**Marco teórico**

**Machine Learning como técnica de pronostico**

Catalina Patiño Forero

Wilfer Manuel Salas Gonzales

Las técnicas de pronóstico tales como el aporte de expertos en los temas requeridos, el estudio de datos históricos y modelos determinísticos o probabilistas tienen como objetivo encontrar un patrón de comportamiento que nos permita conocer el comportamiento de este en un futuro.

En un momento como lo es el actual en donde existe un flujo continuo de diferentes tipos de datos es necesario analizarlos con el fin de hallar información relevante sobre estos. Para el análisis de estas grandes cantidades de datos se utiliza la minería de datos la cual mediante una serie de algoritmos de inteligencia artificial y técnicas estadísticas se extrae la información de las bases de datos y se convierte en conocimiento que puede ser interpretado posteriormente. A esto también se le conoce como la técnica cuantitativa de predicción la cual requiere información para poder efectuar la predicción.

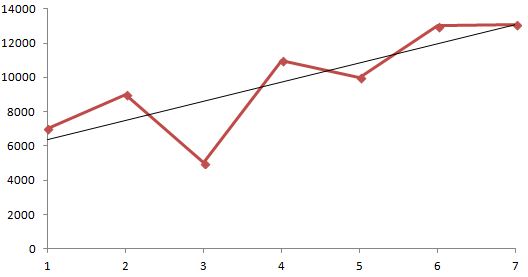
La técnica cualitativa es creada por expertos y se basa en el conocimiento del negocio. No existen datos, sino que se basa en la intuición de los expertos. Estas dos técnicas, la cualitativa y la cuantitativa son la base de los métodos de pronósticos.

1. Métodos por análisis de regresiones

La regresión tiene como fin predecir una media basándose en el conocimiento de otra y esto lo logra al crear una relación entre la variable de entrada y la variable de salida, para poder así predecir la salida cuando se presenta una entrada nueva.

* + Regresión lineal

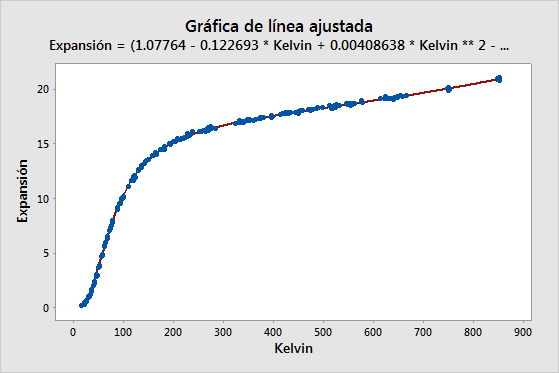
La regresión lineal se construye a partir de los puntos de entrada y en función a estos se traza una line que atraviesa los puntos con la mínima variación entre las observaciones y el punto predicho por la función.



Modelo de regresión lineal para un pronóstico de ventas. Tomado de <https://www.ingenieriaindustrialonline.com/herramientas-para-el-ingeniero-industrial/pron%C3%B3stico-de-ventas/regresi%C3%B3n-lineal/>

* + Regresión no lineal

La regresión no lineal genera una ecuación que describe la relación entre una variable continua y una o más variables de predicción.



Modelo de regresión no lineal que busca entender la relación que existe entre el coeficiente de expansión térmica del cobre y la temperatura en grados Kelvin. Tomado de <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/nonlinear-regression/before-you-start/example/>

1. Métodos basados en Instancias:

Son Modelos de Aprendizaje para problemas de decisión con instancias o ejemplos (muestras) de datos de entrenamiento que son importantes o requeridos por el modelo. También son llamados Algoritmos “Ganador se lleva todo” y aprendizaje basado-en-memoria en el que se crea un modelo a partir de una base de datos y se agregan nuevos datos comparando su similitud con las muestras ya existentes para encontrar “la mejor pareja” y hacer la predicción.

* + k-Nearest Neighbor:

El algoritmo clasifica cada dato nuevo en el grupo que corresponda, según tenga k vecinos más cerca de un grupo o de otro. Es decir, calcula la distancia del elemento nuevo a cada uno de los existentes, y ordena dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenecer. Este grupo será, por tanto, el de mayor frecuencia con menores distancias.

El K-NN es un algoritmo de aprendizaje supervisado, es decir, que a partir de un juego de datos inicial su objetivo será el de clasificar correctamente todas las instancias nuevas. El juego de datos típico de este tipo de algoritmos está formado por varios atributos descriptivos y un solo atributo objetivo (también llamado clase). El algoritmo K-NN y su importancia en el modelado de datos. [1]

* + Self-Organizing Map:

Un mapa auto-organizado (SOM por sus siglas en inglés) o un mapa auto-organizado de características (SOFM por sus siglas en inglés) es un tipo de red neuronal artificial (ANN por sus siglas en inglés), que es entrenada usando aprendizaje no supervisado para producir una representación discreta del espacio de las muestras de entrada, llamado mapa. Los mapas auto-organizados son diferentes de otras redes neuronales artificiales, en el sentido que estos usan una función de vecindad para preservar las propiedades topológicas del espacio de entrada. [2]

1. Modelos basados en Árbol de Decisión:

Modelan la toma de Decisión basado en los valores actuales (reales) de los atributos que tienen nuestros datos. Se utilizan sobre todo para clasificación de información, bifurcando y modelando los posibles caminos tomados y su probabilidad de ocurrencia para mejorar su precisión. Una vez armados, los arboles de decisión ejecutan muy rápido para obtener resultados. Los Algoritmos de árbol de decisión más usados son.

* + Arboles de Clasificación y Regresión:

Los árboles de clasificación y regresión son métodos que proporcionan modelos que satisfacen objetivos tanto predictivos como explicativos. Dos de los puntos fuertes de este método son, por un lado, la sencilla representación gráfica mediante árboles y, por otro, el formato compacto de las reglas de lenguaje natural. Se distinguen dos casos en que estas técnicas de modelado deben utilizarse: - Utilizar árboles de clasificación para explicar y predecir la pertenencia de los objetos (observaciones, individuos) a una clase, sobre la base de variables explicativas cuantitativas y cualitativas. - Utilizar un árbol de regresión para crear un modelo explicativo y predictivo para una variable cuantitativa dependiente basada en variables explicativas cuantitativas y cualitativas. [3]

* + Decisión de Árbol condicional

Un árbol de decisión1​ es un modelo de predicción utilizado en diversos ámbitos que van desde la inteligencia artificial hasta la Economía. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. [4]

1. Métodos basados en Bayesianos

Son algoritmos que utilizan explícitamente el Teorema de Bayes de probabilidad para problemas de Clasificación y Regresión

* + Naive Bayes:

Se trata de una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados. Constituye una técnica supervisada porque necesita tener ejemplos clasificados para que funcione, como ya veremos. Clasificador Naïve Bayes. [5]

* + Multinomial Naive Bayes:

En teoría de la probabilidad y minería de datos, un clasificador Bayesiano ingenuo es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de ingenuo. [6]

1. Métodos basados en Clustering:

Se utilizan para agrupar datos existentes de los que desconocemos sus características en común o queremos descubrirlas.  
Estos métodos intentan crear “puntos centrales” y jerarquías para diferenciar grupos y descubrir características comunes por cercanía.

* + K-Means:

K-means es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos.

La agrupación del conjunto de datos puede ilustrarse en una partición del espacio de datos en celdas de Voronoi.

El problema es computacionalmente difícil (NP-hard). Sin embargo, hay eficientes heurísticas que se emplean comúnmente y convergen rápidamente a un óptimo local. Estos suelen ser similares a los algoritmos expectation-maximization de mezclas de distribuciones gausianas por medio de un enfoque de refinamiento iterativo empleado por ambos algoritmos. Además, los dos algoritmos usan los centros que los grupos utilizan para modelar los datos, sin embargo, k-means tiende a encontrar grupos de extensión espacial comparable, mientras que el mecanismo expectation-maximization permite que los grupos tengan formas diferentes. [7]

* + K-Medians:

En estadística y minería de datos, la agrupación en k-medians es un algoritmo de análisis de conglomerados. Es una variación de clustering k-means donde en lugar de calcular la media de cada cluster para determinar su centroide, uno calcula el medio. Esto tiene el efecto de minimizar el error en todos los conglomerados con respecto a la métrica de distancia de 1 norma, a diferencia del cuadrado de la métrica de distancia de 2 normas (lo que hace k-means).

1. Métodos basados en Reducción de Dimensión:

Buscan explotar la estructura existente de manera no supervisada para simplificar los datos y reducirlos o comprimirlos.  
Son útiles para visualizar datos o para simplificar el conjunto de variables que luego pueda usar un algoritmo supervisado.

* + Principal Component Analysis

En estadística, el análisis de componentes principales (en español ACP, en inglés, PCA) es una técnica utilizada para describir un set de datos en términos de nuevas variables ("componentes") no correlacionadas. Los componentes se ordenan por la cantidad de varianza original que describen, por lo que la técnica es útil para reducir la dimensionalidad de un conjunto de datos.

Técnicamente, el ACP busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. Esta convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas componentes principales.

El ACP se emplea sobre todo en análisis exploratorio de datos y para construir modelos predictivos. El ACP comporta el cálculo de la descomposición en autovalores de la matriz de covarianza, normalmente tras centrar los datos en la media de cada atributo.

Debe diferenciarse del análisis factorial con el que tiene similaridades formales y en el cual puede ser utilizado como un método de aproximación para la extracción de factores. [8]

1. Métodos por Machine Learning:

Son métodos computacionales que, en base a la información recolectada que se usa para entrenamiento, mejoran los resultados de las predicciones, logrando así un resultado más exacto.

* + Ridge Regresión:

Es un método para regresión que incluye una función de minimización del error en términos de penalización.

* + Kernel Ridge Regression:

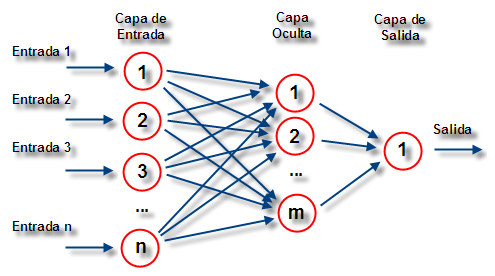
Analiza los datos de entrada que se consideran como no lineales obteniendo errores de ajustes muchos menores que los conseguidos en la entrada inicial. Se utiliza cuando puede no haber una relación no lineal en los datos de entrenamiento

* + Redes Neuronales:

Son modelos de aprendizaje que se componen de una serie de unidades llamadas neuronas, que simulan la funcionalidad de las neuronas biológicas.

Una neurona artificial recibe una serie de entradas a través de interconexiones y con estas emite unas salidas que se dan a través de tres funciones:

* + 1. Función de propagación la cual suele calcular combinaciones para cada entrada modificada por el peso de su interconexión.
    2. Función de activación tiene como parámetro de entrada a la anterior, aunque es opcional utilizarla.
    3. Función de transferencia salida del valor devuelto por la función



Ejemplo de red neuronal artificial. Tomada de <http://www.cs.us.es/~fsancho/?e=72>

Las redes neuronales son una gran herramienta de reconocimiento de patrones, clasificación y agrupación de datos debido a su gran capacidad de aprender de la experiencia, a la capacidad de poder reconocer datos que habían sido ignorados y organizarlos, analizar y utilizar información con alto nivel de ruido y que está incompleta.

El entrenamiento de las redes neuronales es clasificado entre entrenamiento supervisado, entrenamiento no supervisado y entrenamiento por refuerzo. El entrenamiento supervisado hace uso de ejemplos de cómo debería ser el comportamiento esperado en base a entradas específicas. El entrenamiento no supervisado no requiere ejemplo de comportamientos adecuados, sino que aprenden a clasificar espacios de entrada determinados mediante un proceso de auto organización. Los algoritmos de entrenamiento por refuerzo no reciben tampoco ejemplos de comportamientos, pero en este caso hay una serie de calificaciones por las acciones tomadas lo cual hace que la meta de la red sea la maximización de calificaciones.

Uno de los modelos más usados en redes neuronales artificiales es el Perceptrón Multicapa que es está formado por múltiples capas de neuronas. Las redes neuronales función base radial se caracteriza porque en su diseño conta de neuronas activadas mediante una función de activación radial de carácter no lineal.

Algunos de los métodos de validación para machine learning son el k-iteraciones y el K-folds que dividen el conjunto de datos en k subconjuntos, de esta forma toma los subconjuntos como datos de validación y el resto como datos de entrenamiento.

# Bibliografía

|  |  |
| --- | --- |
| [1] | S. Ruiz, «Analitica Web,» [En línea]. Available: http://www.analiticaweb.es/algoritmo-knn-modelado-datos/. |
| [2] | «Wikipecia,» [En línea]. Available: https://es.wikipedia.org/wiki/Mapa\_autoorganizado. [Último acceso: 0429 2018]. |
| [3] | «Wikipedia,» Xistat, [En línea]. Available: https://www.xlstat.com/es/soluciones/funciones/arboles-de-clasificacion-y-de-regresion. [Último acceso: 29 04 2018]. |
| [4] | «Wikipedia,» Wikipedia, [En línea]. Available: https://es.wikipedia.org/wiki/%C3%81rbol\_de\_decisi%C3%B3n. [Último acceso: 29 04 2018]. |
| [5] | C. CH, «Neiva bayers,» [En línea]. Available: http://naivebayes.blogspot.com.co/. [Último acceso: 20 04 2018]. |
| [6] | «Wikipedia,» [En línea]. Available: https://es.wikipedia.org/wiki/Clasificador\_bayesiano\_ingenuo. [Último acceso: 27 04 2018]. |
| [7] | «Wikipedia,» [En línea]. Available: https://es.wikipedia.org/wiki/K-means. [Último acceso: 27 04 2018]. |
| [8] | «Wikipedia,» [En línea]. Available: https://es.wikipedia.org/wiki/An%C3%A1lisis\_de\_componentes\_principales. [Último acceso: 28 04 2018]. |
| [9] | S. Ruiz, El algoritmo K-NN y su importancia en el modelado de datos, 2017. |

Referencias:

Jamie Areli-Toral Barrera *Redes Neuronales.* Disponible en: <http://www.cucei.udg.mx/sites/default/files/pdf/toral_barrera_jamie_areli.pdf>

Ana María Yepes de Castaño y Martha Eugenia Alvarez VillaP*ronóstico mediante modelos probabilísticos: una herramienta en la toma de decisiones.*  Editorial de EAFIT. Disponible en: <http://publicaciones.eafit.edu.co/index.php/revista-universidad-eafit/article/download/1150/1040/>

Pedro Domingos *A Few Useful Things to Know about Machine Learning.* Disponible en:<https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf>

ARTHUR E. HOERL AND ROBERT W. KENNARD *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Disponible en: <https://pdfs.semanticscholar.org/910e/d31ef5532dcbcf0bd01a980b1f79b9086fca.pdf>

Max Weding *Kernelidge Regression.* Disponible en: <https://www.ics.uci.edu/~welling/classnotes/papers_class/Kernel-Ridge.pdf>