Técnicas Escalables de Análisis de Datos en entornos Big Data: Clasificadores

Curso 2025-26. 21 de octubre del 2025

Entregable Individual: Selección sencilla de un modelo de clasificación

Contenidos

Enunciado

Este es un **trabajo individual** donde cada estudiante trabajará con los conjuntos de entrenamiento y prueba elaborados por su grupo en la primera entrega del proyecto software.

El ejercicio consiste en entrenar varios árboles de decisión mediante la selección de distintos valores de los parámetros *maxDepth y maxBins*, buscando el árbol que cometa un menor error, dejando los restantes parámetros con su valor por defecto. Debéis explorar, al menos, 9 combinaciones distintas de valores válidos de esos parámetros y razonar sobre dicho proceso de exploración. Todo el proceso de exploración debe ser justificado en una breve memoria en PDF que debéis adjuntar.

El error de resubstitución (error sobre el conjunto de entrenamiento) no es una buena estimación del error. **Y el conjunto de prueba no podéis utilizarlo para seleccionar el mejor valor de ningún parámetro**.

Lo que si podéis hacer es crear una segunda partición sobre el conjunto de entrenamiento, creando un segundo conjunto de entrenamiento (para crear árboles) y un conjunto de validación (para estimar tasas de error). Con estos conjuntos debéis explorar las distintas combinaciones de los parámetros *maxDepth y maxBins*. Finalmente, con los valores finales que consideréis mejor para esos parámetros, creáis el modelo final con todo el conjunto de entrenamiento y los restantes parámetros con su valor por defecto.

Para terminar, evaluáis el modelo final sobre el conjunto de prueba utilizando las métricas de ML/MLlib.

Para simplificar la elaboración de los modelos y el cálculo del error, podéis convertir los problemas de clasificación multiclase en problemas de clasificación binaria, agrupando por un lado todas las instancias de la clase minoritaria y por otro lado las del resto de las clases **(hay otras opciones según el conjunto de datos: por ejemplo, en KDD Cup 99, el no ataque vs ataque, y en NYC Taxis definir sólo viajes cortos vs largos)**.

En este entregable cada estudiante debe crear su propio conjunto de árboles de decisión, seleccionando los valores de los parámetros que le parezcan razonable. Es decir, **no uséis las utilidades que proporciona ML para crear y seleccionar modelos**.

Recordad: éste es un trabajo individual, no forma parte del proyecto software, es una parte de la evaluación sumativa.

Scripts que debéis entregar

Debéis entregar un único script en Scala que realice todo el procesamiento necesario del conjunto de datos (el mismo que hayáis realizado en el primer entregable del proyecto software, salvo que detectéis que habéis cometido algún error y los queráis modificar, aspecto que debe ser debidamente justificado), cree los conjuntos de entrenamiento y prueba, y después realice todas las operaciones que se solicitan en este entregable. También hay que entregar el modelo final generado en el directorio: .\Modelo.

El script debe mostrar por pantalla los valores de los parámetros que se modifican, así como el número de nodos y el error de los sucesivos árboles creados (primero para el subconjunto de datos de entrenamiento y para el conjunto de validación, y después para el conjunto de entrenamiento completo y para el conjunto de test).

Breve memoria que debéis entregar

La memoria debe incluir:

1. Vuestros datos personales.
2. El resumen ejecutivo del conjunto de datos.
3. Una descripción del proceso de exploración de parámetros que habéis seguido, indicando por qué habéis considerado esas combinaciones de parámetros y no otras, así como los valores del error que vayáis calculando.
4. El tiempo que habéis empleado en realizar el entregable.
5. La bibliografía que consideréis necesaria.

IMPORTANTE: El script de Scala ha de poder ser ejecutado por Spark 3.5.6 con el comando *":load"* sin necesidad de edición por parte de los profesores, una vez descargados en nuestras máquinas virtuales. La **única** excepción es la localización del fichero de datos original. Para ello, se utilizará la variable "PATH" para indicar la ruta de acceso al fichero en la máquina local. Nosotros modificaremos el valor de esta variable para ejecutar vuestro software en nuestras máquinas. No haremos más modificaciones a vuestro software. **Si con esta única modificación el script no funciona, la calificación de este entregable será de CERO puntos.**

Fecha de entrega:

3 de noviembre de 2025 a las 09:00 h