SASTRA

DEEMED TO BE UNIVERSITY

School of Computing

FIRST CIA Test – Aug 2025

Course Code: INT317

Course Name: DATA MINING AND ANALYTICS

Duration: 90 minutes Max Marks: 50

## PART - A

**Answer All the Questions** **5X10 =50 Marks**

1.A dataset contains two features $X_1$ and $X_2$, recorded for four entities:

| Features | $E_1$ | $E_2$ | $E_3$ | $E_4$ |
|---|---|---|---|---|
| $X_1$ | 2 | 3 | 4 | 5 |
| $X_2$ | 4 | 6 | 8 | 10 |

Perform a Principal Component Analysis to reduce the dataset from two dimensions to one dimension.

2. A database has 10 transactions with min_support=40% and min_confidence=70%

| **TID** | **Items Purchased** |
|---|---|
| T1 | {Milk, Bread, Eggs} |
| T2 | {Milk, Bread, Butter} |
| T3 | {Bread, Butter, Jam} |
| T4 | {Milk, Bread, Butter, Eggs} |
| T5 | {Bread, Butter} |
| T6 | {Milk, Eggs} |
| T7 | {Milk, Bread, Butter, Jam} |
| T8 | {Bread, Butter, Eggs} |
| T9 | {Milk, Bread, Jam} |
| T10 | {Bread, Butter, Eggs, Jam} |

a) Find all frequent item set using Apriori Algorithm (5)

b) List all the strong association rules (5)

3. Generate the frequent pattern from the following data set using FP growth, where minimum support = 3.

| TID | Items Bought |
|---|---|
| 100 | f, a, c, d, g, i, m, p |
| 200 | a, b, c, f, l, m, o |
| 300 | b, f, h, j, o |
| 400 | b, c, k, s, p |
| 500 | a, f, c, e, l, p, m, n |

4. (i)The dataset below shows three numerical attributes for 5 products:

| Product | Weight (kg) | Price ($) | Rating (1–5) |
|---|---|---|---|
| P1 | 5 | 500 | 4.5 |
| P2 | 12 | 1500 | 3.8 |
| P3 | 8 | 1000 | 4.2 |
| P4 | 20 | 3000 | 2.5 |
| P5 | 15 | 2000 | 3 |

a) Apply Min–Max normalization to transform the Weight value 20 in the range [0.2, 0.8]. (2)

b) Apply Z-score normalization to transform the Price value 1000. (3)

(ii) Explain the steps involved in KDD Process (5)

5. You are provided with a dataset of patient medical records that contains missing values, duplicate records, and noisy data. Explain the preprocessing steps you would implement to prepare the dataset for predictive modeling.

SASTRA
DEEMED TO BE UNIVERSITY
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA
THANJAVUR · KUMBAKONAM | CHENNAI

School of Computing
Second CIA Exam –Sept 2025
Course Code: INT317
Course Name: DATA MINING AND ANALYTICS
Duration: 90 minutes Max Mark: 50

## PART A

**Answer any FIVE of the following questions** **5x10=50**

1.Find the naïve bayes probability computation on the given data for the test instance X= {Weather:” rainy”, Road condition:” Good”, Traffic:” Normal”, Engine Issue: No}.

| Weather | Road Condition | Traffic | Engine Issue | Accident |
|---|---|---|---|---|
| Rainy | Bad | High | No | Yes |
| Cloudy | Average | Normal | Yes | Yes |
| Clear | Bad | Light | No | No |
| Clear | Good | Light | Yes | Yes |
| Cloudy | Good | Normal | No | No |
| Rainy | Average | Light | No | No |
| Rainy | Good | Normal | No | No |
| Cloudy | Bad | High | No | Yes |
| Clear | Good | High | Yes | No |
| Clear | Bad | High | Yes | Yes |

2.We have two features of the following data points:

X1: (2.5, 1.5, 1.7, 1.9, 2.9, 2.3, 2.8,1.6)

X2: (545, 438, 489, 429, 528, 503, 563, 445)

the corresponding target values: (1, 0, 0, 0, 1, 1, 1, 0), using the K-Nearest Neighbors algorithm with k=5, determine the target value for the new data point A= (1.8,415) by calculating the Euclidean distances between A and each of the given data points.

3. Dr. Bob is developing a model to predict whether patients have cancer based on their medical data. After training the model, the results for a group of 3895 patients are summarized. out of the 3895 patients, 368 were diagnosed with cancer, and 3527 were healthy. The model correctly identified 266 patients with cancer and 3419 healthy patients. However, 102 patients with cancer were misclassified as healthy, and 108 healthy

patients were incorrectly predicted to have cancer. Plot the confusion matrix and infer the performance measures of the model.

4. Describe the Generalized Linear Model (GLM) framework and explain the commonly used link functions for the following probability distributions:

i) Poisson
ii) Binomial
iii) Inverse Binomial
iv) Gamma.

5. Find the Root node of the decision tree for the following Dataset

| Age | Cough | Fever | Cold | Viral Infection |
|---|---|---|---|---|
| Youth | High | No | No | No |
| Youth | High | No | Yes | No |
| Adult | High | No | No | Yes |
| Senior | Medium | No | No | Yes |
| Senior | Low | Yes | No | Yes |
| Senior | Low | Yes | Yes | No |
| Adult | Low | Yes | Yes | Yes |
| Youth | Medium | No | No | No |
| Youth | Low | Yes | No | Yes |
| Senior | Medium | Yes | No | Yes |
| Youth | Medium | Yes | Yes | Yes |
| Adult | Medium | No | Yes | Yes |
| Adult | High | Yes | No | Yes |
| Senior | Medium | No | Yes | No |

6. Describe the iterative methods used in Nonlinear Least Squares (NLS) estimation, specifically:

i) Grid Search
ii) Newton–Raphson Method
iii) Steepest Descent Method
iv) Marquardt's Method.

| | |
|---|---|
| SASTRA<br>DEEMED TO BE UNIVERSITY<br>THINK MERIT \| THINK TRANSPARENCY \| THINK SASTRA<br>THANJAVUR \| KUMBAKONAM \| CHENNAI | School of Computing<br>Third CIA Exam – Nov 2025<br>Course Code: INT317<br>Course Name: DATA MINING AND ANALYTICS<br>Duration: 90 minutes Max Mark: 50 |

LJC -318 / DMA-18

## PART A

**Answer any FOUR of the following questions** **4x10=40**

1. Create the distance matrix between the following data points such as O1, O2, O3 and O4.

| | | | | |
|---|---|---|---|---|
| O1 | 12 | 23 | 12 | 23 |
| O2 | 45 | 12 | 45 | 87 |
| O3 | 12 | 45 | 21 | 45 |
| O4 | 24 | 54 | 28 | 37 |

2. A database has five transactions. Let min sup = 60% and min.conf=80%.

| Transaction id | Items |
|---|---|
| T100 | {N, P, O, L, F, Z} |
| T200 | {E, P, O, L, F, Z} |
| T300 | {N, B, L, F} |
| T400 | {N, V, D, L, F, Z} |
| T500 | {D, P, L, J, F} |

(a) Find all frequent item sets using Apriori algorithm

(b) List all the strong association rules.

3. Compare Linear and Logistic regression. Derive the equation for sigmoid function in logistic regression.

4. Describe with suitable equations and diagrams, how ARIMA models overcome the limitations of ARMA models when applied to non-stationary time series data.

5.Create the dissimilarity matrix between the items using a simple distance measure based on the various forms of data. The ordinal traits are ranked as follows Excellent – 1, Average – 2, and Bad – 3.

| Object | Attribute 1 (Nominal) | Attribute 2 (Ordinal) | Attribute 3 (Numerical) |
|---|---|---|---|
| $O_1$ | B1 | Excellent | 40 |
| $O_2$ | B2 | Average | 55 |
| $O_3$ | B1 | Bad | 72 |
| $O_4$ | B3 | Excellent | 48 |

6. A bakery sells muffins at different prices each week. The baker records the price per muffin (x pence) and the number of muffins sold (y) during six consecutive weeks:

| x(pence) | 12 | 18 | 24 | 30 | 36 | 42 |
|---|---|---|---|---|---|---|
| y (sold) | 95 | 82 | 70 | 58 | 45 | 38 |

i Calculate the least square regression line y on x.

ii Predict the number of muffins when he sells for 50

iii Calculate the coefficient of determination R2

## PART-B

**Answer the following question** **1x10=10**

7. Find the covariance and correlation between the Stock Prices of Company A and Company B over a 6-month period:

| Month | Company A stock price | Company B stock Price |
|---|---|---|
| Jan | 320 | 340 |
| Feb | 350 | 360 |
| Mar | 370 | 380 |
| Apr | 390 | 400 |
| May | 410 | 420 |
| Jun | 430 | 440 |