Práctica 3: Integración de *KNIME* y *WEKA*

Sistemas de Ayuda a la Toma de Decisiones

Pedro Allué Tamargo (758267) Juan José Tambo Tambo (755742)
Jesús Villacampa Sagaste (755739)

2 de noviembre de 2020

Índice

1. Ejercicio 1

Se ha creado el *workflow* ilustrado en la Figura **numFigura** para trabajar con los datos del conjunto de datos *yellow-small.data*.
Se ha utilizado un nodo *Rule Engine* para crear una nueva columna *"class"*. El contenido de este nodo son las siguientes reglas:

```
$Color$ MATCHES "YELLOW" AND $Size$ MATCHES "SMALL" => "inflated"
TRUE => "not inflated"
```

Tras este nodo se ha utilizado un nodo *String manipulation* para concatenar los valores de las columnas *class* e *inflated (true/false)* utilizando la expresión:

```
string($class$ + " is " + $Inflated (True/False)$)
```

Meter figura aqui

2. Ejercicio 2

Se va a proceder a analizar un conjunto de datos que describe el número de visitantes de un sitio web en los meses de junio/julio de 2010 (fichero *website1.txt*).

Para calcular los parámetros de media, desviación típica, Kurtosis se ha utilizado el *workflow* mostrado en la Figura **numFigura**.
La *Kurtosis* es una medida estadística que muestra la forma de una distribución de probabilidad. Una *Kurtosis* grande implica una mayor concentración de valores de la variables o muy cerca de la media de la distribución (pico) o muy lejos de ella (colas de la distribución), al mismo tiempo que existe una menor frecuencia de valores intermedios.
¿Cómo ilustramos esto en este conjunto de datos?

Para entrenar la red Bayesiana (Figura **numFigura**) se deben preparar los datos. Para ello se debe crear una nueva columna *isWeekend* para ilustrar si es fin de semana o no. Utilizando el nodo *Rule engine* se utilizarán las siguientes reglas:

```
$weekday$ MATCHES "Sat" => "Yes"
$weekday$ MATCHES "Sun" => "Yes"
TRUE => "No"
```

Se ha utilizado un nodo *Column filter* para eliminar la columna *weekday* ya que la red Bayesiana presenta un mejor rendimiento si conoce este valor ya que si se entrena con esta variable reconoce la regla de creación de la columna *isWeekend*.
Para llegar a esta conclusión se han probado las distintas combinaciones de columnas utilizando el *Column Filter*.
¿Explica más pruebas realizadas?

Para dibujar la curva *ROC* se ha utilizado un nodo *ROC Cuve (local)* a la salida del nodo *Naive Bayes Predictor*. Se puede observar en la Figura **numFigura** que la forma de la gráfica... **(?)**.

Meter figura aqui
Meter gráfica aqui

3. Ejercicio 3

Para la evaluación de los distintos conjuntos de datos se han creado 2 *workflows*. Uno de ellos (Figura **numFigura**) utiliza las herramientas de *KNIME* para evaluar los datos. El otro (Figura **numFigura**) utiliza las herramientas de *WEKA* para evaluar los datos.

Para el conjunto de datos *wine* se han entrenado las distintas herramientas con el 80 %, 50 % y 30 %.
Para el conjunto de datos *iris* se han entrenado las distintas herramientas con el 80 %, 50 % y 30 %.

Para el conjunto de datos *adult* se han entrenado las distintas herramientas con el 80 %, 50 % y 30 %. En el anexo se pueden observar las tablas comparativas de las distintas ejecuciones del *workflow* tanto con *KNIME* (K) como con *WEKA* (W). Se puede apreciar que conforme se reduce el número de datos de entrenamiento se pierde precisión, identificando datos pertenecientes a la otra clase de datos (>50K) como datos de clase <=50K. El método que más ha errado en la identificación de las clases ha sido el perceptrón multicapa.
El método que menos ha variado la precisión conforme se entrenaba con menos datos ha sido el *J48* y en especial utilizando los nodos de *WEKA* siempre ha mantenido el porcentaje de *True Positives* por encima del 93 % en la clase <=50K y del 60 % en la clase >50K.

Meter datos en el Anexo 1

En cuanto a los problemas que han surgido, con el *dataset adult* se necesitaba una columna *Class* y por lo tanto, se ha renombrado la última columna a *"Class"*. Para ello se ha utilizado un nodo *Column Filter* para eliminar todas las variables no numéricas.

Meter figura workflow 1 aqui
Meter figura workflow 2 aqui

4. Anexo 1: Cuadros de datos ejercicio 3

4.1. Dataset Adult

4.1.1. Naive Bayes

Clase	True Positives (K)	False Positives (K)	True Positives (W)	False Positives (W)
<=50K	0.942	0.058	0.936	0.064
>50K	0.511	0.489	0.515	0.485

Cuadro 1: Datos de entrenamiento del dataset *Adult* con el 80 % usando *Naive Bayes*

Clase	True Positives (K)	False Positives (K)	True Positives (W)	False Positives (W)
<=50K	0.941	0.059	0.933	0.067
>50K	0.513	0.487	0.513	0.487

Cuadro 2: Datos de entrenamiento del dataset *Adult* con el 50 % usando *Naive Bayes*

Clase	True Positives (K)	False Positives (K)	True Positives (W)	False Positives (W)
<=50K	0.939	0.061	0.934	0.066
>50K	0.508	0.492	0.522	0.478

Cuadro 3: Datos de entrenamiento del dataset *Adult* con el 30 % usando *Naive Bayes*

4.1.2. J48

Clase	True Positives (K)	False Positives (K)	True Positives (W)	False Positives (W)
<=50K	0.900	0.100	0.936	0.064
>50K	0.603	0.397	0.621	0.379

Cuadro 4: Datos de entrenamiento del dataset *Adult* con el 80 % usando *J48*

Clase	True Positives (K)	False Positives (K)	True Positives (W)	False Positives (W)
<=50K	0.897	0.103	0.937	0.063
>50K	0.598	0.402	0.607	0.392

Cuadro 5: Datos de entrenamiento del dataset *Adult* con el 50 % usando *J48*

Clase	True Positives (K)	False Positives (K)	True Positives (W)	False Positives (W)
<=50K	0.890	0.110	0.933	0.067
>50K	0.599	0.401	0.600	0.400

Cuadro 6: Datos de entrenamiento del dataset *Adult* con el 30 % usando *J48*

4.2. Perceptrón multicapa

Clase	True Positives (K)	False Positives (K)	True Positives (W)	False Positives (W)
<=50K	0.901	0.099	0.927	0.073
>50K	0.377	0.623	0.602	0.398

Cuadro 7: Datos de entrenamiento del dataset *Adult* con el 80 % usando *Perceptrón multicapa*

Clase	True Positives (K)	False Positives (K)	True Positives (W)	False Positives (W)
<=50K	1.000	0.000	0.930	0.070
>50K	0.000	1.000	0.605	0.395

Cuadro 8: Datos de entrenamiento del dataset *Adult* con el 50 % usando *Perceptrón multicapa*

Clase	True Positives (K)	False Positives (K)	True Positives (W)	False Positives (W)
<=50K	0.890	0.110	0.932	0.068
>50K	0.389	0.611	0.587	0.413

Cuadro 9: Datos de entrenamiento del dataset *Adult* con el 30 % usando *Perceptrón multicapa*