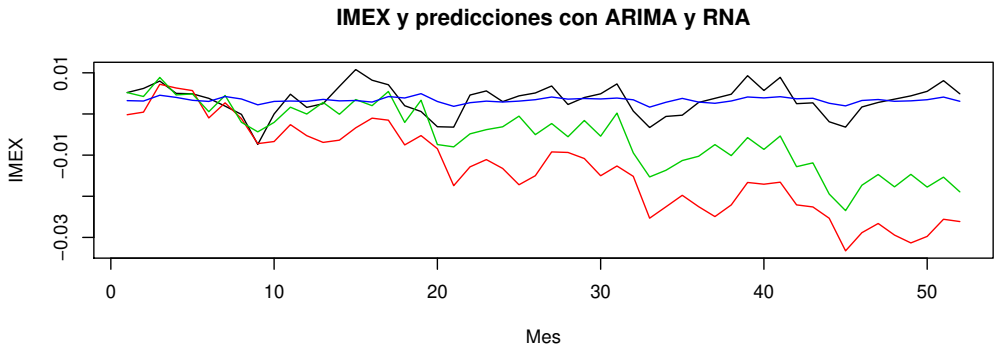

Figura 6.31: En esta gráfica se pueden visalizar las predicciones para la serie de tiempo IMEX que se muestra de color negro. En color rojo está la predicción del modelo ARIMA(12,1,0) con periodo 12 usando el método de diferenciación. De verde se puede ver la predicción hecha con el modelo ARIMA(0,1,24) y periodo 12 usando el método de suavizamiento. Finalmente en color azul está la predicción del modelo de RNA con 5 neuronas en la capa de entrada, usando los datos hacia atrás 11, 12, 22, 23 y 24 y 80 neuronas en la capa oculta.

Modelo	CDP	TDE	EE
ARIMA(0,1,24)	75.00 %	25.08 s	0.0003
ARIMA(12,1,0)	76.92 %	2.213 s	0.0001
RNA	82.00 %	5.151 s	1.112176e-05
MSV	83.33 %	15.39 s	-

Cuadro 6.5: Resultados de los diferente modelos ajustados a la serie IMEX.

6.2.5. Serie de tiempo IUSA

Análisis de la serie

En la Figura 6.32 se muestra la gráfica de la serie de tiempo IUSA, esta no parece tener tendencia y revisando su ACF y PACF en la parte izquierda de la Figura 6.33 podemos verificar que ésta solo muestra la existencia de periodos de longitud 12. Para retirar los periodos de la serie IUSA haremos uso del método de suavizamiento y del de diferenciación. En la parte derecha de la Figura 6.33 podemos ver la ACF y la PACF de los residuos de IUSA después de eliminar los periodos con el método de suavizamiento, estas sugieren un comportamiento para la serie IUSA según un modelo AR(2) con ciclos de periodo 12.

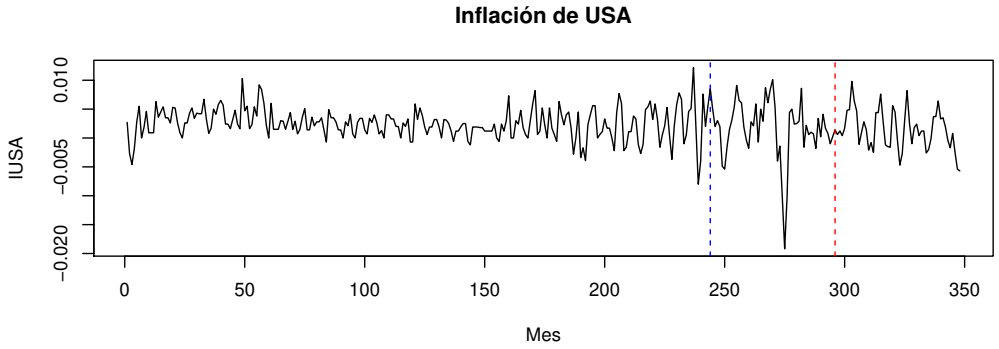

Figura 6.32: Esta gráfica muestra la serie de tiempo de la IUSA, dividida por las lineas verticales de izquierda a derecha en el conjunto de entrenaiento, el conjuto de validación y el de prueba.

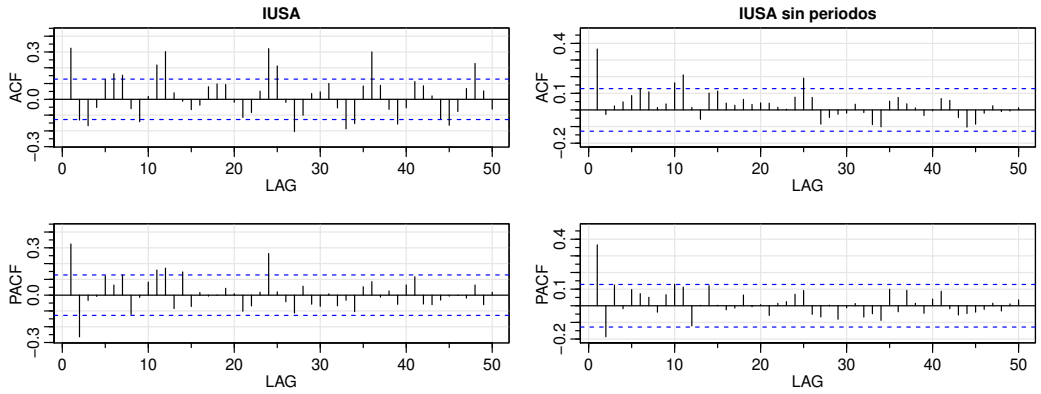

Figura 6.33: En la parte izquierda de la figura podemos ver las graficas de la ACF y la PACF de la serie IUSA, éstas muestran la presencia de ciclos de periodo 12. En la parte derecha se muestran la ACF y la PACF de la serie IUSA después de retirar los periodos con el método de suavizamiento.

En la Figura 6.34 podemos revisar la ACF y la PACF de los residuos de IUSA después de diferenciar una vez a una distancia 12 para eliminar los periodos, éstas indican que los datos tienen un comportamiento según un modelo AR(12).

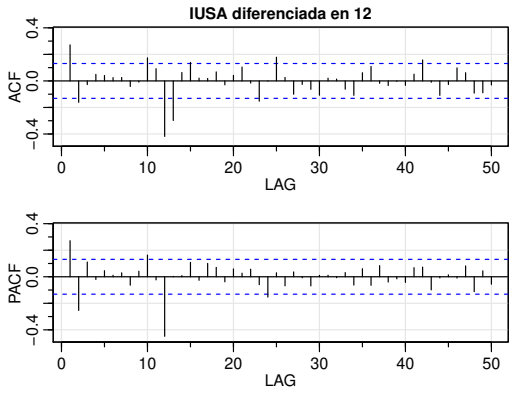

Figura 6.34: En esta figura podemos ver la ACF y la PACF IUSA diferenciada en 12, de esta forma asumimos que la serie IUSA tiene un comportamiento AR(12) y periodos de distancia 12.

Entrenamento de la RNA

Para el entrenamiento de las RNAs se hicieron experimentos utilizando los valores con mayor correlación que se observaron durante el análisis, es decir, los datos a una distancia 1, 11 y 12 o bien los 12 datos que anteceden a la predicción. En la Figura 6.35 se muestran las distribuciones de la frecuencia de la CDP de las RNAs variando los datos de la capa de entrada, como se puede ver, los mejores resultados los dieron las RNAs que se entrenaron con los valores de distancia 11 y 12.

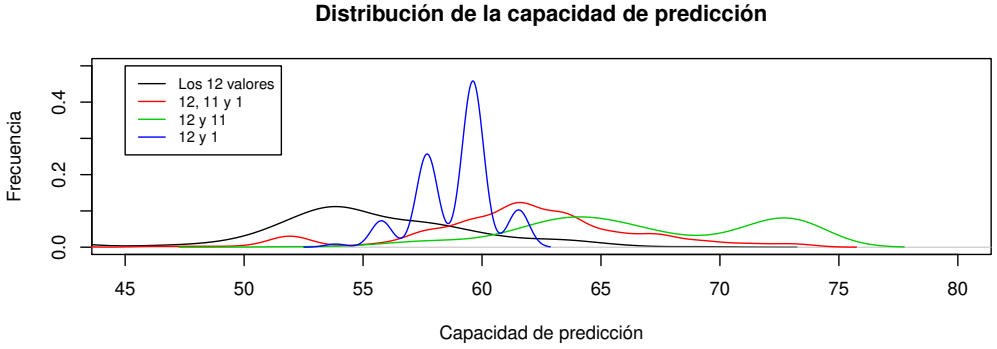

Figura 6.35: Aquí podemos observar las distribciones de la freccuencia de la CDP de la serie IUSA de los modelos de RNAs variando los datos de la capa de entrada. En el recuadro se indican la distancia de los valores con respecto al valor que se está prediciendo con los que se entrenaron las redes.

Cada RNA con diferentes datos de entrada se probó con diferente número de neuronas en la capa oculta. En la Figura 6.36 podemos ver el comportamiento de las distribuciones

de la frecuencia de la IUSA de las RNAs con 2 neuronas en la capa de entrada que usan los valores de distancia hacia atrás 11 y 12. En este caso la RNA con mejor desempeño en la CDP fue la red con 65 neuronas en la capa oculta, obteniendo un mínimo del 57.69 %, una media del 70.49 % y un máximo del 73.07 % de CDP.

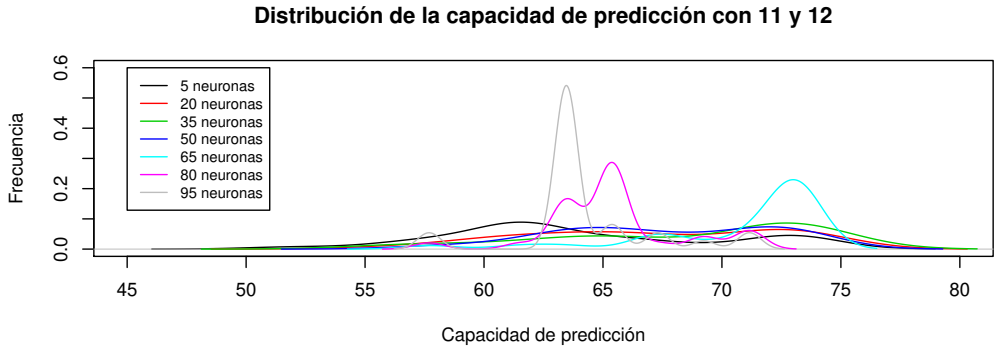

Figura 6.36: Aquí podemos observar las distribuciones de freccuencia de la serie IUSA de los modelos de RNAs con 2 neuronas en la capa de entrada, usando los datos hacia atrás 11 y 12, variando las neuronas de la capa oculta.

Entrenamento de la MSV

Para entrenar la MSV se utilizó en mismo arreglo de datos que en las RNA's, en esta se obtvieron los mejores resultados utilizando los 12 datos hacia atrás del valor que se quiere predecir.

Resultados

En el Cuadro 6.6 se pueden ver los resultados de los modelos implementados para la serie IUSA, es decir, el AR(2) quitando periodos con suavizamiento, el AR(12) eliminando periodos por diferenciación, la RNA con los valores de entrada 11 y 12 y 65 neuronas en la capa oculta y finalmente la MSV con los 12 valores hacia atrás. En la Figura 6.37 también se pueden visualizar las predicciones hechas con cada modelo.

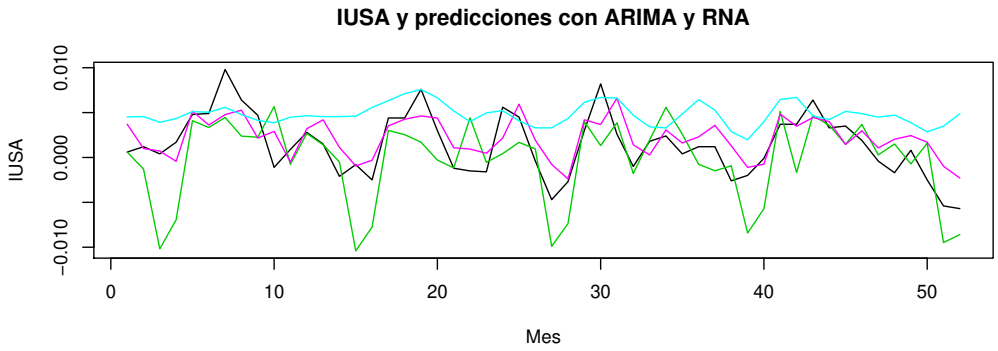

Figura 6.37: En esta gráfica se pueden visalizar las predicciones para la serie de tiempo IUSA que se muestra de color negro. En color verde está la predicción del modelos AR(24) con periodo 12 usando el método de diferenciación. De morado se puede ver la predicción hecha con el modelo AR(2) con periodo 12 usando el método de suavizamiento y finalmente en color azul claro está la predicción del modelo de RNA con 3 neuronas en la capa de entrada, usando los datos hacia atrás 11 y 12, y con 63 neuronas en la capa oculta.

Modelo	CDP	TDE	EE
ARIMA(12,0,0)	55.76 %	2.111 s	1.611082e-05
ARIMA(2,0,0)	57.69 %	0.409 s	5.792974e-06
RNA	71.15 %	2.747 s	1.998536e-05
MSV	66.66 %	78.66 s	-

Cuadro 6.6: Resultados de los diferente modelos ajustados a la serie IUSA.

6.3. Serie de tiempo del covid-19

Como mencionamos antes, en epidemiología surgen series de tiempo que son de interés estudiar. En este caso nos gustaría estudiar la serie de tiempo del número de infectados por covid-19 en el estado de Michoacán, México, del 12 de Enero del 2020 al 14 de enero del 2021. Los datos de esta serie fueron registrados una vez por día y pueden encontrarse en https://datos.covid-19.conacyt.mx/. Al igual que en los ejemplos anteriores de series de tiempo en economía, aquí se analizarán los datos de la serie de tiempo para ajustar un modelo ARIMA y entrenar una RNA y una MSV, con la intención de realizar predicciones del movimiento de esta serie. También para realizar el análisis y hacer el ajuste del modelo estudiaremos solo el 70 % de los datos y con el 30 % restante definiremos el criterio de paro (15 %) y calcularemos la CDP del modelo (15 %).

Análisis de la serie

En la Figura 6.38 podemos ver la serie de tiempo del número de infectados por covid-19 en el estado de Michoacán, ésta nos deja ver claramente la presencia de ciclos de periodo 7, revisando la parte izquierda de la Figura 6.39, donde se muestra la ACF y la PACF de la serie del covid, podemos comprobar la existencia de periodos de longitud 7 en la serie de tiempo de la Figura 6.38.

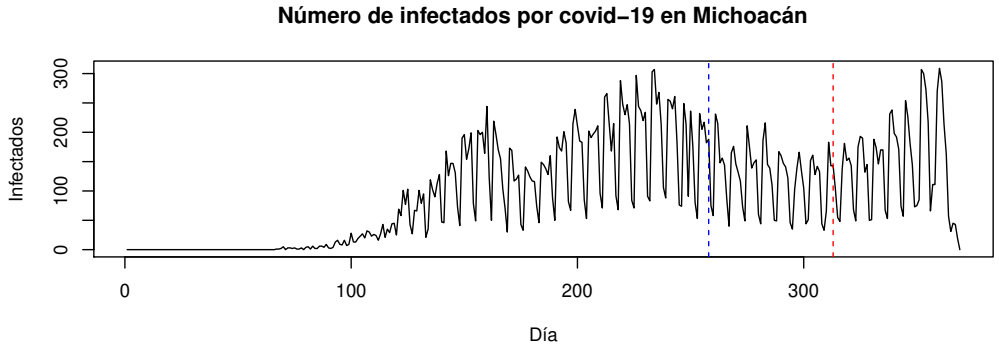

Figura 6.38: En esta gráfica se muestran la cantidad de infectados por día de covid-19 en el estado de Michoacán. Las lineas verticales azul y roja indican la división de los datos en el conjunto de entrenamiesto, validación y prueba de izquierda a derecha.

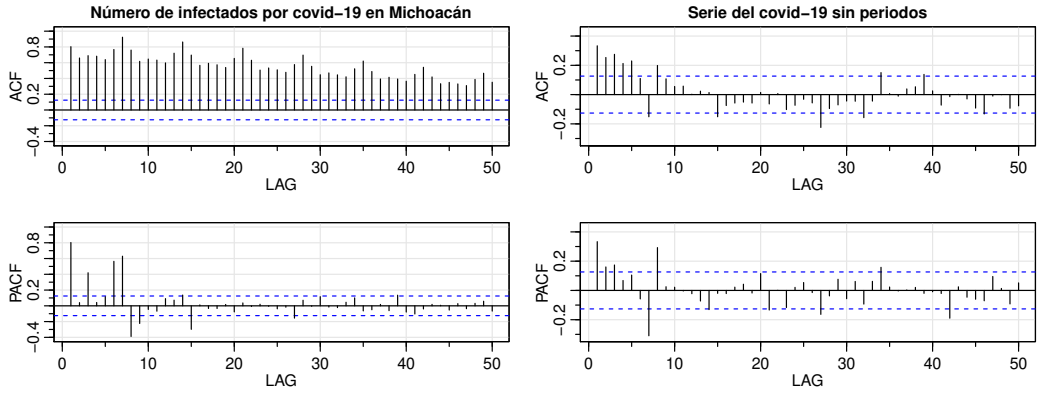

Figura 6.39: En la parte izquierda de esta figura se muestra la ACF y la PACF de la serie de tiempo del número de infectados por covid-19 en Michoacán. En la parte derecha se puede observar la ACF y la PACF de la serie de tiempo del número de infectados por covid-19 en Michoacán después de haber quitado los periodos de longitd 7 diferenciando.

Para eliminar los periodos de la serie del covid-19 en Michoacán, se utilizó el método de diferenciación, en la parte derecha de la Figura 6.39 se muestra la ACF y la PACF de

la serie sin periodos, éstas sugieren un comportamiento AR(8) de la serie de covid sin periodos. En la Figura 6.40 podemos ver la ACF y la PACF de los residuales del ajuste del modelo AR(8) a la serie del covid en Michoacán sin periodos, éstas se aproximan al comportamiento del ruido blanco.

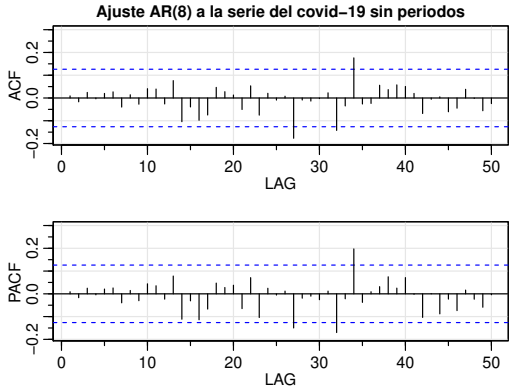

Figura 6.40: En esta figura se pueden ver la ACF y la PACF de los residuales del ajuste del modelo AR(8) a la serie de tiempo del número de infectados por covid-19 en Michoacán, después de haber eliminado los periodos con un método de diferenciación.

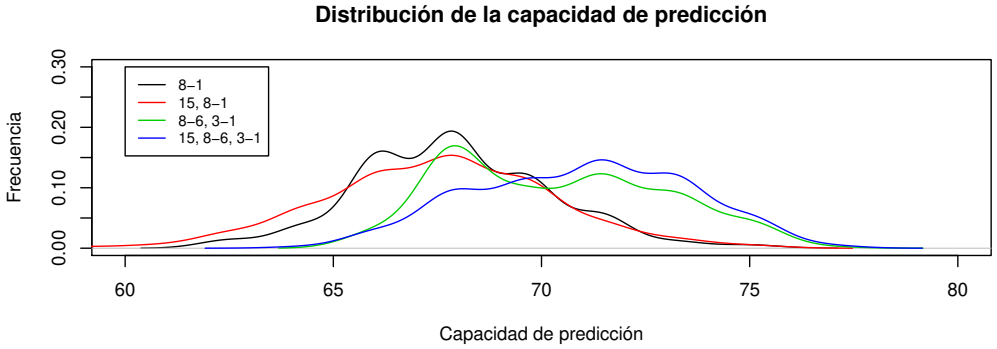

Figura 6.41: En esta figura se muestran las distribuciones de la frecuencia de la CDP de las RNAs que fueron entrenadas en la capa de entrada con los valores que se encuentran en el recuadro de arriba a la izquierda. Los valores que se señalan son las distancias de los datos que se usaron para predecir cierto valor.

Entrenamento de la RNA

Para el entrenamiento de la RNA se probó utilizándo los datos que se encontraban a una distancia del valor que se buscaba predecir, de 1, 2, 3, 4, 5, 6, 7, 8 y 15 que como pudimos ver en el análisis, fueron los valores más correlacionados. En la Figura 6.41 se muestran las

distribuciones de la frecuencia de la CDP, donde se varió el número de neuronas en la capa de entrada, así como los valores que fueron introducidos, obteniendo los mejores resultados en la RNA que utilizó 7 neuronas en la capa de entrada con los valores 1, 2, 3, 6, 7, 8 y 15.

En la Figura 6.42 se muestran las distribuciones de la frecuencia de la CDP de la RNAs entrenadas con los valores 1, 2, 3, 6, 7, 8 y 15 en la capa de entrada y variando las neuronas de la capa oculta. El modelo que mejores resultados arrojó fue la RNA con 50 neuronas en la capa oculta, con un mínimo del 67.85 %, una media del 72.58 % y un máximo del 76.78 % de CDP.

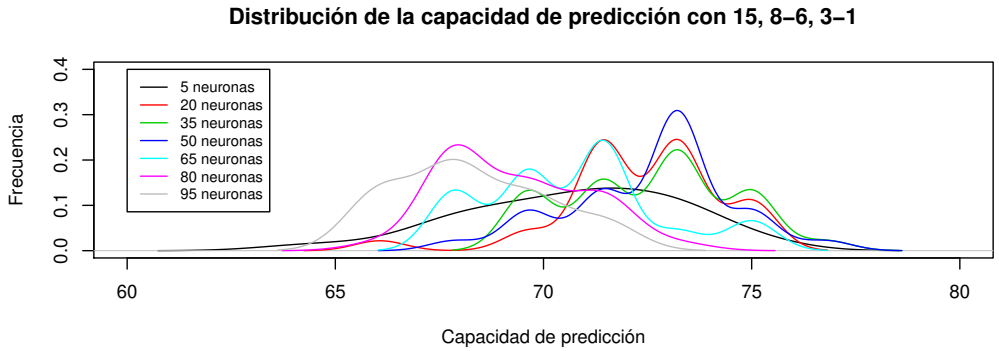

Figura 6.42: Distibuciones de frecuencia de la CDP para las RNAs entrenadas con los valores 1, 2, 3, 6, 7, 8 y 15 y variando las neuronas de la capa oculta.

Entrenamento de la MSV

Para el entrenamiento de las MSV se utilizó la misma estructura de los datos de entrada que para entrenar las RNAs, se obtuvo, al igual que en las RNAs, que los mejores resultados los dio a MSV entrenada con los valores a una distancia 1, 2, 3, 6, 7, 8 y 15 del valor que se quiere predecir.

Resultados

En la Figura 6.43 se muestran las predicciones hechas con los modelos que mejores resultados obtuvieron en la región de prueba, además, en el Cuadro 6.7 se muestran los resultados de la CDP, el tiempo de entrenamiento y el error empírico de cada modelo. Los modelos de los que se muestran resultados son el AR(8) para la serie diferenciada a distancia 7, el modelo de la RNA con 50 neuronas en la capa ocualta y los valores 1, 2, 3, 6, 7, 8 y 15 en la capa de entrada y la MSV con estos mismos valores. En este caso, tanto el modelo de la RNA como el AR(8) mostraron un EE parecido, pero la RNA una vez más sobrepasó a al AR(8) y a la MSV en cuanto a CDP.

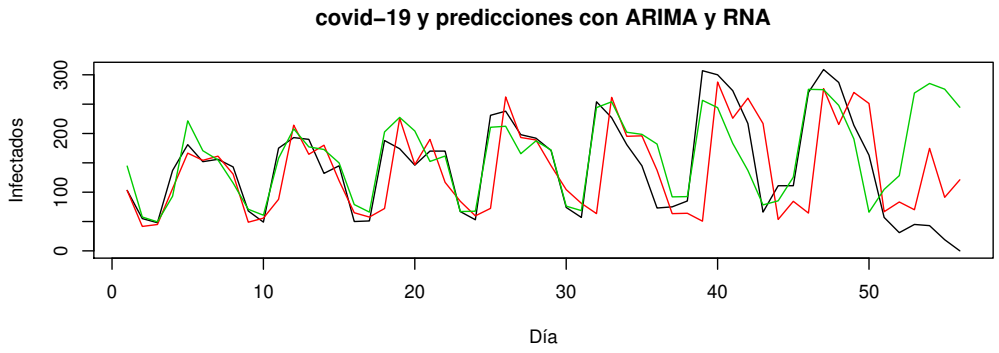

Figura 6.43: En esta gráfica podemos ver la serie de tiempo del covid en la región de prueba en color negro, las predicciones de la RNA en color verde y las predicciones del modelo AR(8) en color rojo.

Modelo	CDP	TDE	EE
ARIMA(8,0,0)	60.714 %	1.015 s	5515.296
RNA	73.21 %	2.944 s	5639.932
MSV	66.06 %	13.36 s	-

Cuadro 6.7: Resultados de los diferente modelos ajustados a la serie del covid-19.

Capítulo 7

Conclusiones

En general, de los problemas tratados en este trabajo se pudo ver un desempeño superior de las RNAs sobre los modelos ARIMA y las MSV en cuanto a la CDP y el TDE, en la mayoría de los casos. Sin embargo, en cuanto al EE, no se obtvieron muy buenos resultados, es decir, las RNAs tuvieron una buena capacidad de aprendizaje del movimiento de las series, pero los resultados estaban en la mayoría de los casos fuera de la realidad.

Entre los principales problemas que se pudieron notar, es que las RNAs tienen problemas para aprender las tendencias, en particular las lineales, un ejemplo de esto lo pudimos ver en la sección 6.2.2 con la serie de tiempo del INPC, que aunque la CPD de la RNA fue mayor y el TDE menor que en los otros modelos, el EE fue excesivamente mayor en la RNA que en el modelo ARIMA (ver Figura 6.17 y Cuadro 6.3). Otro caso parecido fue el de la serie de tiempo del CPI, en este caso aunque el modelo de la RNA no tuvo ni el mejor desempeño en la CDP ni el mejor TDE, estos resultados fueron aceptables, pero en cuanto al EE, definitivamente resultó pésimo (ver Figura 6.24 y Cuadro 6.4).

Otra de las cosas que se pudieron observar en las RNAs, fueron problemas en el aprendizaje de la varianza de los datos. Esto se pudo ver en todos los problemas de series temporales abordados, pues las gráficas de las predicciones en la región de prueba se ven muy suaves. Este problema se puede ver más claramente en las series de tiempo IMEX e IUSA, en estas parece haber aprendido bien el movimiento de la serie ya que la CDP fue mejor que los modelos ARIMA, y tanto el TDE como el EE fueron menores que algunos de los modelos ARIMA y muy parecidos, aunque mayores, en otros (ver Cuadro 6.5 y cuadro 6.6), sin embargo si observamos la Figura 6.31 y la Figura 6.37 podremos ver el exceso de suavidad en las predicciones.

En los problemas de series de tiempo, las MSV no mostraron el mejor desempeño en cuanto a la CDP, pero tampoco fue el peor, en general su desempeño se mantuvo mayor que el de los modelos ARIMA y menor que el de las RNAs. Estas resultaron eficientes en problemas donde los datos eran homogéneos, es decir, donde el tamaño de los grupos de clasificación no difería en gran medida, por ejemplo en la serie de la IMEX. Resultaron pésimas donde se excedían los datos de alguna clase, un ejemplo de esto se pudo ver en la

serie de tiempo del INPC, donde el 80.39 % de los datos eran crecientes y todo el conjunto de prueba lo clasificó como creciente, esto ocurrió también en la serie CPI. Uno de los puntos en contra de las MSV fue el TDE, incluso en el problema de clasificación que fue donde mejor desempeño tuvo, el tiempo de entrenamiento resultó tres veces mayor que el de la RNA y su CDP ligeramente mayor que el 1 %.

De los problemas analizados pudimos ver que los modelos ARIMA no tuvieron muy buena CDP, este hecho ya era esperado dado que estos modelos son lineales y justamente se querían tratar series no lineales, sin embargo, estas se mantuvieron más cerca de la realidad que las RNAs y además, en las series que tenían comportamiento lineal fueron sumamente eficientes, tanto en CDP como en EE, como es el caso de las series INPC y CPI. El TDE de los modelos ARIMA(p,d,q) varia dependiendo del tamaño de p y de q, puede ser muy rápido si p y q son pequeñas y muy lento e incluso ineficiente si son muy grandes.

El análisis de series de tiempo fue esencial, no solo para la selección de los parámetros de los modelos ARIMA, sino para la elección de lo valores de entrenamiento para las RNAs y las MSV, pues ya que sin esto no se habría podido obtener buenos resultados de los algoritmos de inteligencia artificial, dado que al agregar neuronas con datos poco correlacionados generaba malos entrenamientos, incrementaba el TDE y disminuía la CDP.

Aunque en algunas de las series se pudieron obtener buenos resultados de CDP, en algunas otras fueron poco confiables los resultados de predicción, estos resultados podrían ser mejorados en futuros trabajos abarcando más en la teoría del análisis de series de tiempo para probar otros modelos, además, dado que las series en economía dependen de diferentes factores, según [3] una opción para mejorar los resultados de predicción podría ser analizando series multivariadas, también se podría trabar con métodos híbridos como lo hace [4], para aprovechar las ventajas de los modelos lineales y los no lineales al mismo tiempo. En cuanto a los algoritmos de inteligencia artificial, se podrían implementar otros métodos de optimización más eficientes y probar con diferentes estructuras, como podrían ser RNAs recurrentes o convolucionales.

Apéndice A

Probabilidad

El término probabilidad se puede interpretar de una manera práctica como una medida de la creencia de que un evento futuro pueda ocurrir, sin embargo, se requiere de un concepto más claro para poder medirla y saber como ayuda a hacer inferencias. El objetivo de la teoría de probabilidades es desarrollar modelos para experimentos que generan observaciones que no se pueden predecir con certeza pero la frecuencia con la que ocurren en una larga serie de intentos es estable. Los eventos que poseen esta propiedad reciben el nombre de eventos aleatorios o estocásticos.

Para el desarrollo de este apéndice se basó en el contenido de [9], el cuál se recomienda revizar en caso de querer profundizar en el tema.

A.1. Espacio de probabilidad

Se puede definir un *experimento* como el proceso por medio del cual se hace una observación. Por ejemplo, observar el número que aparece en la cara superior de un dado al lanzarlo, registrar el precio diario de una acción en particular, medir el cociente de inteligencia (IQ) de una persona o determinar el número de bacterias por centímetro cúbico en una porción de alimento procesado.

El modelo fundamental para describir experimentos gobernados por el azar es el *espacio de probabilidades* y está descrito por tres componentes principales.

La primera es un conjunto Ω, conocido como el *espacio muestral*, este es el conjunto de todos los resultados posibles de un experimento. Los elementos de un espacio muestral se les conoce como *eventos o sucesos simples* y se denotan como ω.

Por ejemplo, en el experimento de lanzar un dado y ver el número que aparece en la parte superior, el conjunto de resultados que se pueden observar en este experimento o bien el espacio muestral es $\Omega = \{1, 2, 3, 4, 5, 6\}$, un evento simple podría ser $\omega = 3$ o cualquier otro número de 1 al 6.

La segunda componente de nuestro modelo es la clase de eventos o sucesos $\mathcal{F}$. Esta

clase está compuesta por subconjuntos de Ω y debe satisfacer las siguientes propiedades

Propiedad A.1.1. $\Omega \in \mathcal{F}$.

Propiedad A.1.2. *Si* $A \in \mathcal{F}$ *entonces* $A^c \in \mathcal{F}$.

Propiedad A.1.3. *Si* $A_n \in \mathcal{F}$ *para* $n \geq 1$ *entonces* $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.

Una colección de subconjuntos de Ω que satisfacen tres propiedades anteriores se conoce como $\sigma - \acute{a}lgebra$. Dado cualquier conjunto A, un ejemplo de sigma algebra es el conjunto $\mathcal{P}(A)$.

La tercera y última componente de nuestro modelo es una *probabilidad* P, la cual es una función que está definida sobre la clase de conjuntos $\mathcal{F}$ y toma valores en el intervalo [0,1], la probabilidad debe satisfacer las siguientes propiedades:

Propiedad A.1.4. *Para cualquier evento* $A \in \mathcal{F}$,

$$0 = P(\emptyset) \leq P(A) \leq P(\Omega) = 1.$$

Propiedad A.1.5. *Si* $\{A_n, n \geq 1\}$ *es una colección de conjuntos disjuntos dos a dos, es decir,* $A_i \cap A_j = \emptyset$ *si* $i \neq j$, *entonces*

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

La terna $(\Omega, \mathcal{F}, P)$ es conocida como *espacio de probabilidad* y la función P es una *medida de probabilidad*. Dado que P está definida sobre $\mathcal{F}$, sólo podemos determinar la probabilidad de los conjuntos que están en esta clase; por eso decimos que estos son los conjuntos medibles. Las propiedades de $\mathcal{F}$ garantizan que si hacemos las operaciones usuales (unión, intersección, diferencia, diferencia simétrica, complemento) con conjuntos medibles, obtenemos conjuntos medibles. Por ejemplo, si $A \subset B$ son medibles entonces $A \backslash B$ también. En consecuencia tenemos que como $B = A \cup (B \backslash A)$ es unión de eventos disjuntos, entonces $P(B) = P(A) + P(B \backslash A) \geq P(A)$, por lo tanto $P(A) < P(B)$ si $A \subset B$.

En el caso de experimentos sencillos, por ejemplo experimentos con un conjunto finito de resultados, normalmente tomamos como $\sigma - \acute{a}lgebra$ el conjunto potencia de Ω. En experimentos más complicados, con una cantidad no numerable de resultados posibles, no siempre es posible tomar esta opción, y es necesario considerar alguna $\sigma - \acute{a}lgebra$ más pequeña.

A.2. Probabilidad condicional

La *probabilidad condicional* $P(A|B)$ del evento A dado el evento B, se define por

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad si \quad P(B) > 0,$$

no está definida o se le asigna un valor arbitrario cuando $P(B) = 0$. A partir de esta definición tenemos la relación

$$P(A \cap B) = P(B)P(A|B). \tag{A.1}$$

Supongamos que saber que el evento B ocurrió no cambia la probabilidad de que A ocurra, es decir $P(A|B) = P(A)$. En este caso la expresión (A.1) se convierte en

$$P(A \cap B) = P(B)P(A) \tag{A.2}$$

y se dice que si los eventos A y B salisfacen la expresión (A.2) son independientes.

A.2.1. Ley de la probabilidad total y teorema de Bayes

Sea $\{B_n, n \geq 1\}$ una partición de Ω, es decir, si $i \neq j$, entonces $B_i \cap B_j = \emptyset$ y

$$\Omega = \bigcup_n B_n.$$

Entonces, para cualquier evento A, teniendo en cunta que los subconjuntos $(A \cap B_j)$, $j \geq 1$ son disjunto dos a dos,

$$P(A) = P(A \cap \Omega) = P(A \cap (\cup_n B_n)) = \sum_n P(A \cap B_n)$$

y ahora, usando la ecación (A.1) en cada sumando, se deduce la conocida *ley de la probabilidad total* que está dada por la siguiente ecuación

$$P(A) = \sum_n P(A|B_n)P(B_n). \tag{A.3}$$

Una consecuencia de la ley de probabilidad total es el Teorema A.2.1 conocido como teorema de Bayes.

Teorema A.2.1. *Si los eventos* $\{B_j, j \geq 1\}$ *forman una partición de* Ω*, para cualquier evento A, con* $P(A) \geq 0$ *entonces para cualquier evento A con* $P(A) > 0$ *y cualquier índice j,*

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_n P(A|B_n)P(B_n)}.$$

Proposición A.2.1. *Sean $A_1, \ldots, A_n$ eventos cualesquiera. Entonces*

$$P(A_1 \cup \ldots \cup A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cup A_1)\ldots P(A_n|A_{n-1} \cup \ldots \cup A_1).$$

A.3. Variables aleatorias

Una *variable aleatoria* (v.a.) X es una función de valor real, definida sobre un espacio muestral que debe de satisfacer ciertas condiciones de medibilidad las cuales serán descritas a continuación.

Ya que una v.a. toma valores en los reales, nos interesa poder calcular $P(X < a)$ para $a \in \mathbb{R}$. Ahora bien, para que estas probabilidades existan es necesario que los conjuntos cuyas probabilidades deseamos calcular sean medibles, es decir, que pertenezcan a un $\sigma - algrbra$ $\mathcal{F}$. Estos conjuntos son de la forma $\{\omega|X(\omega) \leq a\}$, entonces queremos que la función X satisfaga la Propiedad A.3.1.

Propiedad A.3.1. *Para todo número $a \in \mathbb{R}$ se tiene que $\{\omega|X(\omega) \leq a\} \in \mathcal{F}$.*

Las funciones $X : \Omega \to \mathbb{R}$ que cumplen esta propiedad son conocidas como *funciones medibles*, por lo tanto, una v.a. es una función medible. La medibilidad de una función depende de la $\sigma - algebra$.

A.4. Distribuciones de una variable aleatoria

Consideremos un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ sobre el cual hemos definido una variable aleatoria X con valores reales. Si A es un intervalo de $\mathbb{R}$ y queremos calcular la probabilidad de que X tome valores en A se considera el conjunto

$$\{\omega \in \Omega|X(w) \in A\} = X^{-1}(A),$$

es decir la preimagen del intervalo A por la función X. Como X es una función medible $X^{-1}(A) \in \mathcal{F}$ y podemos calcular su probabilidad

$$P_X(A) = P(X \in A) = P(\{\omega \in \Omega|X(w) \in A\}) = P(X^{-1}(A)).$$

Esta relación nos permite definir una *medida de probabilidad* inducida a X, donde $P_X(A)$ es conocida como *distribución* y contiene toda la información probabilística de la variable aleatoria X.

A.5. Funciones de distribición

Sea $(\Omega, P, \mathcal{F})$ un espacio de probabilida, donde $\mathcal{F}$ es la $\sigma - \textit{álgebra de Borel } \mathcal{B}$, es decir la $\sigma - \textit{álgebra}$ generada por los intervalos de $\mathbb{R}$, entonces los intervalos de la forma $(-\infty, x] \in \mathcal{F}$, para $x \in \mathbb{R}$ y podemos definir una función $F : \mathbb{R} \to \mathbb{R}$ como

$$F(x) = P_X((-\infty, x]) = P(X \leq x). \tag{A.4}$$

La funcion de la ecuación (A.4) se llama *función de distribución* de la variable aleatoria X y se abrevia (f.d.). Conociendo la distribución de una variable aleatoria se puede conocer su f.d. y también el recíproco es cierto. Una f.d. se caracteriza por las sigientes propiedades:

Propiedad A.5.1. *F es continua por la derecha y tiene límite por la izquierda.*

Propiedad A.5.2. *F es monótona no decreciente.*

Propiedad A.5.3. *Si F es una función de distribucion entonces*

$$\lim_{x \to -\infty} F(x) = 0 \;\; y \;\; \lim_{x \to \infty} F(x) = 1.$$

Si una función cumple las propiedades de una f.d. entonces existe una v.a. X definida sobre un espacio de probabilidad $(\Omega, P, \mathcal{F})$ tal que F es la f.d. de X.

A.6. Variables discretas

Una variable aleatoria X es *discreta* si toma una cantidad finita de valores $\{x_1, \ldots, x_n\}$ o a lo más numerable $\{x_1, x_2, \ldots\}$. En el primer caso existen números positivos $p_1, \ldots, p_n$ con $p_1 + \ldots + p_n = 1$ tales que

$$P(X = x_i) = p_i. \tag{A.5}$$

Llamaremos a la función $p_X(x_i) = p_i$ *función de probabilidad o densidad* de X.

De manera similar, en el segundo caso tenemos números positivos $p_i, i \geq 1$ que satisfacen $\sum_{i=1}^{\infty} p_i = 1$ y la ecuación (A.5) para $i \geq 1$.

En ambos casos, las funciones de distribución son funciones de saltos, es decir, que solo crece en los puntos x_i y son constantes entre puntos consecutivos. La función de distribución para las variables aleatorias discretas está definida como

$$F(x) = \sum_{x_i \leq x} x_i.$$

A.7. Variables continuas

Una variable aleatoria X es *continua* si su función de distribución F es continua. Otra definición es que si x es cualquier evento elemental de la v.a. X, entonces $P(X = x) = 0$. Para la mayoria de estas variables aleatorias existe una función $f : \mathbb{R} \to \mathbb{R}$ con $f(x) \geq 0$ para todo x y $\int_{-\infty}^{\infty} f(x)dx = 1$ que satisface

$$F(x) = \int_{-\infty}^{x} f(u)du \ \ (-\infty < x < \infty), \tag{A.6}$$

para todo $x \in \mathbb{R}$. La función f es conocida como *densidad* de la variariable aleatoria X. Hay algunas variables continuas que no tienen esta propiedad, es decir, que no existe una función cuya integral sea la función de distribución de la variable aleatoria, sin embargo, el interés de este tipo de variableas aleatorias es meramente teórico. Por lo tanto, al hablar de variables continuas nos estaremos refiriendo a esas que cumplen con la propiedad de la ecuación (A.6). Si F es diferenciable en x entonces la densidad de probabilidad está dada por

$$f(x) = \frac{dF(x)}{dx} = F'(x) \quad -\infty < x < \infty.$$

A.8. Valores esperados y momentos

Si X es una variable aleatoria discreta, su momento de orden n está dado por

$$E(X^n) = \sum_i x_i^n P(X = x_i),$$

siempre que la serie converja absolutamente, en caso de no converger se dice que el momento de orden n no existe. Si X es una variable aleatoria continua con función de dencidad $f(x)$ entonce su momento de orden n se define como:

$$E(X^n) = \int_{-\infty}^{\infty} x^n f(x)dx,$$

siempre y cuando esta integral converja absolutamente.

El primer momento corresponde a $n = 1$, se conoce como *media, valor esperado o esperanza matemática* de X, y lo denotamos como μ_X. El *momento central* de orden n está definido como el momento de orden n de la variable aleatoria $X - \mu_X$ siempre y cuando μ_X exista. El primer momento central es cero, el segundo es conocido como *varianza* y se denota $Var(X)$. La raiz cuadrada se le llama *desviación estandar*.

Teorema A.8.1. *$Var(X) = E(X^2) - \mu_X^2$, donde $\mu = E(X)$*

Si X es una variable aleatoria y g es una función medible, entonces $g(X)$ es también una variable aleatoria. Si X es una v.a. discreta el valor esperado de $g(X)$ es

$$E(g(X)) = \sum_i g(x_i)P(X = x_i),$$

siempre que la suma converja absolutamente. Si X es una v.a. continua con densidad $f(x)$ el valor esperado de $g(X)$ está dado por

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Una fórmula general que abarca ambos casos e incluso el mixto está dada por la ecuación (A.7), que es una integral de Lebesgue-Stieltjes cuya definición va más allá del propósito de este apéndice, por lo tanto, nos limitaremos a ver variables aleatorias discretas y continuas.

$$E(g(X)) = \int g(X)dF_X(x) \tag{A.7}$$

A.9. Distribuciones conjuntas e independencia

Si tenemos un par de variables aleatoria (X,Y) definidas sobre un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, su función de distribución conjunta F_{XY} está definida por

$$F_{XY}(x, y) = F(x, y) = P(X \leq x, Y \leq y).$$

Si ambas variables aleatorias son discretas y toman valores $x_i, i \geq 1$ y $y_i, i \geq 1$, entonces su función de probabilidad conjunta es

$$p_{XY}(x_i, y_j) = P(X = x_i, Y = y_j).$$

Una función de distribución conjunta tiene densidad conjunta si existe una función f_{XY} de dos variables que satisface

$$F_{XY}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(u, v)dudv \text{ para todo } x, y.$$

La función $F_X(x) = \lim_{y\to\infty} F_{XY}(x, y)$ es un f.d. que se conoce como *función de distribución marginal* de X. Si las variables aleatorias son ambas discretas, las funciones de probabilidad marginal están dadas por

$$p_X(x_i) = \sum_j p_{XY}(x_i, y_j) \text{ y } p_X(y_j) = \sum_i p_{XY}(x_i, y_j).$$

Si la f.d F tiene una densidad conjunta f, entonces las funciones de distribución marginal de X y Y están dadas respectivamente por

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \;\; y \;\; f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx.$$

Teorema A.9.1. *Si X e Y son variables aleatorias con distribución conjunta y c es una constante, entonces*

$$E(cX + Y) = cE(X) + E(Y),$$

siempre y cuando cada momento exista.

Si para todos los valores x e y se tiene que $F_{XY}(x,y) = F_X(x)F_Y(y)$ decimos que las variables X e Y son *independientes*. Si las variables son discretas y tienen función de probabilidad conjunta p_{XY} entonces son independientes si y solo si

$$p_{XY}(x,y) = p_X(x)p_Y(y).$$

De manera similar, si las variables son continuas y tienen función de densidad conjunta f_{XY} entonces son independientes si y solo si

$$f_{XY}(x,y) = f_X(x)f_Y(y).$$

Teorema A.9.2. *Si X,Y son v.a. independientes con primer momento finito, entonces el producto XY también tiene primer momento finito y*

$$E(XY) = E(X)E(Y).$$

Este resultado se extiende a cualquier colección finita de variables independientes.

En general, para una colección de variables aleatorias $(X_1, \ldots, X_n)$ o bien, un vector aleatorio definido en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, la distribución conjunta se define como

$$F(X_1, \ldots, X_n) = F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

y se dice que las variables $X_1, \ldots, X_n$ son independientes si y solo si para todos los valores posibles $x_1, \ldots, x_n$ se cumple

$$F_{X_1 \ldots X_n}(x_1, \ldots, x_n) = F_{X_1}(x_n) \ldots F_{X_n}(x_n).$$

Una función de distribución conjunta $F(X_1, \ldots, X_n)$ tiene función de densidad conjunta $f(t_1, \ldots, t_n)$ si

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} f(t_1, \ldots, t_n)dx_1 \ldots dx_n.$$

A.10. Covarianza y correlación

Si X e Y son variables aleatorias con distribución conjunta, medias μ_X, μ_Y y varianzas fininitas σ_X^2, σ_Y^2 entonces la *covarianza* de X e Y, denotada como Cov(X,Y) o σ_{XY} se define como

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]. \tag{A.8}$$

Teorema A.10.1. *Si X y Y son variables aleatorias con medias μ_1 y μ_2, respectivamente, entonces*

$$Cov(X, Y) = E[XY] - \mu_X\mu_Y.$$

La covarianza es una medida del grado de dependencia lineal entre las variables aleatorias y si $Cov(X, Y) = 0$ se dice que X e Y no están correlacionadas.

Teorema A.10.2. *Si X e Y son v.a independientes entonces $Cov(X, Y) = 0$*

El recíproco del Teorema A.10.2 no es cierto, pues hay variables aleatorias que no están correlacionadas pero no son independientes. Un ejemplo de este caso lo podemos encontrar en el Ejemplo 5.24 de [9].

Teorema A.10.3. *Sean $\{X_j\}$ y $\{Y_k\}$ variables aleatorias con varianza finita, si U y V son combinaciones lineales de $\{X_j\}$ y $\{Y_k\}$ respectivamente, es decir*

$$U = \sum_{j=1}^{m} a_j X_j, \quad V = \sum_{k=1}^{r} b_k Y_k,$$

entonces

$$Cov(U, V) = \sum_{j=1}^{m} \sum_{k=1}^{r} a_j b_k Cov(X_j, Y_k).$$

Además $Var(U) = Cov(U, U)$.

Demostración.

$$\begin{aligned}
Cov(U,V) &= E(UV) - E(U)E(V) \\
&= E(\sum_{j=1}^{m} a_j X_j \sum_{k=1}^{r} b_k Y_k) - \sum_{j=1}^{m} a_j E(X_j) \sum_{k=1}^{r} b_k E(Y_k) \\
&= \sum_{j=1}^{m}\sum_{k=1}^{r} a_j b_k E(X_j Y_k) - \sum_{j=1}^{m}\sum_{k=1}^{r} a_j b_k E(X_j)E(Y_k) \\
&= \sum_{j=1}^{m}\sum_{k=1}^{r} a_j b_k (E(X_j Y_k) - E(X_j)E(Y_k) \\
&= \sum_{j=1}^{m}\sum_{k=1}^{r} a_j b_k Cov(X_j, Y_k)
\end{aligned}$$

□

Es difícil utilizar la covarianza como medida absoluta de dependencia porque su valor depende de la escala de medición. En consecuencia, es difícil determinar a primera vista si una covarianza particular es grande o pequeña. Este problema se puede eliminar al estandarizar su valor y usar el *coeficiente de correlación*, una cantidad relacionada con la varianza y que se define como

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Teorema A.10.4. *Sean X e Y variables aleatorias y ρ_{XY} el coeficiente de correlación de X e Y, entonces $-1 \leq \rho_{XY} \leq 1$.*

Demostración. Si U y V son v.a. con distribución conjunta, entonces $E(UV)$ define un producto interno. De este modo, si X e Y son variables aleatorias, μ_X y μ_Y sus medias respectivamente, $U = (X - \mu_x)$ y $V = (Y - \mu_y)$, tenemos por la desiguldad de Cauchy-Schwarz que

$$\begin{aligned}
|E(UV)|^2 &\leq E(UU)E(VV) \\
|E[(X-\mu_X)(Y-\mu_Y)]|^2 &\leq E[(X-\mu_X)^2]E[(Y-\mu_Y)^2] \\
|\sigma_{XY}|^2 &\leq \sigma_X \sigma_Y \\
\frac{|\sigma_{XY}|}{\sqrt{\sigma_X \sigma_Y}} &\leq 1 \\
|\rho_{XY}| &\leq 1
\end{aligned}$$

por lo tanto $-1 \leq \rho_{XY} \leq 1$, que es lo que queriamos demostrar. □

El signo del coeficiente de correlación es igual al signo de la covarianza. Entonces, $\rho > 0$ indica que Y aumenta a medida que X aumenta y $\rho = +1$ implica correlación perfecta, con todos los puntos cayendo en una recta con pendiente positiva. Un valor de $\rho = 0$ implica cero covarianza y que no hay correlación. Un coeficiente negativo de correlación implica una disminución en Y cuando X aumenta, y $\rho = -1$ implica correlación perfecta, con todos los puntos cayendo en una recta con pendiente negativa.

A.11. Probabilidad y esperanza condicional

A.11.1. Caso discreto

Sean X e Y variables aleatorias discretas. La función de probabilidad condicional de X dado Y=y se define como

$$p_{X|Y}(x|y) = \frac{P(X = x, Y = y)}{P(Y = y)} =, \quad si \;\; P(Y = y) > 0$$

y no está definida, o se asigna un valor arbitrario, si $P(Y = y) = 0$. En terminos de la densidad conjunta $p_{XY}(x, y)$ y de la densidad marginal de Y, $p_Y(y) = \sum_x p_{XY}(x, y)$, la definición es

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}, \quad si \;\; p_Y(y) > 0.$$

Para cada valor fijo y, $p_{X|Y}(x|y)$ es una dencidad de probabilidad , pues

$$p_{X|Y}(x|y) > 0 \;\; y \;\; \sum_x p_{X|Y}(x, y) = 1, \quad para \;\; todo \;\; y.$$

La ley de probabilidad total es

$$P(X = x) = \sum_y P(X = x|Y = y)P(Y = y) = \sum_y p_{X|Y}(x|y)p_Y(y).$$

Sea g una función tal que $E[g(X)] < \infty$. Definamos la esperanza condicional de $g(X)$ dado $Y = y$ por la formula

$$E[g(X)|Y = y] = \sum_x g(x)p_{X|Y}(x|y), \quad para \;\; p_Y(y) > 0,$$

la esperanza condicional no está definida para valores en los que $p_Y(y) = 0$. La ley de la probabilidad total en la esperanza condicional es:

$$E[g(X)] = \sum_y E[g(X)|Y = y]p_Y(y).$$

La esperanza condicional es una función real, es decir, $E[g(X)|Y = y] = \phi(y)$. Si evaluamos la función ϕ en la variable aleatoria Y, tenomos que $\phi(Y)$ es también una variable aleatoria que denotaremos como $E[g(X)|Y]$, entonces

$$E[g(X)|Y](\omega) = E.$$

Podemos escribir ahora la ley de probabilidad total como

$$E[g(X)] = E[E[g(X)|Y]].$$

A.11.2. Caso continuo

Sean X e Y vaiables aleatorias con distrubución conjunta continua y función de dencidad $f_{XY(x,y)}$, definimos la densidad condicional de X dado $Y = y$ como

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_y(y)}, \quad para \quad f_Y(y) > 0$$

y la densidad condicional no está definida si $f_Y(y) = 0$. Si g una función tal que $E[g(X)] < \infty$, La esperanza condicional de $g(X)$ dado que $Y = y$ se define por

$$E[g(X)|Y = y] = \int g(x) f_{X|Y}(x|y), \quad para \quad f_Y(y) > 0.$$

También tenemos en este caso la propiedad de que

$$E[g(X)] = E[E[g(X)|Y].$$

A.12. Algunas distribuciones importantes

A.12.1. Distribuciones discretas

Distribución de Bernoulli. Una variable aleatria Bernoulli toma valores 0 y 1 con probabilidades respectivas p y $q = 1 - p$, es decir, la función de probabilidad está dada por $P(X = 1) = p$ y $P(X = 0) = q$, donde $0 \leq p \leq 1$. Si el resultado del experimento es 1 decimos que ocurrió un éxito con probabilidad p. La media y la varianza respectivas son:

$$E(X) = p \text{ y } Var(X) = pq.$$

Distribución Binomial. Consideremos $X_1, \ldots, X_n$ una colección de n variables aleatorias independientes Bernoulli con probabilidad de éxito p. Sea X el total de éxitos de las n variables aleatorias, entonces $X = \sum_{i=1}^{n} X_i$. La distribución de X es binomial con parámetros

n y p o bien $X \sim \mathcal{B}in(p, q)$ si la función de probabilida de X es

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ para } k = 1, 2, \ldots, n.$$

La media y la varianza de esta distribución son respectivamente

$$E(X) = np \;\; y \;\; Var(X) = np(1-p).$$

Distribución Geométrica. En el mismo contexto del experimento anterior, sea Z la variable aleatoria que cuenta el número de ensayos antes lograr el primer éxito, es decir, si $Z = k$, quiere decir que los $k-1$ ensayos anteriores fueron fracasos, cada uno con probabilidad $(1-p)$ y el k-ésimo intento fue un éxito con probabilidad p. Si Z tiene distribución geométrica entonces su función de probabilidad es

$$p_Z(k) = P(Z = k) = (1-p)^{k-1} p, \;\; para \;\; k = 1, 2, \ldots,$$

por lo tanto el valor esperado y la varianza de una variable aleatoria con esta distribución están dados como

$$E(Z) = \frac{1}{p} \;\; y \;\; Var(Z) = \frac{1-p}{p^2}.$$

Distribución de Poisson. Una variable aleatoria X con una distribución de Poisson de parámetro $\lambda > 0$ o bien $X \sim \mathcal{P}ois(\lambda)$, tiene función de probabilidad

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda} \;\; para \;\; k = 1, 2, \ldots.$$

Esta distribución tiene las siguientes media y varianza respectivamente

$$E(X) = \lambda \;\; y \;\; Var(X) = \lambda.$$

Distribución multinomial. Las variables aleatorias $X_1, \ldots X_k$ con valores en $\{0, 1, \ldots, n\}$, tienen una distribución multinomial si su función de probabilidad conjunta es

$$P(X_1 = r_1, \ldots, X_k = r_k) = \begin{cases} \frac{n!}{r_1!, \ldots, r_k!} p_1^{r_1} \ldots p_k^{r_k} & si \;\; r_1 + \ldots + r_k = n \\ 0 & si \;\; no \end{cases},$$

donde $p_i > 0$, para $i = 1, \ldots, k$ y $p_1 + \ldots p_k = 1$.

Para esta distribución se tiene que

$$E(X_i) = p_i, \;\; Var(X_i) = np_i(1-p_i) \;\; y \;\; Cov(X_i, X_j) = -np_i p_j.$$

A.12.2. Distribuciones continuas

Distribución Normal. Una variable aleatoria X tiene una distribución normal con parámetros μ y σ^2 si su función de densidad está dada por

$$\phi(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty.$$

Para denotar las variables aleatorias con esta distribución usaremos $X \sim \mathcal{N}(\mu, \sigma^2)$, donde μ y σ^2 son la media y la varianza respectivamente de la variable X. Cuando $\mu = 0$ y $\sigma^2 = 0$ se dice que X tienen una *distribución normal estandar*.

Distribución Lognormal. Si $\log(V)$ tiene una distribución normal, decimos que V tiene una distribución lognormal. Recíprocamente, si $X \sim \mathcal{N}(\mu, \sigma^2)$, entonces $V = e^X$ es una variable con distribución lognormal. Haciendo un cambio de variable para la densidad obtenemos

$$f_V(v) = \frac{1}{\sqrt{2\pi}\sigma v} \exp -\frac{1}{2}\left(\frac{\log v - \mu}{\sigma}\right)^2, \quad v \geq 0.$$

La media y la varianza son, respectivamente

$$E(V) = \exp \mu + \frac{1}{2}\sigma^2 \;\; y \;\; Var(V) = \exp 2(\mu + \frac{1}{2}\sigma^2)(e^{\sigma^2} - 1).$$

Distribución exponencial. Una variable aleatoria T tiene distribución exponencial con parámeto $\lambda > 0$ y se denota $T \sim \mathcal{E}xp(\lambda)$ si su función de densidad es

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & si \;\; t \geq 0 \\ 0 & si \;\; t < 0 \end{cases}.$$

La media y la varianza están dadas por

$$E(T) = \frac{1}{\lambda} \;\; y \;\; Var(T) = \frac{1}{\lambda^2}.$$

Distribución Uniforme. Una variable aleatoria U tiene distribución uniforme en el intervalo [a,b] y se denota $U \sim \mathcal{U}nif[a, b]$, con $a < b$ si U tiene función de densidad

$$f_U(u) = \begin{cases} \frac{1}{b-a} & \text{para } \; u \in [a, b] \\ 0 & \text{para } \; u \notin [a, b] \end{cases}.$$

La media y la varianza están dadas como

$$E(U) = \frac{1}{2}(a + b) \;\; y \;\; Var(U) = \frac{(b-a)^2}{12}.$$

Distribución normal multivariada. El vector aleatorio $X = (X_1, \ldots, X_n)$ se dice que

tiene una distribución normal multivariada si su función de densidad conjunta está dada por

$$f_X(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp{(x-\mu)^T \Sigma^{-1}(x-\mu)},$$

donde $|\Sigma|$ es el determinante de la matriz de covarianzas y $\mu = (\mu_1, \ldots, \mu_n)^T$ es el vector de medias del vector aleatorio. La notación de esta variable aleatoria es $X \sim \mathcal{N}(\mu, \Sigma)$.

Apéndice B

Optimización

En este apéndice daremos una breve introducción a la teoría de optimización, sin enmbargo para mejores referencias se puede consultar [15] y [16], donde se aborda ampliamente el tema de optimización convexa, un gran número de sus aplicaciones y los métodos numéricos que podrian ser utiles para resolver los problemas planteados. También se puede consultar [17] para revisar más sobre optimización numérica. En la bibliografía mencionada en este párrafo se basó el contenido de este apéndice.

B.1. Optimización

Para describir un problema de optimización en su forma estandar estaremos usando la siguiente notación:

$$\begin{aligned} &\text{minimizar} && f_0(x) \\ &\text{sujeta a} && f_i(x) \leq 0 \;\; i = 1, \ldots, m, \\ &&& h_i(x) = 0 \;\; i = 1, \ldots, p, \end{aligned} \tag{B.1}$$

donde $x \in \mathbb{R}^n$, $f_i : \mathbb{R}^n \to \mathbb{R}$ para $i \in \{0, 1, \ldots m\}$ y $h_i : \mathbb{R}^n \to \mathbb{R}$ para $i \in \{1, 2, \ldots p\}$. Llamaremos a x *variable de optimización*, la función $f_0(x)$ es conocida como *función objetivo*, mientras que $h_i(x) = 0$ para $i \in \{1, 2, \ldots p\}$ y $f_i(x) \leq 0$ para $i \in \{1, 2, \ldots m\}$ como *restricciones* (de igualdad y desigualda respectivamente). Un *problema sin restricciones* es aquel para el cual $m = p = 0$.

El *dominio de optimización* estará definido por el conjunto de funciones del problema (B.1), es decir, tanto la función objetivo como las restricciones. El dominio de optimización está dado por la ecuación (B.2) donde $\mathbf{dom}(h_i)$ y $\mathbf{dom}(f_i)$ es el dominio de h_i para $i \in \{1, 2, \ldots p\}$ y f_i para $i \in \{1, 2, \ldots m\}$ respectivamente.

$$\mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom}(f_i) \cap \bigcap_{i=1}^{p} \mathbf{dom}(h_i) \tag{B.2}$$

Si $x \in \mathcal{D}$ satisface las restricciones $f_i(x) \leq 0$, $i = 1, \ldots, m$ y $h_i(x) = 0$, $i = 1, \ldots, p$, entonces se dice que x es *factible*. Por lo tanto, si existe al menos un punto factible, el problema (B.1) es *factible*, en caso contrario se dice que el problema es *no factible*.

Un *valor óptimo* $p^\star$ del problema (B.1) está definido como

$$p^\star = inf\{f_0(x) \mid f_i(x), i = 1, \ldots, p, \;\; h_i(x) = 0, i = 1, \ldots, m\},$$

donde se permite que $p^\star$ tome los valores $+\infty$ y $-\infty$, $+\infty$ para el caso de un problema no factible y $-\infty$ en el caso de que exista una sucesión de puntos factibles x_k tal que

$$f_0(x_k) \to -\infty \;\text{ cuando }\; k \to \infty.$$

Decimos que $x^\star$ es un punto un *punto óptimo* si $x^\star$ es factible y $f_0(x^\star) = p^\star$. Al conjunto de todos los puntos óptimos se le nombra *conjunto óptimo* y de denota como

$$X_{opt} = \{x \mid f_i(x) \leq 0, i = i, \ldots, p, h_i(x) = 0, i = 1 \ldots, m, f_0(x) = p^\star\},$$

si el conjunto $X \neq \emptyset$ entonces se dice que el valor óptimo es alcanzado y el problema (B.1) tiene solución, en caso contrario no la tiene. Si $x \in X_{opt}$ entonces se dice que x es un *punto óptimo global* mientras que x es un *punto óptimo local* si existe $R > 0$ tal que

$$f_0(x) = inf\{f_0(z) \mid f_i(z) \leq 0, i = 1, \ldots, p, h_i(z) = 0, i = 1, \ldots, m, \| z - x \| < R\}.$$

Definición B.1.1. *Una matriz* $\boldsymbol{A}$ *de dimención* $n \times n$ *se dice que es* ***positiva definida*** *si* $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$*,* ***semipositiva definida*** *si* $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0$*,* ***negativa definida*** *si* $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} < 0$ *y* ***seminegativa definida*** *si* $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \leq 0$*, para todo vector* $\boldsymbol{x} \neq 0$*.*

Recordemos que si $f : \mathbb{R}^n \to \mathbb{R}$ tal que $f \in C^1$ entonces el *gradiente* de la función f está definido como

$$\mathbf{g(x)} = \nabla f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right]^T \tag{B.3}$$

y $Df(x) = \nabla f(x)^T$, además, si $f \in C^2$ el *Hessiano* de f, está definido como

$$\mathbf{H(x)} = \nabla^2 f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1^2} & \frac{\partial f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial f}{\partial x_1 \partial x_n} \\ \frac{\partial f}{\partial x_2^2} & \frac{\partial f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_n^2} & \frac{\partial f}{\partial x_n \partial x_2} & \cdots & \frac{\partial f}{\partial x_n^2} \end{bmatrix} \tag{B.4}$$

Propiedad B.1.1. ***Regla de la cadena:*** *Sean $f : \mathbb{R}^m \to \mathbb{R}^k$ y $g : \mathbb{R}^n \to \mathbb{R}^m$ funciones diferenciables, con $\boldsymbol{x} \in \mathbb{R}^n$, entonces*

$$D_{\boldsymbol{x}}(f \circ g) = Df(g(\boldsymbol{x}))Dg(x),$$

donde $Df(g(x)) \in \mathbb{R}^{k\times m}$, $Dg(x) \in \mathbb{R}^{m\times n}$ y $D_{\boldsymbol{x}}(f \circ g) \in \mathbb{R}^{k\times n}$ son matrices y $(f \circ g) : \mathbb{R}^n \to \mathbb{R}^k$.

Teorema B.1.1. *Si $\boldsymbol{x}^*$ es un punto optimo local y f es una función continuamente diferenciable en una vecindad abierta de $\boldsymbol{x}^*$ entonces*

$$\nabla f(\boldsymbol{x}^*) = 0.$$

Teorema B.1.2. *Si $\boldsymbol{x}^*$ es un punto mínimo local de f y $\nabla^2 f(\boldsymbol{x}^*)$ existe y es continua en una vecindad abierta de $\boldsymbol{x}^*$, entonces $\nabla f(\boldsymbol{x}^*) = 0$ y $\nabla^2 f(\boldsymbol{x}^*)$ es positiva semidefinida.*

Se pueden definir varias clases de problemas de optimización, las cuales están definidas por la forma de su función objetivo y sus restriciones. Un ejemplo muy importante y que está bastante estudiado es la clase de los problemas de optimización lineal o también conocindos como *problema de programación lineal*. A esta clase de de problemas se les caracteriza porque tanto su función objetivo f_0, como sus restricciones f_i, $i = 1, \ldots, m$ y h_i, $i = 1, \ldots, p$ son lineales. Recordemos que una función $f : \mathbb{R}^n \to \mathbb{R}$ es lineal si y solo si

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y),$$

donde $\alpha, \beta \in \mathbb{R}$ y $x, y \in \mathbb{R}^n$. En el caso en el que las funciones f_i, $i = 0, 1, \ldots, p$, h_i, $i = 1, \ldots, m$ no son lineales se dice que el problema de optimización no es lineal o que es un *problema de programación no lineal*.

Los problemas de programación no lineal desafortunadamente no son problemas fáciles de resolver, no existen métodos eficientes para aproximar sus soluciones. Sin embargo, existe otra clase de problemas de optimización en los cuales ha habido grandes aportaciones, estos son los *problemas de optomización convexa*, los cuales se caracterizan porque su función objetivo y restricciones son funciones convexas. Un ejemplo problema de optimización convexa es el ya bien conocido problema de los mínimos cuadrados.

B.2. Convexidad

Decimos que un conjunto $A \subseteq \mathbb{R}^n$ es *convexo* si $y \in A$, donde y está dado por la ecuación (B.5), $x_1, x_2 \in A$, con $x_1 \neq x_2$ y $\theta \in \mathbb{R}$ tal que $0 \leq \theta \leq 1$.

$$y = \theta x_1 + (1 - \theta)x_2 \tag{B.5}$$

Una función $f : \mathbb{R}^n \to \mathbb{R}$ es convexa si $\mathbf{dom}(f)$ es un conjunto convexo y si para toda $x, y \in \mathbf{dom}(f)$ existe $0 \leq \theta \leq 1$ tal que se cumple la siguiente ecuación:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y).$$

El concepto de convexidad se puede entender geométricamente en $\mathbb{R}^2$ como en la Figura B.1, es decir, que el segmento de recta que va de $(x, f(x))$ a $(y, f(y))$ está sobre la gráfica de la función $f(x)$. La función f se llama *cóncava* si $-f$ es convexa.

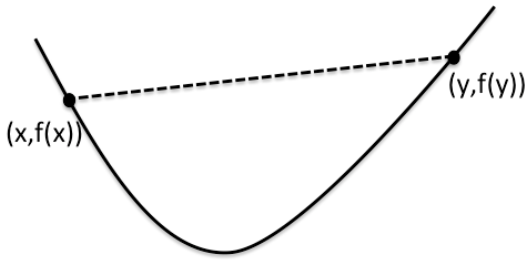

Figura B.1: En esta figura podemos ver un ejemplo de gráfico convexo en $\mathbb{R}^2$.

Teorema B.2.1. *Una función $f(x) \in C^2$ es convexa sobre un conjunto convexo, si y solo si, el Hessiano **H(x)** de f es positivo semidefinido.*

La importancia de los problemas de optimización convexa radica en que estos sólo tienen un óptimo global y son más extensos que el conjunto de problemas de optimización lineal.

B.3. Multiplicadores de Lagrange

El método de *Multiplicadores de Lagrange* es un método muy conocido y útil para encontrar extremos en problemas de optimización sujetos a restricciones, cabe mencionar que la dificultad de este método incrementa con el número de restricciones y la complejidad de su función objetivo y restricciones, sin embargo, este método se vuelve bastante amigable para el caso de problemas de programación convexa.

Consideremos el problema de optimización en su forma estandar

$$\begin{aligned} &\text{minimizar} \;\; f_0(x) \\ &\quad \text{sujeta a} \;\; f_i(x) \leq 0 \;\; i = 1, \ldots, m, \\ &\qquad\qquad h_i(x) = 0 \;\; i = 1, \ldots, p, \end{aligned} \tag{B.6}$$

con $x \in \mathbb{R}^n$, asumiendo que su dominio de optimización $\mathcal{D}$ es no vacío y $p^\star$ es el valor óptimo de (B.6) (sin asumir que el problema es convexo).

La *función Lagrangiana* o el *Lagrangiano* asociado al problema (B.6) también conocido como problema primal se define como la función:

$$L_P : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R},$$

la cual tiene la siguiente forma:

$$L_P(x, \lambda, \upsilon) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \upsilon_i h_i(x), \tag{B.7}$$

donde la variable $\lambda_i \geq 0$ es el multiplicador de Lagrange asociado a la i-ésima restricción de desigualdad $f_i(x) \leq 0$, de la misma forma, la variable $\upsilon_i \geq 0$ es el multiplicador de Lagrange asociado a la i-ésima restricción de igualdad $h_i(x) = 0$. En general los vectors $\lambda = (\lambda_1, \ldots, \lambda_m)$ y $\upsilon = (\upsilon_1, \ldots, \upsilon_m)$ son los multiplicadores de Lagrange asociados al problema (B.6).

Se define la función dual asociada al lagrangiano primal (B.7) como

$$L_D : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R},$$

que es el valor mínimo de la función Lagrangiana sobre x, para $\lambda \in \mathbb{R}$ y $\upsilon \in \mathbb{R}$, es decir,

$$L_d(\lambda, \upsilon) = \underset{x \in \mathcal{D}}{Inf} L_p(x, \lambda, \upsilon) = \underset{x \in \mathcal{D}}{Inf} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \upsilon_i h_i(x) \right).$$

El interés del problema dual asociado es que al encontrar su solución se puede obtener también la solución del problema primal, además que normalmente es más fácil resolver el problema dual.

B.4. Condiciones de Karush-Kuhn-Tucker (KKT)

Consideremos el problema de optimización (B.6) donde las funcines f_i, $i = 0, \ldots, p$, h_i, $i = 1, \ldots, m$ son diferenciables (no necesariamente convexas) y están definidas sobre un conjunto abierto. Las condiciones necesarias para que un punto $(x^\star, \lambda^\star, \upsilon^\star)$ sea valor óptimo (local) de la ecuación (B.7), es que se satisfagan las condiciones de *Karush-Kuhn-Tuker* (KKT).

$$\vec{\nabla} L_P(x, \lambda, \upsilon) = \vec{\nabla} f_0(x^\star) + \sum_{i=1}^{n} \lambda_i^\star \vec{\nabla} f_i(x^\star) + \sum_{i=1}^{p} \upsilon^\star \vec{\nabla} h_i(x^\star) = 0 \tag{B.8}$$

$$\lambda^\star f_i(x^\star) = 0, \quad i = 1, \ldots, p \tag{B.9}$$

$$\upsilon^\star h_i(x^\star) = 0, \quad i = 1, \ldots, m \tag{B.10}$$

La ecuación (B.8) es la primera condición de KKT y las ecuaciones (B.9) y (B.10) son conocidas como condiciones complementaria de KKT.

Para el caso en el que el problema (B.6) es convexo, las condiciones KKT no son solamente son condiciones necesarias sino suficientes.

Bibliografía

[1] Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer, 2017.

[2] Yakup Kara, Melek Acar Boyacioglu, and Ömer Kaan Baykan. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications*, 38(5):5311–5319, 2011.

[3] Yaser S Abu-Mostafa and Amir F Atiya. Introduction to financial forecasting. *Applied intelligence*, 6(3):205–213, 1996.

[4] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

[5] Chris M Bishop. Neural networks and their applications. *Review of scientific instruments*, 65(6):1803–1832, 1994.

[6] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[7] Wei Huang, Yoshiteru Nakamori, and Shou-Yang Wang. Forecasting stock market movement direction with support vector machine. *Computers & operations research*, 32(10):2513–2522, 2005.

[8] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[9] William Mendenhall, Richard L Scheaffer, and Dennis D Wackerly. *Estadística matemática con aplicaciones, Séptima Edición*. Grupo Editorial Iberoamérica,, 2008.

[10] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. springer, 1996.

[11] Peter J Brockwell, Richard A Davis, and Stephen E Fienberg. *Time series: theory and methods*. Springer Science & Business Media, 1991.

[12] Anil K Jain, Jianchang Mao, and KM Mohiuddin. Artificial neural networks: A tutorial. *Computer*, (3):31–44, 1996.

[13] Enrique Carmona Suárez. Tutorial sobre máquinas de vectores soporte (svm). *ResearchGate: Madrid, España. Recuperado de https://www. researchgate. net/publication/263817587_Tutorial_sobre_Maquinas_de_Vectores_ Soporte_SVM*, 2014.

[14] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[15] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[16] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[17] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.