MIDTERM – Sample

CS583: Data Ming and Text Mining

Name:_______________________________ UID____________________________

Section:______________________

Instruction:

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

	Marks
Q1	
Q2	
Q3	
Q4	
Q5	
Q6	
Q7	
Total	

1. (10%) Given the actual classes and the classified classes in the table at the bottom of the page, (a) fill the confusion matric below, (b) compute *the precision*, *recall* and *F* score of the **positive class**, and the **overall *accuracy***.

Classified as			**Actual**
Positive	Negative	Neutral	
			Positive
			Negative
			Neutral

No.	Actual class	Classified as
1	Positive	Positive
2	Positive	Positive
3	Positive	Positive
4	Positive	Positive
5	Positive	Positive
6	Positive	Positive
7	Positive	Positive
8	Positive	Positive
9	Positive	Negative
10	Positive	Neutral
11	Neutral	Neutral
12	Neutral	Neutral
13	Neutral	Neutral
14	Neutral	Neutral
15	Neutral	Neutral
16	Neutral	Neutral
17	Neutral	Neutral
18	Neutral	Neutral
19	Neutral	Positive
20	Neutral	Negative
21	Negative	Negative
22	Negative	Negative
23	Negative	Negative
24	Negative	Negative
25	Negative	Negative
26	Negative	Negative
27	Negative	Negative
28	Negative	Negative
29	Negative	Neutral
30	Negative	Positive

2. (10%) Sequential pattern mining.

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

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

3. (10%) Given the following 7 training documents (which are already represented as bags of words and their classes) and the test document *d*: “Baseball, Coach”
   a. (6%) compute the probabilities needed for text classification. Ignore smoothing.
   b. (4%) what is the predicted probability *Pr*(Education | *d*) and what is the predicted probability *Pr*(Sport | *d*)?

Document	**Class**
Teacher, Class, Book	: Education
Teacher, Lecture, Class	: Education
Teacher, Class, Class, Book	: Education
Baseball, Football	: Sport
Football	: Sport
Baseball, Coach, Coach, Team	: Sport
Football, Team	: Sport

4. (10%) Given the following data set,

Transactions	**Class**
1, 2	: S
1, 2	: S
3, 2	: S
4, 5	: E
5, 6, 7	: E
4, 8, 9, 10	: E
5, 10	: E

a. Find all class association rules (CAR) with minsup = 20% and minconf = 60%.

b. Build a classifier using the CBA method below. *S* is the set of CAR rules discovered above and *D* is the dataset.

```
Algorithm CBA(S, D)
S = sort(S);      // sorting is done according to the precedence defined in the book
RuleList = ∅;     // the rule list
for each rule r ∈ S in sequence do
    if D ≠ ∅ AND r classifies at least one example in D correctly then
        delete from D all training examples covered by r;
        add r at the end of RuleList
    end
end
add the majority class as the default class at the end of RuleList
```

5. (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 gives us the predicted 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	-	0.5
d_2	+	0.2
d_3	-	0.1
d_4	+	0.8
d_5	+	0.9

Rank						
Actual class						
TP						
FP						
TN						
FN						
FPR						
TPR						

6. (10%) Given the following dataset with two classes, yes and no, and that "student" has already been selected as the root node of the decision tree, use the information gain criterion to compute the gain value for attribute "credit_rating" under "student" = no. 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

7. (20%) Mark the **most appropriate** answer for the following questions. There is **only one best answer** for each question.

(1). What are the parameters of a multinomial distribution?

1. The probability of each document.
2. The probability of each outcome and the number of independent trials.
3. The probability of each word in a document.
4. The probability of each outcome.

(2). Overfitting is often manifested as follows:

1. The model classifies the training data very well, but classifies the test data poorly.
2. The model classifies both the training data and the test data very well.
3. The model classifies both the training data and the test data poorly.
4. The model classifies the training data poorly, but classifies the test data very well.

(3). In a generative model for classification of classes $c_1, \ldots, c_N$, the probability of generating a data point d is.

1. $\Pr(d) = \sum_{i=1}^{N} \Pr(c_i) \Pr(d|c_i)$
2. $\Pr(d) = \prod_{i=1}^{N} \Pr(c_i) \Pr(d|c_i)$
3. $\Pr(d) = \sum_{i=1}^{N} \Pr(d|c_i)$
4. $\Pr(d) = \sum_{i=1}^{N} \Pr(c_i) \Pr(c_i|d)$

(4) In naïve Bayesian classification, which of the following is assumed?

1. $\Pr(A_1 = a_1|C = c_i) = \Pr(A_1 = a_1)$
2. $\Pr(A_1 = a_1|A_2 = a_2, C = c_i) = \Pr(A_1 = a_1|C = c_i)$
3. $\Pr(A_1 = a_1|C = c_i) = \Pr(C = c_i)$
4. $\Pr(A_1 = a_1, A_2 = a_2, |C = c_i) = \Pr(A_1 = a_1|C = c_i)$

(5). Given the dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, which one of the following forms a Kernel matrix for SVM?

1. $\langle \mathbf{w} \cdot \mathbf{x}_i \rangle$
2. $\langle \mathbf{x}_j \cdot \mathbf{x}_i \rangle$
3. $\langle \mathbf{x}_j + \mathbf{x}_i \rangle$
4. $\langle \mathbf{x}_i \times \mathbf{x}_j \rangle$

(6). In $\frac{\partial L_P}{\partial w_j} = w_j - \sum_{i=1}^{n} y_i \alpha_i x_{ij} = 0, \;\; j = 1, 2, ..., r$, what is r?

1. The number of errors
2. The number of support vectors.
3. The number of data points.
4. The number of features

(7). Given $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) > 1$ and $\xi_i = 0$ in SVM, which of the following is correct?

1. These data points are a subset of the support vectors.
2. These data points are not support vectors.
3. These data points have $\alpha_i > 0$.
4. These data points do not have $\alpha_i = 0$.

(8). Let $\mathbf{x} = (x_1, x_2)$ and $\mathbf{z} = (z_1, z_2)$. We use $K(\mathbf{x}, \mathbf{z}) = <\mathbf{x} \bullet \mathbf{z}>^2$ as the kernel function. What is $\phi(\mathbf{x})$?

1. $(x_1^2, x_2^2, \sqrt{3}x_1x_2)$
2. $(x_1^2, x_2^2, \sqrt{2}x_1x_2)$
3. $(z_1^2, z_2^2, \sqrt{3}x_1x_2)$
4. $(x_1^2, x_2^2, \sqrt{2}z_1z_2)$

(9). In AdaBoost, we compute the errors for each iteration as follows:

AdaBoost(D, Y, BaseLeaner, k)
1. Initialize $D_1(w_i) \leftarrow 1/n$ for all i; // initialize the weights
2. **for** $t = 1$ to k **do**
3. $f_t \leftarrow$ BaseLearner(D_t); // build a new classifier f_t
4. $e_t \leftarrow \sum_{?} D_t(w_i)$ // compute the error of f_t

What should the '?' in line 4 be replaced with?

1. i: $D_t(\mathbf{x}_i) \neq y_i$
2. i: $f_t(D_t(\mathbf{x}_i)) \neq y_i$
3. i: $D_t(f_t(\mathbf{x}_i)) \neq y_i$
4. i: $f_t(D_t(y_i)) = \mathbf{x}_i$

(10). In Bagging, given n training data points, we obtain m bootstrap samples and use them to build m classifiers. Which of the following is corrected?

1. Each sample S consists of n data points sampled without replacement from the n training data points.
2. Each sample S consists of m data points sampled from the n training data points.
3. Each sample S consists of n data points sampled with replacement from the n training data points.
4. Each sample S consists of m data points sampled with replacement from the n training data points.