B.Tech (Computer Engineering), 7th Semester, Annual Examination 2020
Data Mining
Paper Code: CEN - 701

Max. Marks: 60 Time: 3 Hours

Instruction to the candidates:

Attempt Any Two parts from each question.
Each part of the question carries 6 marks.
Write your roll no. & Name on top of each page of your answer sheet.
Write your answer in your own words. Answer matching with other classmates will not be marked.

1(a) Indicate whether the following statement is true or false. Give reasons for your answer. If the statement is false then change the statement to make it true.

i) The chi-square distribution is left skewed.
ii) Outliers are influential observations
iii) The principal components are always mutually exclusive and exhaustive of the variables
iv) A 2-nearest neighbor model is more likely to overfit than a 20-nearest neighbor model.
v) With k-fold cross-validation, larger k is always better.
vi) Overfitting is a danger when learning a classifier, but not when doing unsupervised learning.

(b) For each of the following questions, provide example of an association rule from two different domains that satisfies the following conditions. Also, describe whether such rules are subjectively interesting or not.

i) A rule that has high support and high confidence.
ii) A rule that has reasonably high support but low confidence.
iii) A rule that has low support and low confidence.
iv) A rule that has low support and high confidence.

(c) Consider the vertical dataset shown below. Assuming minimum support = 3 find all frequent itemsets using FP-tree.

A	B	C	D	E
1	2	1	1	2
3	3	2	4	3
5	4	3	6	4
6	5	5		5
	6	6		

2.(a) Most frequent pattern mining algorithms consider only distinct items in a transaction. However, multiple occurrences of an item in the same shopping basket, such as four cakes and three chocolates, can be important in transactional data analysis. How can one mine frequent itemsets efficiently considering multiple occurrences of items? Propose modifications to the Apriori, to adapt to such a situation.

(b) i) Let 'C' be the set of all closed frequent itemsets and 'M' the set of all maximal frequent itemsets for some database. Prove that $M \subseteq C$.

ii) Give a good example with at least 5 attributes and 10 instances and find the closed and max patterns from it.

(c) An algorithm has to designed which will display only those frequent items which lies between MaxSupport and MinSupport whose values are supplied by the user. Modify any frequent pattern mining algorithm such that the frequent items are displayed between these two values in decreasing value of their support.

3.a) Consider the following dataset. Compute the information gain of attribute a1 and a2. If we want to find the information gain of a3 what should be done?

Instance	**a1**	**a2**	**a3**	**Class**
1	T	T	5.0	Y
2	T	T	7.0	Y
3	T	F	8.0	N
4	F	F	3.0	Y
5	F	T	7.0	N
6	F	T	4.0	N
7	F	F	5.0	N
8	T	F	6.0	Y
9	F	T	1.0	N

(b) i) A binary classifier achieves 95% accuracy on a test set consisting of 95% positive and 5% negative instances. If we use the same classifier on a test set composed of 50% positive and 50% negative instances, what can we say on the accuracy of the classifier.

ii) How boosting improves the accuracy of the classifier? Use a proper example to illustrate your answer.

(c) Given below is the sample dataset for classification of mammals and non-mammals. Classify the test date whose values are **Give birth = Yes, Can Fly = No, Live in Water = Yes, Have Legs = No.**

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals

4.a) Given two examples of Bayesian Belief Network with proper diagram. Using any one of them explain how this network is used to calculate the probabilities.

b) Consider the training examples shown below for a binary classification problem.

i) What is the entropy of this collection of training examples with respect to the positive class?

ii) What are the information gains of a1 and a2 relative to these training examples?

iii) What is the best split (between a1 and a2) according to the classification error rate?

Instance	a1	a2	a3	Class
1	**T**	**T**	**3**	+
2	**T**	**T**	**6**	+
3	**T**	**F**	**5**	-
4	**F**	**F**	**4**	+
5	**F**	**T**	**7**	-
6	**F**	**T**	**4**	+
7	**F**	**F**	**8**	-
8	**T**	**F**	**7**	+
9	**F**	**T**	**6**	+
10	**F**	**F**	**9**	-

(c) Using decision tree classifier what should be the root of the decision tree for the following data whose classifying attribute is "Class".

Number	Outlook	Temperature	Humidity	Windy	Class
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Overcast	Mild	High	True	P
12	Overcast	Hot	Normal	False	P

5.a) Hierarchical clustering is sometimes used to generate K clusters, K > 1 by taking the clusters at the K^{th} level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.

The following is a set of one-dimensional points: {6, 10, 18, 24, 32, 42, 48, 54}.

For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.

i) (18,40) ii) (15,45) (iii) (10, 32)

(b) For a given k=2, cluster the following data set using K-mediod clustering algorithm.

Point	X-Axis	Y-Axis
1	7	6
2	2	6
3	3	8
4	8	5
5	7	4
6	4	7
7	6	2
8	7	3
9	6	5
10	3	5

(c)

i) How can use DBSCAN algorithm to detect outliers?

ii) You are given a data set with 100 records and are asked to cluster this data using K-means algorithm. It is found that for all values of K, 1<=K<=100, the K-means algorithm returns only one non-empty cluster? What is the reason for this.
How would DBSCAN handle such data?