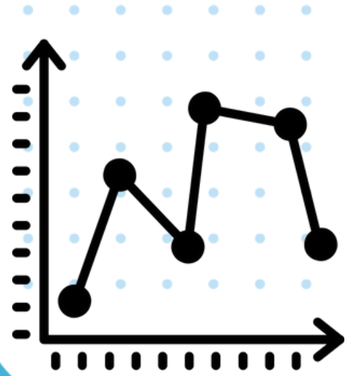
COLOMBIA POTENCIA DE LA VIDA

TIC

Lección 3: Modelo de regresión lineal



TALENTO TECH

AZ | PROYECTOS EDUCATIVOS

UTP Universidad Tecnológica de Pereira

La regresión es una técnica estadística para modelar la relación entre una variable dependiente y una o múltiples variables independientes. Lo que se pretende es encontrar un modelo matemático que mejor se ajuste a los datos observados. Es importante recalcar que los datos son observaciones, o mediciones.

Cuando se habla de una observación o una medición siempre se debe pensar que hay ruido asociado al proceso, por lo que los datos están desplazados de su coordenada real.

Sin embargo, el valor real de las mediciones no se puede conocer, solamente se puede estimar. Por este motivo, no es posible encontrar un modelo que explique perfectamente los datos, sino que, se debe recurrir al ajuste de variables que tengan en cuenta el ruido inherente a las medidas y minimicen el error al intentar encontrar un modelo que explique los datos.

Cuando los datos se explican con un modelo de la forma:

$$y = mx + b$$

Se habla de una regresión lineal, es decir, un modelo de línea recta. En este caso particular es de interés conocer los parámetros m y b que modelan una única recta en el espacio de dos dimensiones. El ajuste debe permitir que, dado un punto independiente x o medición se pueda encontrar el resultado y con el modelo de la forma más aproximada a la realidad.

Cuando se intenta resolver el problema de regresión, lo que se hace es encontrar un modelo matemático y un conjunto de parámetros que debe tener ese modelo tal que el error entre los puntos de datos del modelo y los datos que se intentan modelar sea el mínimo.

Es decir, el problema de regresión se puede resolver minimizando el error entre dos valores, los puntos de datos del modelo y los datos que se intentan modelar. Por ejemplo, si se tienen los datos de el precio de las casas y se quiere encontrar el precio que puede tener una casa con base en el número de habitaciones, la ubicación, el tamaño, entre otros. La tarea de encontrar el precio dados los datos es resolver un problema de regresión.

Existen técnicas de modelado de datos como el ajuste de mínimos cuadrados (lineales, no lineales), modelos de máxima verosimilitud, regresión linear bayesiana, y modelos más elaborados como las máquinas de soporte vectorial y las redes neuronales.

El análisis de regresión lineal se utiliza para estimar el valor de una variable en relación con el valor de otra variable. La variable que desea predecir se llama variable dependiente. La variable que utiliza para predecir el valor de la otra variable se llama variable independiente.

Los modelos de regresión lineal son relativamente simples y proporcionan una fórmula matemática fácil de interpretar que puede producir predicciones.

La regresión lineal se puede aplicar a una variedad de campos de los estudios académicos y empresariales. Debido a que la regresión lineal es un procedimiento estadístico bien establecido, es fácil entender las propiedades de un modelo y entrenar diferentes tipos de modelos según sea el caso.

Regresión lineal simple

El modelo de regresión lineal simple es un método utilizado para estimar la variable dependiente con la ayuda de la variable independiente cuando existe una **relación lineal** entre la variable independiente y la variable dependiente. Es decir, los datos se pueden modelar como una línea recta.

En su forma canónica el modelo se puede escribir como:

$$y = A + Bx + e$$

En donde la pareja de datos (x, y) son los puntos de datos (provienen del conjunto de datos). Los coeficientes A y B modelan el intercepto de la recta con el cero y la pendiente de la recta respectivamente y el valor ‘e’ se agrega para modelar el error del modelo. Nota: Ya que se trabajan con datos medidos, siempre habrá un error asociado a los datos, generado en la adquisición o en el procesamiento, que debe modelarse.

Regresión polinómica

La regresión polinómica se trata de encontrar un modelo que permita explicar datos no lineales mediante un polinomio. Para esto, se selecciona el grado del polinomio a ajustar y se entrenará el modelo, es decir, se encontrarán los coeficientes que mejor se ajusten a los datos. El modelo polinómico es de la forma:

$$y = A + Bx + Cx^2 + Dx^3 + \ldots + W_n x^n + e$$

En donde los coeficientes A, B, C… W son el grado del polinomio y son los valores para ajustar de acuerdo con los datos. La variable x representa los valores independientes y la variable y los valores dependientes (x y y provienen de los datos).