# MIDTERM – Fall 2016

## CS583: Data Ming and Text Mining

Name:_____ UID_____

## Instruction:

1. This is a close book test.
2. The paper has 7 questions and the full mark is 80.

| | **Marks** |
|---|---|
| Q1 | |
| Q2 | |
| Q3 | |
| Q4 | |
| Q5 | |
| Q6 | |
| Q7 | |
| **Total** | |

1. (10%) In multiple minimum support association rule mining, we can assign a minimum item support (MIS) to each item. Given the transaction data below:

   {1, 7, 3}
   {1, 7, 5, 4}
   {1, 7, 5, 2}
   {7, 2}
   {5, 6}
   {7, 3}
   {7, 5, 2}
   {7, 2, 3}

   and the MIS assignments for the items in the data:

   MIS(2) = 50%  MIS(7) = 80%
   The MIS values for the rest of the items in the data are all 25%,

   use the MSapriori algorithm to produce the set of frequent itemsets in $F_1$, $F_2$, $F_3$.

2. (10%) Sequential pattern mining.

   Given the minimum support of 25% and the following sequence data, generate all sequential patterns.

| Customer ID | Customer sequence |
|---|---|
| 1. | <{20} {100, 80} {50}> |
| 2. | <{30} {70}> |
| 3. | <{90} {40}> |
| 4. | <{20, 30} {10} {100, 70, 80}> |
| 5. | <{10, 70, 80}> |

3. (10%) Given the following training data, which has two attributes A and B, and a class C, compute all the probability values required to build a naïve Bayesian classifier. Ignore smoothing.

| A | B | C |
|---|---|---|
| n | x | Y |
| h | m | Y |
| g | m | Y |
| n | m | Y |
| n | x | Y |
| h | x | N |
| g | m | N |
| g | m | N |
| n | x | N |
| n | x | N |

4. (10%) Assume we have built a naïve Bayesian classifier *h* using some training data, and have used *h* to classify the following test data, which give us the corresponding probability for each data point $d_i$. Fill up the table below with appropriate values for the test data and draw the ROC curve.

**Test data**

| $d_i$ | *Actual class* | Pr(+\|$d_i$) |
|---|---|---|
| $d_1$ | - | 0.2 |
| $d_2$ | + | 0.8 |
| $d_3$ | - | 0.4 |
| $d_4$ | + | 0.1 |
| $d_5$ | + | 0.5 |
| $d_6$ | + | 0.7 |

| Rank | | | | | | | |
|---|---|---|---|---|---|---|---|
| Actual class | | | | | | | |
| TP | | | | | | | |
| FP | | | | | | | |
| TN | | | | | | | |
| FN | | | | | | | |
| TPR | | | | | | | |
| FPR | | | | | | | |

5. (10%) Given the following dataset with two classes (yes and no) for decision tree building, the "credit_rating" attribute has already been selected as the root node. We want to grow the tree further. Use the information gain criterion to compute the gain value for attribute "student" for the tree branch of "credit_rating = excellent". Give the detailed computation.

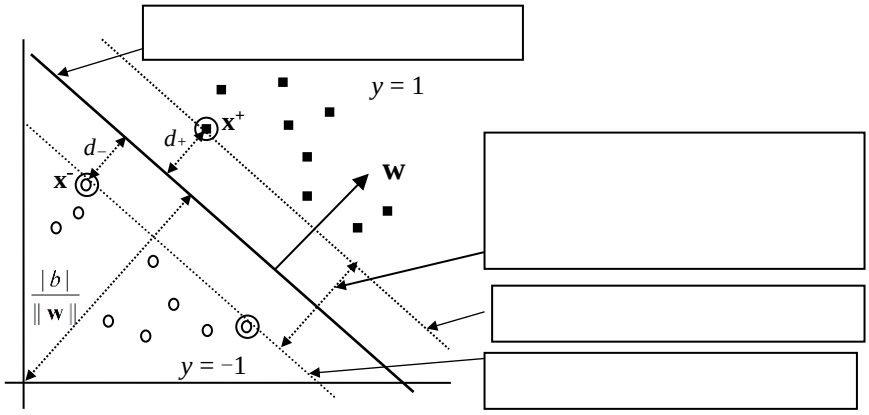
| age | income | student | credit_rating | Class |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |

6. (10%) Given the classification results in the following confusion matrix, compute *the precision*, *recall* and *F*-score for the ***negative class***, and the overall ***classification accuracy***.

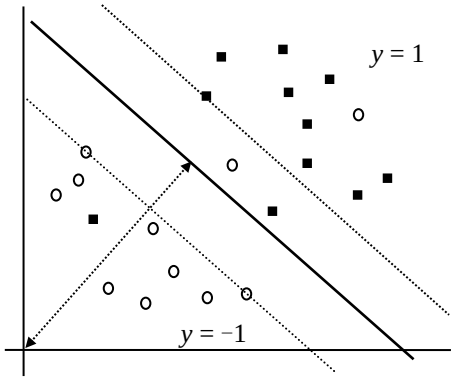
| **Classified as** | | | **Correct** |
|---|---|---|---|
| Positive | Neutral | Negative | |
| 60 | 25 | 20 | Positive |
| 10 | 10 | 180 | Negative |
| 10 | 35 | 5 | Neutral |

7. (20%) Answer the following questions.

(a). (4%) Write the equation in each of the boxes in the figure for SVM.




(b). (3%) Please circle the support vectors for linear non-separable SVM in the figure below.




(c) (3%) In deriving the naïve Bayesian classification (not the text classification version), an important assumption was made. What is it called? Give the equation.

(d). (3%) Boostsing is a kind of weighted voting, what is the weight?

(e). (3%) Give two reasons for missing data. Which classification algorithm can be used for attribute discretization?

(f). (4%) Given the two data points, $\mathbf{x} = (x_1, x_2)$, $\mathbf{z} = (z_1, z_2)$, we want to use the following kernel function,

$$K(\mathbf{x}, \mathbf{z}) = <\mathbf{x} \bullet \mathbf{z}>^3 = <\phi(\mathbf{x}) \bullet \phi(\mathbf{z})>$$

Compute $\phi(\mathbf{x})$ and $\phi(\mathbf{z})$.