Final Exam – Spring 2017

CS583: Data Mining and Text Mining

Name:_____ UID_____

Instructions:

1. This is a close book examination
2. The paper has 8 questions and the full mark is 100.

	Marks
Q1	
Q2	
Q3	
Q4	
Q5	
Q6	
Q7	
Q8	
Total	

1. (40%) Answer the following questions. There is only one best answer for each question.

(1) The downward closure property in association rule mining means

(a) If a set is frequent, then any of its supersets must be frequent
(b) If a superset is frequent, then its subsets must be frequent
(c) If a set is frequent, then its subsets must not be frequent.
(d) If a set is infrequent, then some of its supersets may be frequent

(2) In sequential pattern mining, the GSP algorithm uses joining and pruning steps to generate candidates. Given the frequent 3-sequences, 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). 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

(a) Each data point has the probability of $(1 - 1/n)$ not being selected into S.
(b) Each data point has the probability of $(n - 1/n)$ not being selected into S.
(c) Each data point has the probability of $(1 - 1/n)/n$ not being selected into S.
(d) Each data point has the probability of $(1 - 1/n)^n$ not being selected into S.

(4). In supervised learning, which algorithm does not build any model?

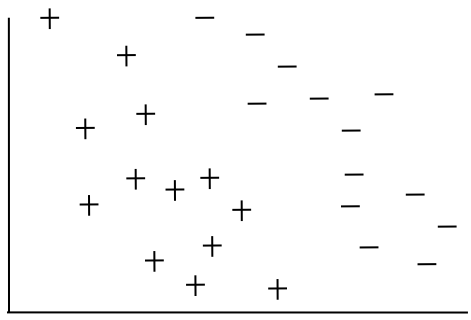
(a) KNN
(b) SVM
(c) Decision tree
(d) Naïve Bayes

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

$$0 < \alpha_i < C \quad \Rightarrow \quad y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) = 1 \quad \text{and} \quad \xi_i = 0$$

(a) For those data points inside the margin area, α_i is non-zero.
(b) For those data points outside the margin area, α_i is non-zero.
(c) For those data points on the margin hyperplanes, α_i is non-zero.
(d) For those data points away from the margin area, α_i is non-zero.

(6). Given the following positive and negative data points, draw a possible decision tree partition.



(7). *K*-means clustering aims to minimize

(a) Sum of squared error
(b) Mean squared error
(c) Root mean squared error
(d) Sum of compactness error

(8). In TF-IDF, what is IDF?

(a) $idf_i = \log (N / df_i)$
(b) $idf_i = \log (df_i / N)$
(c) $idf_i = \log (df_i)$
(d) $idf_i = \log (N)$

(9). Which is the following statement is true for co-training?

(a) Classifier f_1 adds examples to the labeled set that is used to learn f_2 based on the X_2 view, and vice versa.
(b) Classifier f_1 adds examples to the labeled set that is used to learn f_1 based on the X_1 view, and classifier f_2 does likewise.
(c) Classifier f_1 adds examples to the labeled set that is used to learn f_2 based on the X_1 view, and vice versa
(d) Classifier f_1 adds examples to the labeled set that is used to learn two new classifiers, and classifier f_2 does likewise.

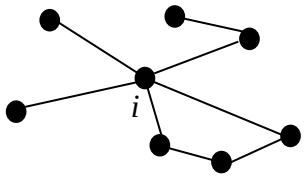
(10). Which of the following statement is true for EM using naïve Bayes?

(a) The E step builds a new classifier and fills the missing values
(b) The M step builds a new classifier and fills the missing values
(c) The E step computes $\Pr(c_j \mid d_i)$
(d) The M step computes $\Pr(c_j \mid d_i)$

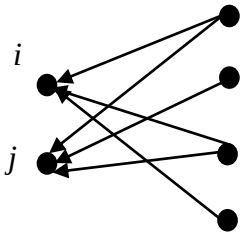
(11). Many PU learning methods have two steps. The first step

(a) identifies a set of reliable negative examples
(b) identifies a set of reliable positive examples
(c) identifies a set of reliable unlabeled examples
(d) identifies a set of good examples

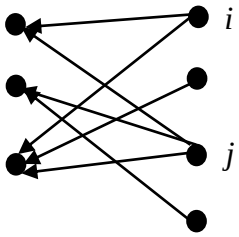
(12). What is the degree centrality of node *i*? What is the closeness centrality of node *i*?



(13). What is the co-citation value of nodes *i* and *j* in the following citation graph?



(14). What is the bibliographic coupling value of *i* and *j* in the following citation graph?



(15). Which of the following statement is true?

(a) PageRank is based on rank prestige
(b) PageRank is based on degree prestige
(c) PageRank is based on proximity prestige
(d) PageRank is based on centrality

(16). Which of the following statement is true for the HITS algorithm?

(a) The authority matrix of HITS is the bibliographic coupling matrix
(b) The hub matrix of HITS is the co-citation matrix
(c) The authority matrix of HITS is the co-citation matrix
(d) The hub matrix of HITS is the bibliometric matrix

(17). Give the quintuple for aspect-based opinion mining.

(18). Collaborative filtering (CF) recommender systems recommend items based on

(a) The user's interests or preferences in the past
(b) The user's interests or preferences in the future
(c) People with similar tastes or preferences in the past
(d) People with similar tastes or preferences in the future

(19). We can use the resulting matrices $\boldsymbol{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_I]$ (user-aspect matrix) and $\boldsymbol{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_J]$ (movie-aspect matrix) to predict the rating for each user-movie pair. Give the formula that does the prediction.

(20). The following is a learning rule for matrix factorization. Which symbol is the learning rate, and what is the gradient?

$$u_{ki}^{t+1} = u_{ki}^{t} + 2\gamma(r_{ij} - p_{ij})m_{kj}^{t}$$

2. (10%) The table below is a test data set together with the classification result of a classifier. "True Class" of each row is the actual class of the data instance. "Predicted Class" is the predicted class by the classifier. X and Y are two attributes of the data. Compute the precision, recall and F score for the class *p*, and also the overall classification accuracy. Draw the confusion matrix.

	X	**Y**	**True Class**	**Predicted Class**
1	1.5	1	p	p
2	1.5	1	p	p
3	1.5	1	p	m
4	1.5	2	p	p
5	1.5	2	p	p
6	1.5	2	p	n
7	1.5	3	p	p
8	1.5	3	p	p
9	1.5	3	p	m
10	1.5	3	p	p
11	2.5	2	n	n
12	2.5	2	n	n
13	2.5	2	n	n
14	2.5	2	n	p
15	2.5	2	n	n
16	2.5	4	n	n
17	2.5	4	n	n
18	2.5	4	n	n
19	2.5	4	n	n
20	2.5	4	n	p
21	5.5	2	m	n
22	5.5	2	m	m
23	5.5	2	m	m
24	5.5	2	m	m
25	5.5	2	m	p
26	5.5	4	m	m
27	5.5	4	m	p
28	5.5	4	m	m
29	5.5	4	m	m
30	5.5	4	m	m

3. (5%) Build an inverted index using the following three sentences (*is*, *in*, *I*, *a*, *with* and *of* do not need to be indexed). You should include the position of each word in the index so that word proximity can be considered in search.

   S1: "Web data mining is useful in many applications"

   S2: "I took a web mining class"

   S3: "Mining with one class of data"

4. (5%) produce all opinion quintuples from the following blog post.

   Id: Abc123 on 5-1-2008 -- "*This morning I bought a Nokia phone and my wife bought a Motorola phone. We called each other when we got home. The voice on my phone was not so clear, but the touch screen is cool. My wife was quite happy with her phone. I wanted a phone with good sound quality. So my purchase was a real disappointment.*"

5. (10%) Given the following examples, we want to learn to extract the area codes using the Stalker algorithm. Give the learned start rules. You should not produce more than two rules (or disjuncts)

1: <li> 205 Willow, Glen, Phone - 773-366-1987
2: <li> 25 Oak, Forest, Phone 800 234-7903
3: <li> 324 Halsted St., Chicago, Phone 800-996-5023
4: <li> 700 Lake St., Oak Park, Phone 708 798-0008 </p>

6. (10%) Use both complete-link and single-link agglomerative clustering to cluster the following one dimensional data points: 1, 2, 5, 7, 11, 20, 24, 28, 33, 38. You are required to draw two separate cluster trees. Use Euclidean distance as the distance measure.

7. (10%) Given the probabilistically labeled data in Table 1 and unlabeled data in Table 2, which has two attributes A and B, and a class C with two classes {1, -1}, we want to use EM to perform semi-supervised learning. Naïve Bayesian (NB) is used as the base classifier for EM (ignore smoothing in NB). Fill in the probabilistic labels of class C for each tuple in Table 2, and list all the specific probabilities needed to build the NB classifier. Note that you should use the naïve Bayesian classifier for normal tableau data rather than for text.

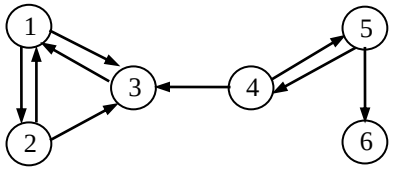
A	B	C ($Pr(1\|d)$)
k	m	0.1
k	s	0.2
g	q	0.8
k	q	0.8
g	q	0.8
g	s	0.2
g	s	0.2
g	m	0.3
g	q	0.8
k	m	0.1

(1) Probabilistically labeled data

A	B	C ($Pr(1\|d)$)
k	m	
k	s	
g	s	
g	q	

(2) Unlabeled data and NB results

8. (10) Given the hyperlink graph below



We want to use the Markov chain model to compute the PageRank value of each node.

a. Give the initial transition probability matrix.

b. If the transition probability matrix is not a stochastic matrix, convert it to a stochastic matrix by using the second method (adding artificial links). Show the resulting matrix.

c. If the resulting matrix is not irreducible, convert it to an irreducible matrix. Show the matrix with $d = 0.9$.