GUIONES DE PRÁCTICAS – MÓDULO 3

Práctica 2_1

Contexto

Los algoritmos *machine learning* necesitan aprender a partir de un conjunto de ejemplos contenidos en un *dataset*. Cuando se trata de aprendizaje supervisado cada ejemplo está formado por un conjunto de valores de los distintos atributos o características de entrada y el último valor es el valor objetivo (salida). Dependiendo de la tarea a realizar la salida puede estar formada por un único atributo o por varios. Dicho *dataset* se particionará de forma que con una parte de los datos el modelo aprenda y con la otra parte se evalúe su bondad.

Vamos a acceder a la página de KEEL (Knowledge Extraction base on Evolutionary Leaning) y nos vamos a descargar un *dataset* pequeño, por ejemplo ***iris*** (dentro de la sección KEEL-data-set + Standard Classification data set).

En este caso, en el *dataset* vienen los ejemplos precedidos de una cabecera que especifica información sobre los atributos.

Los *dataset* en este formato KEEL tienen una cabecera (Ejemplo 1) con los siguientes campos:

@relation: a continuación el nombre del *dataset.*

@attribute: a continuación nombre del atributo.

- Si este es numérico se especificará su tipo (*real, integer*) y a continuación un intervalo de dos valores entre corchetes especificando el rango de sus posibles valores.
- Si es nominal no se especifica tipo y los valores que puede tomar se especifican entre llaves.

@inputs: se especifican los nombres de los atributos de entrada.

@outpus: se especifican los nombres del/los atributo/s de salida.

@data: indica que se termina la cabecera y empiezan las líneas de instancias o ejemplos.

```
@relation iris
@attribute SepalLength real [4.3, 7.9]
@attribute SepalWidth real [2.0, 4.4]
@attribute PetalLength real [1.0, 6.9]
@attribute PetalWidth real [0.1, 2.5]
@attribute Class {Iris-setosa, Iris-versicolor, Iris-virginica}
@inputs SepalLength, SepalWidth, PetalLength, PetalWidth
@outputs Class
@data
```

Ejemplo 1. Cabecera del dataset iris

Cada ejemplo o instancia, como se ha comentado, es una línea formada por los valores que toma cada uno de los atributos para dicho ejemplo, siendo el último valor el valor de la clase de salida. Un ejemplo de tales instancias se puede ver en Ejemplo 2.

```
5.1, 3.8, 1.6, 0.2, Iris-setosa
4.6, 3.2, 1.4, 0.2, Iris-setosa
5.3, 3.7, 1.5, 0.2, Iris-setosa
5.0, 3.3, 1.4, 0.2, Iris-setosa
7.0, 3.2, 4.7, 1.4, Iris-versicolor
6.4, 3.2, 4.5, 1.5, Iris-versicolor
6.9, 3.1, 4.9, 1.5, Iris-versicolor
5.5, 2.3, 4.0, 1.3, Iris-versicolor
```

Ejemplo 2. Muestra de un conjunto de instancias del dataset iris

Algunas veces no se conoce el valor de algún atributo para algún ejemplo, a este valor se le conoce como "*valor perdido*" y se identifica porque en su lugar aparece un signo de interrogación, por ejemplo: `5.3, ?, 1.5, 0.2, Iris-setosa`

En el ejemplo visto siempre detrás de una coma hay un espacio en blanco pero puede ocurrir que no siempre sea así, al igual que delante de las llaves o corchetes de los valores de los atributos.

Ejercicios

Crea un RDD con el *dataset* y a continuación ve haciendo las siguientes operaciones:

1. Crea un RDD con los datos de la cabecera y otro con los datos de las instancias. Muestra el número de líneas de cabecera que hay.
2. Muestra el nombre del *dataset*.
3. Cuenta el número de atributos de entrada.
4. Obtén los nombres de los atributos de entrada a partir de las líneas que empiezan por `@atribute` y también a partir de la línea que empieza por `@inputs`. Comprueba que coindicen.

5. Cuenta el número de instancias (puedes comprobar mirando en Keel que el número es correcto).
6. Igual que antes pero considerando que puede haber líneas en blanco entre los ejemplos que no deben ser contadas.

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.