Data mining (2013/11/21)

1. (20%) (Association rules) given a transaction database in the following. Answer the follow questions with min\_sup = 40%.

|  |  |
| --- | --- |
| Transaction ID | Items Bought |
| T100 | {a,b,d,e} |
| T200 | {a,b,c} |
| T300 | {a,b,d,e} |
| T400 | {b,d,e} |
| T500 | {a,b,c,d,e} |
| T600 | {b,d,e} |

1. Use the Apriori algorithm to find all the maximal itemsets (10%)
2. Compute the confidence and lift of the rule of ”b→e”. (5%)
3. Draw the FP-Tree of this transaction dataset.
4. (20%) Given a dataset in the following where "+ " and "-" signs denote the class labels. Answer the following questions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Instance | a1 | a2 | a3 | Target Class |
| 1 | T | T | 1.0 | + |
| 2 | T | T | 6.0 | + |
| 3 | T | F | 5.0 | - |
| 4 | F | F | 4.0 | + |
| 5 | F | T | 7.0 | - |
| 6 | F | T | 3.0 | - |
| 7 | F | F | 8.0 | - |
| 8 | T | F | 7.0 | + |
| 9 | F | T | 5.0 | - |

1. Using the Gini function as attribute selection measure; select the best splitting attribute from attribute a1 and a2 for the root. (5%)
2. For a3, which is a continuous attribute, compute the information gain for every possible split. What is the splitting condition for this attribute? (10%)
3. According to the results of questions a and b, draw a two-level decision tree and calculate its classification accuracy. (5%)

An example.

scn0001.tif

1. (20%) (Model selection) according to the following table, draw the ROC charts for classifiers M1 and M2. According to the charts, which classifier performs better?

|  |  |  |  |
| --- | --- | --- | --- |
| Instance | True Class | P(+|A,…,Z,M1) | P(+|A,…,Z,M2) |
|  | + | 0.73 | 0.61 |
|  | - | 0.69 | 0.03 |
|  | + | 0.44 | 0.68 |
|  | - | 0.55 | 0.31 |
|  | + | 0.67 | 0.45 |
|  | - | 0.47 | 0.09 |
|  | + | 0.08 | 0.38 |
|  | - | 0.15 | 0.05 |
|  | + | 0.45 | 0.01 |
|  | - | 0.35 | 0.04 |

4. (20%) （Data preprocessing）

1. What are the steps for data preprocessing？(15%)
2. According the contingency table in the following, Determine whether smoking will induce lung cancer or not? The symbol L is for having a lung cancer (L=1) or not having a lung cancer (L=0) while S is for smoking (S=1) or not smoking (S=0). Assuming that the significant level is set ot 0.01. (5%)

|  |  |  |  |
| --- | --- | --- | --- |
|  | L=1 | L=0 | Total |
| S=1 | 8 | 19 | 27 |
| S=0 | 1 | 10 | 11 |
| Total | 9 | 29 | 38 |

Related equation: ，oi is the observation and eis the expectation value of cell i, respectively.

5. (20%) (Accuracy estimation methods) What are the different methods for calculating the classification accuracy of a classifier?