Taller 4 – Clasificación en Spark

Prof. Fabio González
Diplomado Big-Data 2016
Universidad Nacional de Colombia

NOTA: *Lea el taller completamente antes de empezar a desarrollarlo, así tendrá una idea clara de lo que toca hacer.*

En este taller se aplicarán técnicas de clasificación sobre conjuntos de datos de la literatura.

1. Descargue el conjunto de datos nba.csv desde https://github.com/eduarc/Diplomado/blob/master/nba.csv. Lea la descripción del conjunto de datos que se encuentra en el contenido del archivo.

2. Cargue el conjunto de datos en Databricks. Tenga en cuenta que los campos del archivo están delimitados por espacios.

3. Entrene un árbol de decisión:

    a. Haga una partición del conjunto de datos usando muestreo, en 70% para entrenamiento y 30% para test.
    b. Entrene el modelo.
    c. Aplique el modelo al conjunto de test.
    d. Mida el desempeño del modelo calculando exactitud, medida F1 e índice de recuperación (recall).
    e. Genere la matriz de confusión.
    f. Interprete el modelo obtenido:
        i. ¿Cuál el el atributo más discriminante? ¿Tiene sentido? De una explicación a partir del conocimiento del problema.
        ii. Genere 3 diferentes reglas de clasificación a partir del árbol. Explíquelas.

4. Complejidad del modelo y sobre-aprendizaje:

    a. Modifique el modelo anterior para que también calcule el desempeño en el conjunto de entrenamiento.
    b. Haga diferentes pruebas cambiando la profundidad máxima del árbol en el operador *Decisión Tree*. Pruebe los valores 1, 2, … ,10.
    c. Grafique la profundidad del árbol contra la evolución del error de entrenamiento y el error de prueba.
    d. De acuerdo con la gráfica, ¿Cuál sería un buen valor de profundidad para el árbol?

5. Comparación de modelos:

    a. Usando los datos de la NBA, entrene un modelo de clasificación Naïve Bayes y evalúelo usando validación cruzada con 10 pliegues (K-Fold).
    b. Entrene un árbol de decisión y evalúelo usando validación cruzada con 10 pliegues (K-folds)
    c. ¿Cuál de los dos modelos es mejor?

6. Cargue el conjunto de datos credit-german.csv a Databricks desde https://github.com/eduarc/Diplomado/blob/master/credit-german.csv. Tenga en cuenta que los campos del archivo están delimitados por ";".

    a. Realice el pre-procesamiento necesario del conjunto de datos.
    b. Haga una partición del conjunto de datos usando muestreo, en 70% para entrenamiento y 30% para test.
    c. Entrene un modelo RandomForest utilizando validación cruzada con 10 pliegues (K-fold).
    d. Aplique el modelo obtenido al conjunto de test.
    e. Mida el desempeño del modelo calculando exactitud, medida F1 e índice de recuperación (recall).
    f. Genere la matriz de confusión.

7. Entregables:

    a. Para los puntos donde sea necesaria realizar una descripción, realícela en una celda de tipo markdown o en un documento PDF aparte.
    b. Descargue el notebook completado en formato IPython Notebook y comprímalo junto con los anexos como nombre_apellido_tallerClasificacion.zip
    c. Cargue el archivo al siguiente enlace Dropbox: https://www.dropbox.com/request/MNgw6TO7cHcVISQFyzK5

Material de Apoyo:

- Talleres y demos desarrollados en la sesiones anteriores.
- http://spark.apache.org/docs/latest/ml-features.html
- https://spark.apache.org/docs/latest/ml-classification-regression.html
- https://spark.apache.org/docs/latest/mllib-evaluation-metrics.html
- ¡Google!