CS685G-Spring 2019

Final Exam

Notes:

(1) 2 hour exam

(2) Please write, write legibly. If I can't read it, you can't get credit for it.

I pledge that I have neither received nor given unauthorized aid on this examination.

Start time: _____

End time: _____

Signed:

_____

Printed Name:

_____

Problem 1 (10 points): Discuss whether or not each of the following activities is a data mining task

(a) Dividing the customers of a company according to their profitability.

Answer:

(b) Predicting the outcomes of tossing a (fair) pair of dice.

Answer:

(c) Predicting the future stock price of a company using historical records.

Answer:

(d) Monitoring the heart rate of a patient for abnormalities.

Answer:

(e) Extracting the frequencies of a sound wave.

Answer:

Problem 2 (20 points): True or False

Answer	No	Statement
	1	Mean is more robust against outliers than median
	2	The selection of the initial centroids for K-means does not affect the final clustering results.
	3	The FP-growth algorithm only requires 2 scans of the transactional database
	4	The set of all maximal frequent sets is a subset of the set of all closed frequent itemsets
	5	Density-based clustering algorithm generally performs better than k-means in the case of irregular cluster shapes
	6	The only acceptable input for clustering algorithms is the data matrix
	7	The time complexity of the apriori algorithm is linear to the number of transactions
	8	Decision tree is a clustering model
	9	Knowledge discovered in the KDD process should always confirm the expectation of domain expert
	10	The higher support an association rule has, the more confidence it is and the more interesting it is.
	11	In general, all features that are used to describe an object are equally important for its classification
	12	Hierarchical clustering is considered to be the most efficient clustering algorithm
	13	Clustering is much more useful than classification in predicting class labels given a training dataset
	14	Naïve Bayes Classifier operates under the assumption that every attribute is dependent upon each other
	15	Decision tree algorithm does not scale with large dataset (>10,000) due to its high computational complexity
	16	Multidimensional scaling (MDS) is a dimensionality reduction technique
	17	The first component of principle component analysis (PCA) of a dataset depicts the dimension with minimum variance.
	18	Overfitting in decision tree means the tree is too big
	19	Cross validation is a method to validate the accuracy of a classification model
	20	Recall for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class

Problem 3 (15 points): A database has 6 transactions as shown

TID	Transaction
1	A, B, C
2	A, C
3	A, B, C D
4	A, C, D
5	A, D
6	A, C, D

(1) Given a support threshold of 50%, please draw the level-wise search lattice for the apriori-based algorithm, where each node represents a *frequent* itemset, and there exists an edge between two nodes if one is a subset of the other. For each node, please label it with the composition of itemset and its support.

(2) Please write down one potential order on how these itemsets might be visited using the apriori-based algorithm.

(3) Please write down one potential depth-first traversal order on how these itemsets might be visited. Alternatively, you may draw the FP-tree and the order to traverse the tree.

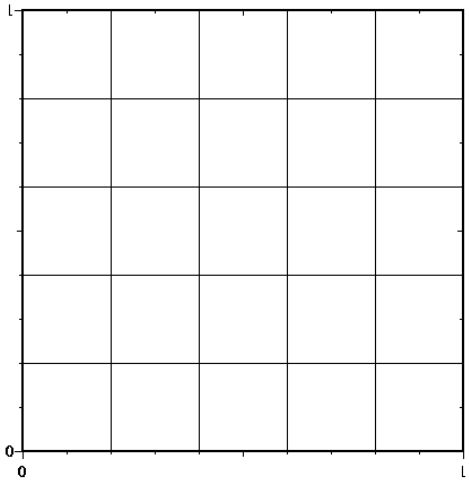
(4) Find all maximal frequent itemsets and closed frequent itemsets in this dataset.

(5) Please compute the confidence of the association rule {AC->D}?

Problem 4 (10 points): There are 5 points in a 2-D space. Here are their coordinates

Points	x	Y
A	4	1
B	4	2
C	3	3
D	2	5
E	1.5	4

(1) (1 points) Draw the above points on a 2-D grid.



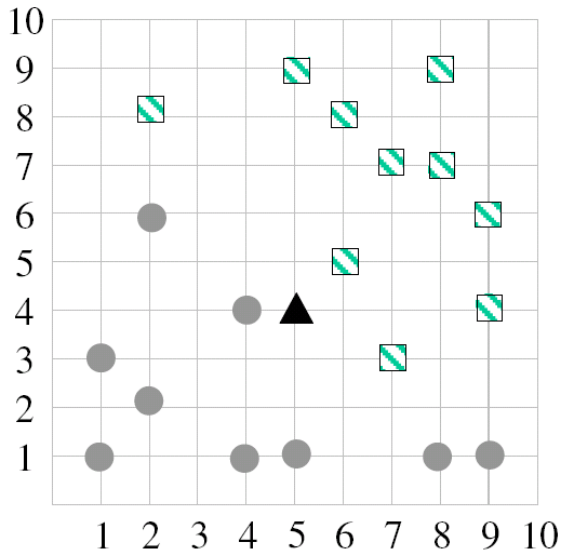
(2) (3 points) Assume Euclidean distance is used for distance measure. Apply the complete-linkage based hierarchical clustering. Please draw the dendrogram of the clustering results and put down the value of linkage used at each merge.

(3) (3 points) With the same distance measure, apply the single-linkage based hierarchical clustering, and draw the dendrogram of the clustering results with the linkage value used at each merge.

(4) (3 points) Given a random dendrogram of the same data points A, B, C, D and E, please name one metric that can be used to evaluate the quality of each dendrogram and explain how.

Problem 5 (20 points): The following figure is a 2D plot of a number of points belonging to two classes: the squares and solid circles. Now there comes another point, marked with a triangle. We would like to classify it into one of the two classes. (Note that their coordinates are all integers)

1. Please list one or more classification algorithm that might be suitable for this type of classification?



2. One approach we have not learned in class is called K-nearest neighbor classification. Basically, the algorithm should retrieve the k-nearest neighbors of the to-be classified point, and classify the point into the class to which the majority of its neighbors belong. What is the labeling of the triangle points for k=1, 3, 5, 7?

3. Assume you are given an input matrix X containing the coordinates of all the points. Please write down the pseudocode for a k-nearest neighborhood classification and analyze its complexity as a function of the number of data points ***n*** and ***k.***

4. Please analyze the complexity of your k-nearest neighborhood algorithm.

5. Please explain the pros and cons of having too small and too large a value for *k* in terms of accuracy and complexity.

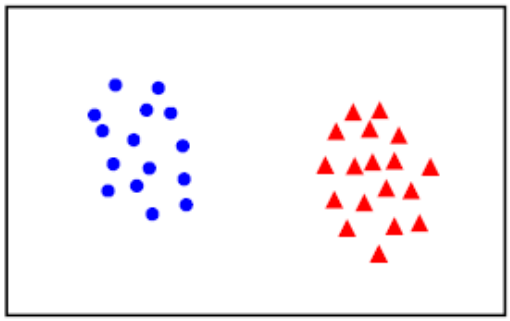
6. For a given dataset, what method you may use to select the *k* that achieves the best predication accuracy?

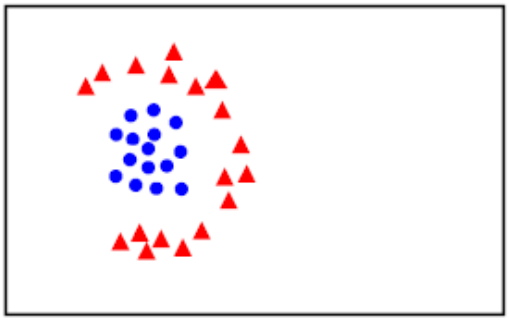
Problem 6 (15 points): Please answer the following questions.

1. (a) Draw a 2-dimensional dataset with two clusters that can be discovered with 100% accuracy by DBSCAN or single linkage hierarchical clustering, but not by k-means.

   (b) Draw a 2-dimensional dataset with two clusters that can be discovered with 100% accuracy by k-means, but not by DBSCAN. Please explain why.

2. Please propose a classification algorithm that is suitable for the classification of circle and triangle points in each of the plots below.





Problem 7 (10 points): Free response. Assuming that you are a data scientist in a financial company and you receive a binary data matrix where each row corresponds to a customer and each column corresponds to yes/no (1/0) answer from a survey consisting of 20 questions about the service provided by the company. In total, the company has received a little under 50, 000 survey results.

1) What are the information you may extract from such a dataset? and what potential data mining techniques can be applied in order to achieve that?

2) (5 Bonus Points). Apparently, some customer does not feel like answering a particular question, so they leave it blank. How does it impact the analysis or what do you propose to address this problem?