Data mining (2016/11/14)

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

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

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A | B | C | Number of Instances | |
| + | - |
| T | T | T | 5 | 0 |
| F | T | T | 0 | 20 |
| T | F | T | 20 | 0 |
| F | F | T | 0 | 5 |
| T | T | F | 0 | 0 |
| F | T | F | 25 | 0 |
| T | F | F | 0 | 0 |
| F | F | F | 0 | 25 |

1. Using Gini function as the attribute selection criterion; select the best splitting attribute for the root. (10%)
2. Build a two-level decision tree. How many instances are misclassified by the resulting decision tree? (10%)

1. (20%) (Model selection) according to the following table, draw the ROC charts for M1 and M2. Based on the two charts, point out which model performs better.

|  |  |  |  |
| --- | --- | --- | --- |
| Instance | True Class | P(+|M1) | P(+|M2) |
|  | + | 0.50 | 0.61 |
|  | - | 0.69 | 0.03 |
|  | + | 0.44 | 0.68 |
|  | - | 0.55 | 0.31 |
|  | + | 0.67 | 0.45 |

A little help: TPR=TP/(TP+FN); FPR=FP/(TN+FP)

4. (20%) （Data preprocessing）

1. What are the steps in data preprocessing？(10%)
2. Draw a boxplot example using symbols of Q1, M, Q3, Max and Min. (5%)
3. According the contingency table in the following, Determine whether smoking and lung cancer are correlated or not? The symbol L stands for having a lung cancer (L=1) or not having a lung cancer (L=0), while S stands for smoking (S=1) or non smoking (S=0). Assuming that the significant level is set to 0.01. (5%)

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

Chi Square distribution table for Probability level (alpha)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Degrees of Freedom | 0.1 | 0.05 | 0.02 | 0.01 | 0.001 |
| 1 | 2.706 | 3.841 | 5.412 | 6.635 | 10.827 |
| 2 | 4.605 | 5.991 | 7.824 | 9.210 | 13.815 |

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

5. (20%) (Bayesian belief network) Given the Bayesian network as shown in the following Figure, calculate the following probabilities:

* 1. P(B=good, F= empty, G=empty, S=yes) (5%)
  2. P(B=bad, F=empty, G=not empty, S=no) (5%)
  3. Given that the battery is bad. Compute the probability that the car will start. (10%) (Note: Gauge 儀表版， Battery 電池 Fuel 燃料，汽油)

