Data Mining I

Corsi di Laurea Magistrale in Business Informatics, Informatica e Informatica Umanistica

Second Part -Test 30.05.2016
Docenti: Dino Pedreschi, Anna Monreale

Exercise 1 (12 Points)

Consider the following transactions

Transaction ID	Itemsets
1	{B,E}
2	{E,R}
3	{B,E,R}
4	{A,R,C}
5	{A,B,E,F}
6	{A,B}
7	{B,R,F}
8	{E,B,F}
9	{A,R,F}
10	{A,R,C,F}

A) Extract the frequent itemsets by *Apriori* using *min sup=20%*, showing and discussing the different steps of the algorithm **(7 points)**

B) Extract the association rules using minimum confidence equal to 70% **(3 points)**

C) Compute the lift for the rules extracted in the previous point and discuss them (**2 points**)

Exercise 2 (16 Points)

Consider the following dataset

Training Data

Height	Weight	Age	Sex	Disease
Short	High	Young	F	No
Short	Low	Young	F	Yes
Short	Low	Old	M	No
Short	Medium	Young	M	Yes
Short	High	Young	M	Yes
Tall	Medium	Old	F	Yes
Short	High	Young	F	No
Tall	Low	Old	M	No
Tall	Low	Old	F	Yes
Short	Medium	Old	M	Yes

A) Use the above training dataset for building a decision tree based on misclassification rate for the variable "DISEASE", expanding the nodes of the tree until the precision is not improved locally, i.e., no split provides a gain. **(10 points)**

B) Provide the confusion matrix and evaluate the accuracy, precision and recall of the tree with respect to the above training set and the following test set **(6 Points)**

Test Data

Height	Weight	Age	Sex	Disease
Tall	High	Old	F	
Short	High	Old	F	
Short	Low	Young	F	
Tall	Low	Young	M	
Tall	Medium	Old	M	

Exercise 3 (4 Points)

Answer to the following questions:

A) How many frequent subsets does the frequent pattern {a, b, c, d, e} contain?

B) Given the test set in Exercise 2, compute the Gini Index, the Entropy and the Misclassification error on the attribute AGE.