Material Imprimible

Curso de Data Science

Módulo 3 - Regresión

Contenidos:

- ¿Qué es una regresión?
- Algoritmos de regresión
- Regresión lineal
- Métricas de evaluación

¿Qué es una regresión?

La regresión es un método estadístico utilizado en el análisis de datos para investigar la relación entre una variable independiente (conocida como variable predictora) y una variable dependiente (también conocida como variable target). El objetivo de la regresión es encontrar una función matemática que pueda predecir el valor de las variables dependientes en función de los valores de la variable independiente.

La regresión es una técnica muy utilizada en el campo del Machine Learning, donde se utiliza para construir modelos predictivos que puedan predecir el valor de una variable de interés en función de otras variables. Por ejemplo, se puede utilizar la regresión para predecir el precio de una casa en función de su tamaño, ubicación y características.

Algoritmos de regresión

Los tipos de algoritmos más famosos y utilizados son:

- Regresión Lineal: Este es el algoritmo más común utilizado en el análisis de regresión. La regresión lineal se utiliza para modelar la relación lineal entre una variable independiente y una variable dependiente. Este modelo utiliza la ecuación de una línea recta para hacer predicciones.

- Regresión Logística: Este algoritmo se utiliza para predecir una variable de respuesta categórica. La regresión logística se utiliza a menudo en la clasificación binaria.

- Regresión de Árbol de Decisión: Este algoritmo utiliza un enfoque basado en árbol para la regresión. Divide los datos en pequeños subconjuntos utilizando una serie de condiciones para modelar la relación entre las variables independientes y dependientes.

- Redes Neuronales: Las redes neuronales son un conjunto de algoritmos que se utilizan para el aprendizaje profundo. Estos algoritmos se utilizan para la regresión y para la clasificación y pueden modelar relaciones no lineales entre las variables independientes y dependientes.

Regresión lineal

Link oficial: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

La regresión lineal es un algoritmo muy común utilizado en el análisis de regresión para modelar la relación lineal entre una variable independiente y una variable dependiente. La biblioteca scikit-learn en Python proporciona una implementación de regresión lineal llamada Linear Regression.

LinearRegression en scikit-learn es una clase que se utiliza para crear modelos de regresión lineal. Se puede utilizar para ajustar un modelo de regresión lineal a un conjunto de datos y hacer predicciones basadas en el modelo. La regresión lineal utiliza la técnica de mínimos cuadrados ordinarios (OLS) para encontrar la mejor línea recta que se ajuste a los datos.

El proceso metodológico para utilizar la regresión lineal, tal como vimos en las clases, es el siguiente:

- Importar la clase LinearRegression de la biblioteca scikit-learn.
- Preparar los datos: Esto incluye dividir los datos en conjuntos de entrenamiento y prueba, y asegurarse de que los datos estén limpios y sean adecuados para la regresión lineal. También se aconseja hacer separación en Train y Test, para hacer una evaluación independiente.
- Crear una instancia del modelo de regresión lineal: Esto se hace utilizando la clase LinearRegression.
- Entrenar el modelo: Esto implica ajustar la línea recta a los datos de entrenamiento utilizando el método fit() de la clase LinearRegression.
- Hacer predicciones: Utilizar el método predict() del modelo para hacer predicciones en los datos de test.
- Evaluar el modelo: Esto implica comparar las predicciones del modelo con los valores reales de la variable dependiente en los datos de prueba utilizando métricas como el R2 o MAE

Métricas de evaluación

Existen distintos tipos de métricas, según el dataset, el problema y la tolerancia al error que tengamos.

Dos de las más utilizadas son el R2 y el MAE.

- R2 es un coeficiente que indica cuanta volatilidad de los datos pueden ser explicados por el modelo. Este dato va de 0 a 1, y cuanto más se acerque a 1, mejor será el algortimo que hayamos entrenado.

- MAE es el valor absoluto medio mide la media de las diferencias absolutas entre los valores predichos y los valores reales. Es muy util para dataset donde los targets no estén normalmente distribuidos.