Scikit Learn

Regresión

CertiDevs

Índice de contenidos

1. Regresión lineal simple

La **regresión lineal simple** es un método estadístico que permite estudiar la relación entre dos variables cuantitativas. En otras palabras, es una técnica que ayuda a entender cómo una variable (llamada **variable dependiente**) cambia en función de otra variable (llamada **variable independiente**).

La regresión lineal simple se basa en la idea de que existe una **relación lineal** entre estas dos variables, es decir, que se pueden representar mediante una **línea recta**.

Para entender mejor este concepto, imaginemos que se quiere analizar la relación entre las horas de estudio y las calificaciones obtenidas en un examen. En este caso, las horas de estudio serían la variable independiente (X) y las calificaciones obtenidas serían la variable dependiente (Y). La regresión lineal simple nos permitiría determinar si existe una relación lineal entre estas dos variables y, de ser así, cuantificar esa relación mediante una ecuación matemática.

La ecuación matemática que describe la relación lineal entre las dos variables se conoce como ecuación de regresión lineal simple y tiene la siguiente forma:

```
Y = a + bX
```

Donde:

- `Y` es la variable dependiente (en nuestro ejemplo, las calificaciones obtenidas).
- `X` es la variable independiente (en nuestro ejemplo, las horas de estudio).
- `a` es el punto de intersección con el eje Y, también conocido como ordenada al origen. Este valor representa la calificación que se obtendría si no se hubiera estudiado nada (X = 0).
- `b` es la pendiente de la línea, también conocida como coeficiente de regresión. Este valor indica cómo cambia la variable dependiente (Y) en función de la variable independiente (X). En nuestro ejemplo, representaría el aumento en la calificación por cada hora adicional de estudio.

El **objetivo** de la regresión lineal simple es encontrar los valores de a y b que mejor se ajusten a los datos observados. Para ello, se utiliza el método de los mínimos cuadrados, que consiste en minimizar la suma de las diferencias al cuadrado entre los valores observados de la variable dependiente (Y) y los valores estimados por la ecuación de regresión lineal simple.

Una vez que se han encontrado los valores de `a` y `b`, se puede utilizar la ecuación de regresión lineal simple para hacer predicciones sobre la variable dependiente (Y) a partir de nuevos valores de la variable independiente (X). Por ejemplo, si se sabe cuántas horas ha estudiado un estudiante, se podría predecir su calificación en el examen utilizando la ecuación de regresión lineal simple.

Es importante tener en cuenta que la regresión lineal simple tiene ciertas limitaciones y supuestos. Entre ellos, se asume que existe una relación lineal entre las variables, que los errores en las mediciones son independientes y que tienen una distribución normal. Además, la regresión lineal simple solo puede analizar la relación entre dos variables, por lo que no es adecuada para estudiar relaciones más complejas que involucren múltiples variables. En estos casos, se pueden utilizar técnicas más avanzadas, como la **regresión lineal múltiple**.

En resumen, la regresión lineal simple es una herramienta estadística que permite analizar y cuantificar la relación lineal entre dos variables cuantitativas. A través de la ecuación de regresión lineal simple, es posible hacer predicciones sobre la variable dependiente a partir de la variable independiente y entender cómo una variable influye en la otra. Sin embargo, es importante tener en cuenta sus limitaciones y supuestos para garantizar la validez de los resultados obtenidos.

2. Regresión lineal múltiple

La **regresión lineal múltiple** es una extensión de la regresión lineal simple que permite analizar la relación entre una variable dependiente y varias variables independientes.

Al igual que en la regresión lineal simple, se busca establecer una relación lineal entre las variables, pero en este caso, se considera el efecto combinado de **múltiples variables independientes** sobre la variable dependiente.

La regresión lineal múltiple es especialmente útil cuando se desea estudiar el impacto de diferentes factores en un resultado específico.

La ecuación matemática que describe la relación lineal en la regresión lineal múltiple tiene la siguiente forma:

```
Y = a + b1X1 + b2X2 + ... + bnXn
```

Donde:
- `Y` es la variable dependiente, es decir, la variable que se quiere predecir o explicar.
- `X1, X2, ⋯, Xn` son las variables independientes, es decir, las variables que se utilizan para predecir o explicar la variable dependiente.
- `a` es el punto de intersección con el eje Y, también conocido como ordenada al origen. Este valor representa la predicción de Y cuando todas las variables independientes son iguales a cero.
- `b1, b2, ⋯, bn` son los coeficientes de regresión, que indican cómo cambia la variable dependiente (Y) en función de cada variable independiente (X1, X2, ..., Xn), manteniendo constantes las demás variables independientes.

El objetivo de la regresión lineal múltiple es encontrar los valores de `a, b1, b2, ⋯, bn` que mejor se ajusten a los datos observados. Para ello, se utiliza el método de los mínimos cuadrados, que consiste en minimizar la suma de las diferencias al cuadrado entre los valores observados de la variable dependiente (Y) y los valores estimados por la ecuación de regresión lineal múltiple.

Una vez que se han encontrado los valores de `a, b1, b2, ⋯, bn`, se puede utilizar la ecuación de regresión lineal múltiple para hacer predicciones sobre la variable dependiente (Y) a partir de nuevos valores de las variables independientes (X1, X2, ..., Xn). Además, los coeficientes de regresión permiten evaluar la importancia relativa de cada variable independiente en la predicción de la variable dependiente.

Al igual que en la regresión lineal simple, la regresión lineal múltiple tiene ciertas **limitaciones** y **supuestos**. Entre ellos, se asume que existe una relación lineal entre las variables, que los errores en las mediciones son independientes y que tienen una distribución normal. Además, es importante tener en cuenta la posibilidad de **multicolinealidad**, que ocurre cuando dos o más

variables independientes están altamente correlacionadas, lo que puede dificultar la interpretación de los coeficientes de regresión y afectar la precisión de las predicciones.

En resumen, la regresión lineal múltiple es una herramienta estadística que permite analizar y cuantificar la relación lineal entre una variable dependiente y varias variables independientes. A través de la ecuación de regresión lineal múltiple, es posible hacer predicciones sobre la variable dependiente a partir de las variables independientes y entender cómo diferentes factores influyen en el resultado de interés. Sin embargo, es importante tener en cuenta sus limitaciones y supuestos para garantizar la validez de los resultados obtenidos.

2.1. Ejemplo

Imaginemos que una empresa de bienes raíces quiere predecir el precio de venta de una casa (variable dependiente) en función de diferentes características de la casa (variables independientes), como el tamaño, la cantidad de habitaciones, la antigüedad y la distancia al centro de la ciudad. La regresión lineal múltiple puede ser utilizada para analizar cómo estas características influyen en el precio de venta y para hacer predicciones sobre el precio de casas no vendidas aún.

Supongamos que se tiene una muestra de casas vendidas con información sobre el precio de venta, el tamaño (en metros cuadrados), la cantidad de habitaciones, la antigüedad (en años) y la distancia al centro de la ciudad (en kilómetros). A partir de estos datos, se puede ajustar un modelo de regresión lineal múltiple con la siguiente ecuación:

```
Precio = a + b1(Tamaño) + b2(Habitaciones) + b3(Antigüedad) + b4(Distancia al centro)
```

Una vez ajustado el modelo, se obtienen los coeficientes de regresión (a, b1, b2, b3 y b4) que mejor se ajustan a los datos observados. Estos coeficientes permiten interpretar cómo cada característica de la casa influye en el precio de venta. Por ejemplo, si b1 es positivo, esto indica que el precio de la casa tiende a aumentar a medida que aumenta el tamaño. Si b3 es negativo, esto sugiere que el precio de la casa tiende a disminuir a medida que aumenta la antigüedad.

Además, el modelo de regresión lineal múltiple puede ser utilizado para predecir el precio de venta de casas no vendidas aún. Por ejemplo, si se tiene información sobre una casa con un tamaño de 150 metros cuadrados, 3 habitaciones, 10 años de antigüedad y a 5 kilómetros del centro de la ciudad, se puede utilizar la ecuación de regresión lineal múltiple para estimar su precio de venta:

```
Precio estimado = a + b1(150) + b2(3) + b3(10) + b4(5)
```

Cabe destacar que, aunque la regresión lineal múltiple puede ser útil para analizar y predecir el precio de venta de una casa en función de sus características, es importante tener en cuenta sus limitaciones y supuestos, así como la posibilidad de que otros factores no incluidos en el modelo puedan influir en el precio de venta.