Department of Mathematics
II Semester 2018-2019
MTL 782 Data Mining
Major Examination Date: 1.5.19 Time: 8 AM – 10 AM
Weightage: 35%

Q1. Suppose we have a neural network, with 1 hidden layer and 1 output layer with 2 nodes. Is it advisable that the hidden layer in neural network should have roughly half the nodes of the input layer. If yes, why it could be possibly so? [2]

Q2. Assume that you have to explore a large data set of high dimensionality. You know nothing about the distribution of the data. Answer the following. [3]

i. How can k-means and DBSCAN be used to find the number of clusters in that data?

ii. Explain how PCA can help find the dimensions where clusters separate.

iii. Explain why PCA might neglect cluster separation in some dimensions.

Q3. Calculate the cosine, correlation, Jaccard, and Extended Jaccard similarity/distance for the vectors x = (1, 1, 0, 1, 0, 1) and y = (1, 1, 1, 0, 0, 1). [2]

Q4. Discuss the differences between dimensionality reduction based on aggregation and dimensionality reduction based on PCA. [2]

Q5. The following table summarizes a data set with three attributes A, B, C and two class labels +, -. [2+2+1]

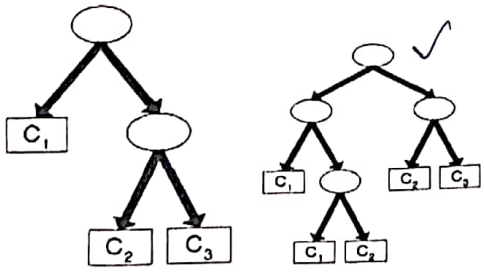
A	B	C	Number of Instances	
			+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

(a) According to classification Error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gain in classification error rate.

(b) Repeat for the two children of the root node.

(c) How many instances are misclassified by the resulting decision tree.

Q6. Consider the decision trees shown in following Figure (a) and (b) . Assume they are generated from a data set that contains 16 binary attributes and 3 classes, C_1 , C_2 , and C_3 . Compute the total description length of each decision tree according to the minimum description length principle.



(a) Decision Tree With 7 Errors (b) Decision Tree with 4 Errors

- The total description length of a tree is given by: Cost (tree, data) = Cost (tree) + Cost (data|tree).
- Each internal node of the tree is encoded by the ID of the splitting attribute. If there are m attributes,
- the cost of encoding each attribute is $\log_2 m$ bits.

- Each leaf is encoded using the ID of the class it is associated with. If there are k classes, the cost of encoding a class is $\log_2 k$ bits.
- Cost(tree) is the cost of encoding all the nodes in the tree. To simplify the computation, you can assume that the total cost of the tree is obtained by adding up the costs of encoding each internal node and each leaf node.
- Cost(data|tree) is encoded using the classification errors the tree commits on the training set. Each error is encoded by $\log_2 n$ bits, where n is the total number of training instances.

Which decision tree is better, according to the MDL principle? [5]

Q7. What is feature selection? Can Decision tree be used for feature selection? Explain. [3]

Q8. Consider the set of points given in Figure 1. Assume that eps = $\sqrt{2}$ and minpts = 3 (including the center point). Using Euclidian Distance find all the density-based clusters in the figure using the DBSCAN algorithm. List the final clusters (with the points in lexicographic order, i.e., from A to J) and outliers. [3]

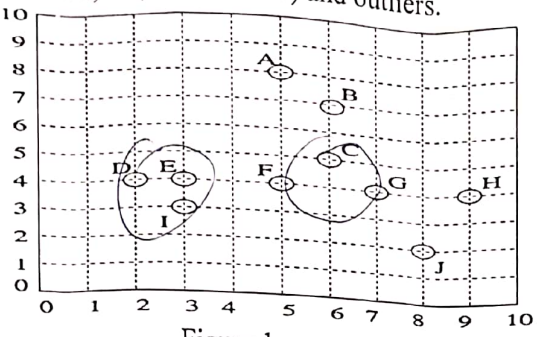


Figure 1

Q9. Consider the following set of frequent 3-itemsets:

{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}. [3]

Assume that there are only five items in the data set.

(a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

(b) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori .

(c) List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

Q10. Many statistical tests for outliers were developed in an environment in which a few hundred observations was a large data set. We explore the limitations of such approaches.

(a) For a set of 1,000,000 values, how likely are we to have outliers according to the test that says a value is an outlier if it is more than three standard deviations from the average? (Assume a normal distribution.)

(b) Does the approach that states an outlier is an object of unusually low probability need to be adjusted when dealing with large data sets? If so, how? [1 +1]

Q11. Consider the data set shown in the table below:

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

Table: Example of market basket transactions.

(a) Compute the support for itemsets {e}, {b,d}, and {b,d,e} by treating each transaction ID as a market basket.

(b) Use the results in part (a) to compute the confidence for the association rules {b,d} → {e} and {e} → {b ,d}. Is confidence a symmetric measure?

(c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

(d) Use the results in part (c) to compute the confidence for the association rules {b, d} → {e} and {e} → {b,d,}.

(e) Suppose s_1 and c_1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s_2 and c_2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s_1 and s_2 or c_1 and, c_2. [1+1+1+1+1]