BLOQUE 1

MÓDULO:
TÉCNICAS DE MACHINE LEARNING

ALGORITMOS SUPERVISADOS

LORENZO MARTÍNEZ MANERO

Ingeniero Industrial Superior por la Universidad Politécnica de Valencia.

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS ESPECÍFICOS

TEMA 1. ÁRBOLES DE DECISIÓN

- Comprender su funcionamiento interno.
- Entender la diferencia entre los dos tipos de cálculo de "impurezas": Gini y Entropy.
- Saber cómo utilizar los parámetros del algoritmo para evitar el overfitting.
- Saber cuáles son las funciones de coste a minimizar en Clasificación y en Regresión.
- Realizar y comprender los grafos.
- Aplicar este tipo de algoritmos a diferentes tipos de datos usando la librería Scikit-Learn de Python.
- Comprender cuándo vale la pena usarlos, sus ventajas y sus desventajas.

TEMA 2. KNN (K-NEAREST-NEIGHBORS)

- Comprender su funcionamiento interno.
- Saber cómo utilizar el parámetro K del algoritmo para evitar el overfitting y el underfitting.
- Aplicar este tipo de algoritmos a diferentes tipos de datos usando la librería Scikit-Learn de Python.
- Aprender a realizar el escalado antes de aplicar el algoritmo.
- Conocer algunos tipos de cálculo de distancias.
- Comprender cuándo vale la pena usarlos, sus ventajas y sus desventajas.

TEMA 3. NAÏVE-BAYES

- Comprender su funcionamiento interno.
- Saber cómo utilizar el parámetro alpha del algoritmo para mejorar su "performance".
- Conocer los tres tipos de algoritmos más usados basados en Naive-Bayes.
- Saber cuándo es más conveniente usar cada uno de ellos.
- Aplicar este tipo de algoritmos a diferentes tipos de datos usando la librería Scikit-Learn de Python.
- Comprender cuándo vale la pena usarlos, sus ventajas y sus desventajas.
- Aprender a trabajar (de manera introductoria) con datos de tipo texto.

ALGORITMOS SUPERVISADOS

En el mundo del Aprendizaje Automático (Machine Learning), existen muchos algoritmos, con diferentes fines y diferentes formas de funcionamiento. La forma más genérica de dividirlos es la siguiente:

- **Algoritmos supervisados**

  Son aquellos que tratan de realizar una predicción/estimación de valores de una variable objetivo:

  - **Clasificación:** cuando la variable objetivo es discreta (spam/no-spam, perro/gato, etc.).
  - **Regresión:** cuando la variable objetivo es continua (precios de productos, etc.)

- **Algoritmos no-supervisados:**

  Son aquellos que no tratan de realizar una predicción/estimación de valores de una variable objetivo:

  - **Clustering** (kmeans, clustering jerárquico, etc.).
  - **Reducción de dimensionalidad** (PCA, etc.).

En este tema, vamos a ver tres algoritmos supervisados en tres temas diferentes:

- **Árboles de decisión:** que pueden trabajar tanto en Clasificación como en Regresión.
- **KNN:** que, también, pueden trabajar en Clasificación y Regresión.
- **Naive Bayes:** aunque existan variantes basados en Bayes (como Bayesian Ridge Regression), el algoritmo que vamos a ver sólo trabaja en modo Clasificación.

A partir de este punto, vamos a trabajar con Colab donde alternaremos teoría con ejemplos programados en Python.

IDEAS CLAVE

TEMA 1: ÁRBOLES DE DECISIÓN

- Los Árboles de decisión son unos algoritmos muy utilizados hoy día debido a la posibilidad de entenderlos: WhiteBox.
- Deben ser bien parametrizados para regularizarlos y evitar overfitting.
- Son capaces de darnos información sobre la importancia de las diferentes variables predictoras, lo cual ayuda a entender mejor el problema que se trata de resolver.
- Son la base de los Algoritmos ensamblados de la segunda parte del módulo, que tan buen resultado dan en muchos contextos dentro del mundo del dato.

TEMA 2: KNN (K-NEAREST-NEIGHBORS)

- El principal aspecto a resaltar es su sencillez de funcionamiento y de aplicación, que nos ayuda, en muchas ocasiones, a usarse como un “baseline” a partir del cual ir mejorando.
- Es muy importante el escalado o estandarización de los datos para su buen funcionamiento.
- Pueden ser usados tanto para supervisado en clasificación y regresión como para el cálculo de distancias entre todos los puntos de un dataset.
- Tiene una gran gama de cálculo de distancias que puede ser interesante conocer según el caso de uso.

TEMA 3: NAIVE-BAYES

- Se trata de algoritmos que funcionan razonablemente bien cuando disponemos de muchas variables predictoras o cuando los datos tienen muchos ceros (sparse data).
- Son ideales como baseline en la clasificación de textos.
- Es interesante saber distinguir el funcionamiento de los tres algoritmos que se ven en el tema y cuándo es mejor usar cada uno de ellos.