



LOYOLA COLLEGE (AUTONOMOUS) CHENNAI – 600 034

M.Sc. DEGREE EXAMINATION – STATISTICS

THIRD SEMESTER – NOVEMBER 2024

PST3MC03 – DATA MINING AND MACHINE LEARNING



Date: 12-11-2024

Dept. No.

Max. : 100 Marks

Time: 01:00 pm-04:00 pm

SECTION A – K1 (CO1)

Answer ALL the questions

(5 x 1 = 5)

1 Fill in the blanks

- The last phase in CRISP-DM is the _____.
- _____ is a graphical method of identifying outliers.
- A decision tree is used for _____ and _____ tasks.
- A logistic regression tree is a machine learning method that partitions the data and fits an _____ model in each partition.
- The three operators used by Genetic Algorithms are _____ crossover, and mutation.

SECTION A – K2 (CO1)

Answer ALL the questions

(5 x 1 = 5)

2 Answer the following

- Define Unsupervised Learning Algorithms.
- What is information gain? Give the formula and the terminologies used.
- Define Naïve Bayes algorithm.
- What is Binary classifier in Machine Learning?**
- What is text mining?

SECTION B – K3 (CO2)

Answer any THREE of the following

(3 x 10 = 30)

- Explain the three methods of data transformation using suitable examples.
- How does Random Forest work? List the key differences between Random Forest and other machine learning algorithms?
- What are Support Vector Machines? Elaborate how they work.
- Explain the construction of Bayesian network with suitable example.
- Outline the foundations of genetic algorithms, search space, fitness score and operators.

SECTION C – K4 (CO3)

Answer any TWO of the following

(2 x 12.5 = 25)

- Explain all the steps in data pre-processing in detail.
- Explain the k Nearest neighbor algorithm with special reference to k and the distance functions
- What is back propagation in machine learning? Explain the working of back propagation algorithm.
- What are the similarities and differences between Bagging and Boosting algorithms in Machine Learning?

SECTION D – K5 (CO4)

Answer any ONE of the following

(1 x 15 = 15)

- What are Neural Networks? List out their applications and explain the components in detail.
- Explain the concept of association rule mining from large datasets. Elaborate on the algorithms used in association mining with advantages and disadvantages.

SECTION E – K6 (CO5)

Answer any ONE of the following

(1 x 20 = 20)

- 14 Grow a decision tree using the C 4.5 algorithm, for the following training set records.

Customer	Savings	Assets	Income (\$1000s)	Credit Risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

- 15 Construct a Naïve Bayes Classifier to predict the probability of “Not playing cricket” given that the weather conditions are “Rainy outlook”, “Temperature is hot”, “high humidity” and “no wind”.

Outlook	Temperature	Humidity	Windy	Play Cricket
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No

\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$