SGDI: Sistema de Gestión de Datos e Información
SparkML - Practica 2

AUTORES: Javier Cortés Tejada, Aitor Cayón Ruano

20 de noviembre de 2018

Javier Cortés Tejada y Aitor Cayón Ruano declaramos que esta solución es fruto exclusivamente de nuestro trabajo personal. No hemos sido ayudados por ninguna otra persona ni hemos obtenido la solución de fuentes externas, y tampoco hemos compartido nuestra solución con nadie. Declaramos además que no hemos realizado de manera deshonesta ninguna otra actividad que pueda mejorar nuestros resultados ni perjudicar los resultados de los demás.

1. Pipeline de procesado del dataframe

Etapas:

- Transformación de los atributos categóricos en continuos.
- Transformación de la clase en valores continuos.
- Ensamblador para combinar los atributos en una sola columna.

Por defecto, la columna con la clase debe renombrarse a "label" para su utilización en los clasificadores.

Por defecto, la columna con los atributos agregados debe renombrarse a "features" para su utilización en los clasificadores.

Para poder utilizar en el pipeline los transformadores mencionados anteriormente los agrupamos en una única lista, correspondiéndose cada uno con una etapa de ejecución en el pipeline.

Transformamos los dataframes de entrenamiento y testeo y les damos el formato necesario para el entrenamiento de los modelos y la posterior clasificación.

2. Pipeline de clasificación

Etapas:

- Dado un clasificador, entrena un modelo con el dataframe de entrenamiento y clasifica el dataframe de prueba.

Creamos una lista de clasificadores sobre la que iteramos.

Creamos un evaluador para analizar los resultados obtenidos.

En cada iteración creamos un pipeline para entrenar un modelo con el correspondiente clasificador, clasificamos el dataframe de testeo y comprobamos la precisión obtenida con el evaluador.

3. Clasificadores

- Logistic Regression: 0.797
- Decision Tree Classifier: 0.802
- Random Forest Classifier: 0.807
- GBT Classsifier: 0.830
- Linear SVC: 0.771
- Naive Bayes: 0.774