Práctica 5

Minería de datos con la herramienta Weka

Marta Blanco Jaime – 562526
Óscar Fraca Ferrández – 566416
Alberto Gómez Climente - 698683

INDICE

Ejercicio 6

Ejecutando el algoritmo J48 sobre Weather.Nominal obtenemos los resultados del anexo 1.
Como se puede apreciar, el 100% de los datos han sido clasificados correctamente siendo la precisión P = 1.0 En la matriz de confusión tenemos todo ceros fuera de la diagonal, lo cual significa que todos los elementos han sido bien clasificados.
Con respecto al árbol generado vemos que el 100% de días que están nublados (28%) se juega partido independientemente de la humedad o viento presente. Cuando hay sol (36%) solo se juega cuando hay humedad normal (14%). Si el día ha resultado lluvioso (36%) entonces solo se juega cuando no hay viento (22%).

Ejercicio 7

Ejecutamos el algoritmo de asociación Apriori y obtenemos los resultados del anexo 2
Todas las reglas tienen la confianza máxima (1), en un primer vistazo a las reglas observamos que hay una variable nueva que había estado oculta en el estudio anterior que es la temperatura.
En la primera regla vemos que si el tiempo es nublado directamente se juega, en caso de no serlo pasamos a la 2ª regla. Si la temperatura es fresca entonces la humedad sera normal y, por la regla 3, si la humedad es normal y no hay viento se puede jugar. Con la regla 4 enlazamos que el clima soleado y el hecho de no jugar solo pasa si la humedad era alta, por lo que con la regla 5, si hay sol y la humedad es alta entonces no se juega. Con la regla 6 enlazamos que el clima lluvioso y el hecho de jugar solo pasa si no hay viento, por lo que con la regla 7, si llueve y no hay viento se juega. En la regla 8 enlazamos, por lo dicho en la regla 2, que si la temperatura es fresca y se ha jugado entonces la humedad era normal. En la regla 9 vemos que si hace sol y la temperatura es alta entonces la humedad sera alta. En la regla 10 tenemos que si la temperatura es alta y no se ha jugado entonces el tiempo era soleado.

Si vamos enlazando varias reglas podemos discernir el árbol obtenido anteriormente.
Con la regla 1 ya queda claro que si hace tiempo nublado se juega siempre.
Con 2, 3 y 8 vemos que si la temperatura es fresca, la humedad sera normal y por lo tanto se jugará si no hay viento, a la vez confirmamos que si la temperatura es fresca y se juega, la humedad sera normal; todo esto sin necesitar saber si hace sol o llueve ya que si hacia sol se juega por humedad normal, y si hacia lluvia se juega por ausencia de viento.
Con 4, 5, 9 y 10 vemos que si hace sol y la temperatura es alta entonces la humedad es alta y no se juega.
Con 6 y 7 vemos que si llueve y no hay viento se juega, y por eliminación el resto de casos resultarían en que llueve y hay viento por lo que no se juega.

Ejercicio 8

Ejecutamos Bayes Ingenuo sobre el fichero de críticas y obtenemos los resultados del anexo 3

Precisión al detectar las opiniones positivas:	0,824
Precisión al detectar las opiniones negativas:	0,794
Recall al detectar las opiniones positivas:	0,784
Recall al detectar las opiniones negativas:	0,832

Precisión promedio:	0,809
Recall promedio:	0,808

Tras quitar símbolos y números (resultado en anexo 4):

Precisión al detectar las opiniones positivas:	0,823
Precisión al detectar las opiniones negativas:	0,800
Recall al detectar las opiniones positivas:	0,792
Recall al detectar las opiniones negativas:	0,830
Precisión promedio:	0,811
Recall promedio:	0,811

Globalmente han mejorado tanto la precisión como la exhaustividad, lo cual es mejor que con la 1ª versión preprocesada.

Se ha procedido a preprocesar la entrada variando cada vez un solo parámetro, eliminando también los símbolos y números y en ninguna configuración se ha obtenido una media superior a 0,811 en ninguna de las dos.
Sin embargo, realizando solo 8 pliegues (en vez de 10) se llega hasta 0,814 en ambos de media.

Metodología

El trabajo se repartió de forma no equitativa para aliviar carga de trabajo respecto a practicas anteriores pendientes.
No ha supuesto mucho tiempo debido a que solo había unos ejercicios que hacer y se ha decidido no hacer la parte optativa.

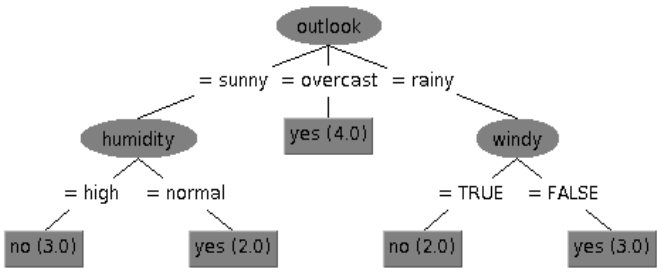
	Ejercicios	Memoria	Total
Marta	0,5h	1h	1,5h
Óscar	2h	1h	3h
Alberto	1h	0,5h	1,5h

Bibliografía

[1] http://www.cs.waikato.ac.nz/ml/weka/index.html (herramienta Weka) Ult. Acc: 14/12/17
[2] http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/review_polarity.tar.gz Ult.Acc: 14/12/17

ANEXO 1

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

outlook = sunny
|   humidity = high: no (3.0)
|   humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :             5

Size of the tree :      8


Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          14               100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     yes
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     no
Weighted Avg.  1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000

=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 0 5 | b = no
```



ANEXO 2

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: weather.symbolic
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.15 (2 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 47

Size of set of large itemsets L(3): 39

Size of set of large itemsets L(4): 6

Best rules found:

1. outlook=overcast 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
2. temperature=cool 4 ==> humidity=normal 4 <conf:(1)> lift:(2) lev:(0.14) [2] conv:(2)
3. humidity=normal windy=FALSE 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
4. outlook=sunny play=no 3 ==> humidity=high 3 <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
5. outlook=sunny humidity=high 3 ==> play=no 3 <conf:(1)> lift:(2.8) lev:(0.14) [1] conv:(1.93)
6. outlook=rainy play=yes 3 ==> windy=FALSE 3 <conf:(1)> lift:(1.75) lev:(0.09) [1] conv:(1.29)
7. outlook=rainy windy=FALSE 3 ==> play=yes 3 <conf:(1)> lift:(1.56) lev:(0.08) [1] conv:(1.07)
8. temperature=cool play=yes 3 ==> humidity=normal 3 <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
9. outlook=sunny temperature=hot 2 ==> humidity=high 2 <conf:(1)> lift:(2) lev:(0.07) [1] conv:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2 <conf:(1)> lift:(2.8) lev:(0.09) [1] conv:(1.29)

ANEXO 3

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes
Relation: _home_a566416_Downloads_review_polarity_txt_sentoken-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"
Instances: 2000
Attributes: 1166
[list of attributes omitted]
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

(...) Desglose de atributos (...)

Time taken to build model: 0.58 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	1616	80.8 %
Incorrectly Classified Instances	384	19.2 %
Kappa statistic	0.616	
Mean absolute error	0.1918	
Root mean squared error	0.4111	
Relative absolute error	38.3507 %	
Root relative squared error	82.2217 %	
Total Number of Instances	2000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,832	0,216	0,794	0,832	0,813	0,617	0,897	0,897	neg
	0,784	0,168	0,824	0,784	0,803	0,617	0,897	0,890	pos
Weighted Avg.	0,808	0,192	0,809	0,808	0,808	0,617	0,897	0,894	

=== Confusion Matrix ===

a	b	<-- classified as
832	168	a = neg
216	784	b = pos

ANEXO 4

Time taken to build model: 0.56 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	1622	81.1 %
Incorrectly Classified Instances	378	18.9 %
Kappa statistic	0.622	
Mean absolute error	0.1917	
Root mean squared error	0.4108	
Relative absolute error	38.3416 %	
Root relative squared error	82.1694 %	
Total Number of Instances	2000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,830	0,208	0,800	0,830	0,815	0,622	0,896	0,896	neg
	0,792	0,170	0,823	0,792	0,807	0,622	0,896	0,890	pos
Weighted Avg.	0,811	0,189	0,811	0,811	0,811	0,622	0,896	0,893	

=== Confusion Matrix ===

a	b	<-- classified as
830	170	a = neg
208	792	b = pos