2323476800005L

Master of Computer Applications
MCAC 105: Data Mining
Unique Paper Code: 223421106
Semester I
January 2024
Year of Admission: 2023

Time: Three Hours **Max. Marks: 70**

Instructions:
1. All questions are compulsory.
2. Attempt all the parts of a question **together**.

1. Consider the following dataset. 10

Data Points	X1	X2
A	7	2
B	9	13
C	2	11
D	8	2
E	3	12
F	5	5

Perform Agglomerative clustering algorithm using
   a. single-link and
   b. average-link.

Also, show the step-wise construction of the dendrogram of the single-link-based Agglomerative clustering.

2. Consider the dataset given below: 9

Humidity	Outlook	Windy	Play
Low	Overcast	No	Yes
Low	Overcast	Yes	Yes
High	Sunny	No	Yes
High	Sunny	Yes	No
Low	Sunny	No	No
Low	Sunny	Yes	Yes
High	Overcast	No	No

Apply k-Nearest-Neighbour (k-NN) classification (k = 3) to classify the tuple <Humidity=Low, Outlook=Overcast, Windy=No>.

3. Consider the data values: {200, 300, 400, 600, 800, 1000}. Normalize using the following: 9
   (i) min-max normalization by setting *min = 0* and *max = 1*
   (ii) z-score normalization
   (iii) normalization by decimal scaling

4. Given the following dataset, apply the ECLAT algorithm to find the frequent itemset using minimum support count = 2. Also, find the closed frequent patterns from frequent itemsets. Show all the intermediate steps. 10

T_{id}	Items
101	Butter, Coffee, Milk
102	Bread, Butter, Coffee, Eggs, Milk, Sugar
103	Bread, Coffee, Eggs, Sugar
104	Bread, Coffee, Eggs, Milk, Sugar
105	Coffee, Eggs, Milk, Sugar
106	Bread, Butter, Coffee, Milk, Sugar
107	Bread, Butter, Milk
108	Coffee, Eggs, Sugar

5. Consider the dataset comprising four features, namely F1, F2, F3, and F4. Perform dimensionality reduction using Principal Component Analysis (PCA). 9

F1	F2	F3	F4
12	8	14	4
8	2	6	10

6. Construct the decision tree using the Iterative Dichotomizer 3 (ID3) classification algorithm. Show all the intermediate steps. 9

Tid	Attribute1	Attribute2	Attribute3	Class
1	Yes	Large	125	No
2	No	Medium	100	No
3	No	Small	70	No
4	Yes	Medium	120	No
5	No	Large	95	Yes
6	No	Medium	60	No
7	Yes	Large	220	No
8	No	Small	85	Yes
9	No	Medium	75	No
10	No	Small	90	Yes

7. Consider the following data set: 9

Data Points	X_1	X_2
P1	2	4
P2	4	6
P3	8	2
P4	10	6

Assuming that k =2 and initial cluster centres for k-means clustering are P1 and P2, compute the cluster centres. What are the drawbacks of using the k-means algorithm?

8. Suppose you are working with a dataset representing customer preferences for different product categories in an e-commerce platform. The dataset contains binary information, where 1 indicates that a customer has purchased a product in a specific category, and 0 indicates no purchase. You want to use the Jaccard measure to compare the similarity of two customer profiles. 5

Customer A's preferences:

[1, 0, 1, 1, 0, 1, 0]

Customer B's preferences:

[0, 1, 1, 1, 0, 0, 1]

Calculate the Jaccard similarity coefficient between Customer A and Customer B based on their preferences for these product categories. How will you interpret the result in the context of customer preferences? Also, discuss how the Jaccard measure can help understand the similarity between different customer profiles.