2323476800016

Master of Computer Applications
MCAO 302: Data Science using Python
Unique Paper Code: 223423304
Semester III
December 2024
Year of Admission: 2023

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

Note: All questions are compulsory. Attempt all the parts of a question together.

1. (a) Consider the following binary class dataset having two attributes A and B: [6]

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Calculate the gain using Gini index when splitting on A and B. Which attribute would be chosen as a decision attribute at the root node? Show all intermediate steps and calculations.

(b) Consider the age of twelve students with the following values: [4]

25, 26, 29, 30, 32, 33, 41, 42, 48, 50, 81, 86

Apply the following approach to map the values to **four** bins.

i. Equal width
ii. Equal frequency binning

(c) Compute Jaccard similarity and Cosine similarity between the random variable **x** and **y**. [4]

$$\mathbf{x} = [1\,1\,1\,1\,1\,0\,0\,0\,1\,1]$$
$$\mathbf{y} = [0\,1\,1\,0\,1\,0\,0\,0\,1\,0]$$

2. (a) An exit poll conducted by an agency collected data from a representative random sample of 6 voters and asked them if they voted for xyz. If the true percentage of voters who vote for xyz is 55.1%, what is the probability that, in your sample, exactly 2 voted for xyz and 4 did not? [2]

(b) Define Poisson probability mass function. Also, find expectation of a random variable X that follows Poisson distribution. [4]

(c) Consider the following dataset of a small hospital's annual revenue (in lakh ₹) over years.

Year	2010	2011	2013	2014	2016	2017	2018	2019	2020
Revenue (in lakh ₹)	1.2	1.15	1.25	1.3	1.4	1.35	1.45	1.5	1.55

i. Use the least squares estimation method to determine the linear relationship between `Year` and `Revenue`. Using this equation, predict the hospital's revenue for the year 2012. [6]

ii. Also, compute Pearson's corelation coefficient between `Year` and `Revenue`. [2]

3. (a) Consider the following transaction dataset:

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	$\{a, b, e\}$

i. Compute the support for itemsets `{e}`, `{b, d}`, and `{b, d, e}` by treating each transaction ID as a market basket. [3]

ii. Compute the confidence and lift for the association rules `{b, d}` → `{e}` and `{e}` → `{b, d}`. [4]

iii. Is confidence a symmetric measure? [1]

(b) Consider the following set of frequent 3-itemsets: [6]

L_3 = {{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.

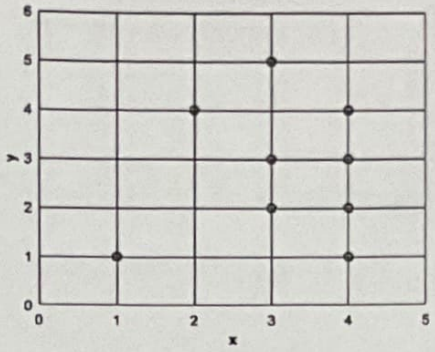
i. List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.

ii. List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm. Justify your answer.

4. (a) Explain CLARA clustering algorithm and compute its time complexity. [4]

(b) Explain the data structure, cluster representation and parameters used in BIRCH algorithm. [4]

(c) Consider the data shown in the following figure. Using the concept of DBSCAN algorithm, examine the data points and identify them as core points or border point or noise point. Use Manhattan distance to calculate the distance between two points, $\epsilon = 2$ and $minPts = 3$ (excluding self). Show all intermediate steps. [6]



5. (a) Perform stemming and lemmatization on the following text data: [3]
"Natural language processing is a fascinating area of computer science"

(b) Consider the following table representing a small corpus with three documents: [5]

Document	Content
Doc1	"Data science is fun and interesting"
Doc2	"Science and data are important for learning"
Doc3	"Learning Python for data science is rewarding"

For the words data and learning, compute the following in each document and interpret the resultant values.

i. Term Frequency (TF)
ii. Term Frequency × Inverse Document Frequency (TF×IDF)

(c) Consider a binary classification model which is used for prediction of class label of 10 samples present in the test dataset. The actual class labels and the predicted labels are shown in the following table. [6]

Actual Label	1	1	1	1	1	0	0	0	0	0
Predicted Label	1	0	1	1	0	0	0	1	0	1

Construct the confusion matrix, and compute the accuracy, sensitivity, specificity, and precision.