Section A

Q1 a) Find the Euclidean distance between data points X (0,-1,0,1) and Y (1,0,-1,0). 2

b) If recall and precision are 0.5 and 0.6 respectively, compute the value of F_1 measure. 2

c) In a given dataset, it is found that an itemset {ab} is infrequent. Will itemset {abc} be infrequent or frequent? Explain why. 2

d) What are the three strategies for handling missing values in a dataset? 3

e) Differentiate between precision and bias on the basis of the quality of the measurement process. 3

f) What is meant by variable transformation? What are its advantages? 3

g) If support of an association rule X -> Y is 80% and confidence is 75%, can we derive support and confidence of the rule Y -> X? If yes, list down the values. If no, state the reason. 3

h) List down two advantages and two disadvantages of leave-one-out approach used in cross-validation for evaluating the performance of the classifier? 4

i) Differentiate between agglomerative and divisive methods of hierarchical clustering with the help of a diagram. 4

j) What are asymmetric attributes? Give an example each of 4

i) asymmetric binary attribute,
ii) asymmetric discrete attribute,
iii) asymmetric continuous attribute.

k) The confusion matrix for a 2-class problem is given below: 5

		Predicted Class	
		Class=1	Class=0
Actual Class	Class=1	400	100
	Class=0	200	300

Calculate the accuracy, sensitivity, specificity, True Positive Rate, and False Positive rate.

			Section B	
Q2	a)		What are the differences between noise and outliers? Are noise objects always outliers? Are outliers always noise objects?	2+1+1
	b)		What is unsupervised learning? Explain with the help of an example application.	2
	c)		Why is K-nearest neighbor classifier a lazy learner?	4
Q3	a)		What is an exhaustive ruleset in Rule based classification? If the ruleset is not exhaustive, what problem arise? How is it resolved?	4
	b)		What is progressive sampling? What are its advantages?	3
	c)		State Bayes' theorem. What assumption is used by the Naïve Bayes classifier?	3
Q4	a)		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}. Assume that there are only five items in the data set. i. List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy. ii. List all candidate 4-itemsets obtained by a candidate generation procedure in Apriori.	6
	b)		Let X denote the categorical attribute having values {awful, poor, OK, good}. What is the representation of each value when X is converted to binary form using i) 2 bits ii) 4 bits?	4
Q5			Consider the following transactional dataset:	8

Transaction ID	**Items Bought**
0001	{a, d, e}
0002	{a, b, c, e}
0003	{a, b, d, e}
0004	{a, c, d, e}
0005	{b, c, e}
0006	{b, d, e}
0007	{c, d}
0008	{a, b, c}

			0009 {a, d, e}
			0010 {a, b, e}
			i. Find out the support of itemsets {e}, {b,d} {a,d} and {b,d,e}. Are these itemsets frequent if minimum support threshold is 30%?
			ii. Find all the rules generated from the 3-itemset {b,d,e}. List down the strong rules among these rules if minimum confidence threshold is 60%.	

	b)		What is the difference between nominal attributes and ordinal attributes? Give an example of each.	2
Q6	a)	i.	Explain the following terms with reference to the DBSCAN clustering algorithm. i) Core point ii) Noise point iii) Border Point	6
	b)		Given the following data points: 2, 4, 10, 12, 3, 20, 30, 11, 25. Assume K = 3 and initial means 2, 4, 6. Show the clusters obtained using K-means algorithm after two iterations and show the new means for the next iteration.	4