**MIDTERM – Fall 2017**

**CS583: Data Ming and Text Mining**

Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ UID\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Instruction:

1. This is a closed-book test.
2. The paper has 9 questions and the full mark is 80.

|  |  |
| --- | --- |
|  | **Marks** |
| Q1 |  |
| Q2 |  |
| Q3 |  |
| Q4 |  |
| Q5 |  |
| Q6 |  |
| Q7 |  |
| Q8 |  |
| Q9 |  |
| **Total** |  |

1. (5 marks) Sequential pattern mining.

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

Customer ID Customer sequence

1. <{60} {90}>

2. <{20, 30} {50} {60, 70, 90}>

3. <{50} {60, 80} {90}>

4. <{50, 60, 80}>

5. <{50} {90}>

2. (5%) In sequential pattern mining, the GSP algorithm uses joining and pruning to generate candidates. Given the frequent 3-sequences in the table, generate candidate 4-sequences.

|  |  |  |
| --- | --- | --- |
| **Frequent**  **3-sequences** | **Candidate 4-sequences** | |
| **after joining** | **after pruning** |
| 〈{2, 5} {4}〉 |  |  |
| 〈{2, 5} {8}〉 |  |  |
| 〈{2} {4, 8}〉 |  |  |
| 〈{2, 4} {6}〉 |  |  |
| 〈{5} {4, 8}〉 |  |  |
| 〈{5} {4} {6}〉 |  |  |

3. (5 marks) Let **x** = (*x*1, *x*2) and **z** = (*z*1, *z*2). We use *K*(**x**, **z**) = <**x** • **z**>3 as the kernel function. What is *φ*(**x**)?

4. (5 marks) Given the classification results in the following confusion matrix, compute *the* *precision*, *recall* and *F* score of the **positive class**, and also the **overall** ***accuracy***.

|  |  |  |  |
| --- | --- | --- | --- |
| **Classified as** | | | **Correct** |
| Positive | Negative | Neutral |
| 60 | 20 | 20 | Positive |
| 10 | 200 | 10 | Negative |
| 5 | 5 | 35 | Neutral |

5. (10 marks) Given the training data in the table with two attributes A and B, and the class C,

1. Training: compute all the probabilities required to build a naïve Bayesian classifier. Ignore smoothing.
2. Testing: given a test data point *d* with A = *m* and B = *m*, what is the predicted probability Pr(C=T | *d*) and what is the predicted probability Pr(C=F | *d*)?

|  |  |  |
| --- | --- | --- |
| A | B | C |
| a | m | T |
| k | m | T |
| m | a | T |
| a | m | T |
| a | a | T |
| k | a | F |
| m | a | F |
| m | m | F |
| a | a | F |
| a | m | F |

6. (10 marks) 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 *di*. Fill up the table below with appropriate values for the test data and draw the ROC curve.

**Test data**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *di* | *Actual class* | Pr(+|*di*) |  | Rank |  |  |  |  |  |  |
| d1 | - | 0.7 |  | Actual class |  |  |  |  |  |  |
| d2 | + | 0.2 |  | TP |  |  |  |  |  |  |
| d3 | - | 0.1 |  | FP |  |  |  |  |  |  |
| d4 | + | 0.6 |  | TN |  |  |  |  |  |  |
| d5 | + | 0.9 |  | FN |  |  |  |  |  |  |
|  |  |  |  | FPR |  |  |  |  |  |  |
|  |  |  |  | TPR |  |  |  |  |  |  |

7. (10 marks) Given the following dataset with two classes, yes and no, use the information gain criterion to compute the gain value for attribute “income”. Give the detailed computation.



8. (10%) Prove the following:

Pr(a1, a2, …, an | c) = Pr(a1 | a2, …, an, c) × Pr(a2, …, an | c)

9. (20 marks) Mark the **most appropriate** answer for the following questions. There is **only** **one best answer** for each question.

(1). Which of the following statements is true about decision trees?

1. Decision tree building is for discretizing continuous attributes into intervals.
2. There are different ways to select the next attribute to partition the data.
3. The complexity of decision tree building with numeric attributes is O(*n*), where *n* is the number of data points.
4. The complexity of decision tree building is O(*k log(n)*), where *n* is the number of data points and *k* is the number of attributes.

(2). Precision and recall usually have the following relationship

1. When the precision is high, the recall is usually low.
2. When the precision is high, the recall is usually high as well.
3. Theoretically, precision and recall are correlated.
4. *F* score is computed because precision and recall are related.

(3). The sequential covering strategy means the following,

1. Find the best rules to cover the data
2. Find the best rules with only one condition.
3. Find the best rules with the highest confidences.
4. Find the best rules with the highest information gains.

(4). In building a classifier using class association rules, which one of the following strategies is reasonable?

1. Select a set of high support rules.
2. Select a set of high confidence rules.
3. Select a set of rules that covers all classes.
4. Select a set of rules with the most conditions.
5. In naïve Bayesian classification, which of the following statements is true?
6. Naïve Bayesian classification makes the independence assumption.
7. Naïve Bayesian classification makes the class independence assumption.
8. Naïve Bayesian classification makes the conditional independence assumption.
9. Naïve Bayesian classification does not make any independence assumption.

(6). In linearly separable SVM, we say that the complementary condition indicates

1. Those data points on the decision hyper-plane are support vectors
2. Those noisy data points are support vectors.
3. Those data points on the margin hyperplanes are support vectors.
4. Those data points inside the margin area are support vectors.

(7). In linearly non-separable SVM, what does the following say?

|  |  |
| --- | --- |
| *αi* = 0 ⇒ *yi*(〈**w ⋅ x***i*〉 *+ b*) ≥ 1 and *ξi* = 0 |  |

1. For most data points inside the margin area, *αi* = 0.
2. For most data points outside the margin area, *αi* = 0.
3. For most data points on the margin hyperplanes, *αi* = 0.
4. For most data points far away from the margin area, *αi* = 0.

(8). In linearly non-separable SVM, what does the following say?

|  |  |
| --- | --- |
| 0 < *αi* < *C* ⇒ *yi*(〈**w ⋅ x***i*〉 *+ b*) = 1 and *ξi* = 0 |  |

1. For those data points inside the margin area, *αi* is non-zero.
2. For those data points outside the margin area, *αi* is non-zero.
3. For those data points on the margin hyperplanes, *αi* is non-zero.
4. For those data points far away from the margin area, *αi* is non-zero.

(9). In AdaBoost, we update weights of training examples after each iteration as follows:

9. *βt* ← *et* / (1− *et*);

10 *Dt*+1(*wi*) ← *Dt*(*wi*) ×  // update the weights

11. *Dt*+1(*wi*) ←  // normalize the weights

What should ‘?’ in line 10 be replaced with?

1. if *ft*(*Dt*(**x***i*)) = *yi*
2. if *ft*(*Dt*(**x***i*)) ≠ *yi*
3. if *ft*(*Dt*(*yi*)) = **x***i*
4. if *ft*(*Dt*(*yi*)) ≠ **x***i*

(10). In Bagging, given *n* training data points, each bootstrap sample *S* consists of *n* data points sampled with replacement from the original *n* training data points. We say

1. Each data point has the probability of (1 – 1/*n*) not being selected into *S*.
2. Each data point has the probability of (*n* – 1/*n*) not being selected into *S*.
3. Each data point has the probability of (1 – 1/*n*)/*n* not being selected into *S*.
4. Each data point has the probability of (1 – 1/*n*)*n* not being selected into *S*.