GUIONES DE PRÁCTICAS – MÓDULO 3

Práctica 5

Contexto

Se va a aplicar un proceso de minería de datos sobre un dataset utilizando métodos de machine learning implementados en MLlib, en este caso para la tarea de clasificación. Al entrenar un método con unos datos se obtiene un modelo que se puede utilizar para clasificar nuevas instancias no vistas en la etapa de entrenamiento. La bondad del modelo obtenido se estima mediante una serie de medidas de evaluación. Pero antes de pasar los datos al modelo para ser entrenado, se tienen que adecuar al mismo (preprocesamiento).

Ejercicio

Partimos del dataset *irisMissing.data*, es el Iris de UCI (*iris.data*), al que se le han sustituido valores en algunos atributos por el carácter ?, para simular un dato perdido. Este dataset contiene los valores de las instancias que teníamos en la práctica 2 pero sin cabecera. Vamos a entrenar dos métodos de clasificación, *logistic regression* y *random forests*, y a evaluar sus resultados sobre este dataset, obteniendo el % de acierto en clasificación (accuracy). Para ello se va a realizar un preprocesamiento del dataset para adecuarlo a los métodos, se aplican los métodos y se evalúan.

Se consideran los siguientes pasos:

1. Crea un RDD, a partir del fichero. Comprueba que hay 150 instancias.
2. Elimina las instancias que contienen valores perdidos y separa los atributos de cada instancia. Comprueba que quedan 144 instancias.
3. Transformar la clase que es nominal a un valor numérico. En el caso de Iris, como la clase de salida tiene tres valores, éstos se transformarán a 0, 1 ó 2. El procesamiento debe ser genérico y si cambia el número de valores de la clase que se adapte a los valores que haya.
4. Convierte todos los valores de los atributos de entrada a *Double* y transforma las instancias al tipo LabeledPoint.
5. Aplica los métodos y obtén las medidas especificadas.

6. Construye un RDD en el que los atributos de entrada (las características que no son la etiqueta) estén estandarizados. Vuelve a entrenar un modelo *logistic regression* y calcula la precisión de este nuevo modelo.

Entrega

Se debe entregar un fichero pdf en el que se especifique el enunciado de cada ejercicio, la solución en spark (con algún comentario explicativo) y captura de pantalla con la ejecución y los resultados obtenidos en la Shell.