H/E/TY

Reg. No:

VIT
Vellore Institute of Technology

Final Assessment Test – April 2025

Course: **MDI3006 - Advanced Data Analytics**

Class NBR(s): **2876/2880** Slot: **G1+TG1**

Time: **Three Hours** Max. Marks: **100**

- **KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

Answer ALL Questions

(10 X 10 = 100 Marks)

1. A company is working on three different machine learning tasks, and you are asked to select and explain the most appropriate kernel functions with formula and plot for each task. The tasks are as follows:

   Task 1: The company wants to classify emails as spam or not spam. The data is text-based, represented in a high-dimensional space using word frequency features. Which kernel would you recommend and why?

   Task 2: The company is designing a model to classify different types of fruits based on their color and shape. The relationship between the features is likely to be non-linear. Which kernel would be most suitable, and how would it work in this scenario?

   Task 3: The company has collected sensor data to monitor machine performance over time. This time-series data has a complex relationship, requiring a kernel that captures temporal patterns effectively. Which kernel would be the best choice, and how would it help?

2. A manufacturing company wants to build a predictive model to determine if a machine is likely to fail (1 = Failure, 0 = No Failure) based on sensor readings. The data consists of two different types of features X1 and X2:

   The company decides to use a Multiple Kernel Learning (MKL) approach to handle the heterogeneity of the features effectively. The kernels used are:

   Gaussian (RBF) kernel for $X1$ (Temperature Sensor Readings), to handle non-linear relationships in continuous data.

   Categorical kernel for X2 (Operational Conditions), to measure similarity between categorical features.

   Explain how the multiple kernels capture different feature types and enable the model to predict machine failures effectively with the help of relevant kernel functions individually and also the combined one.

3. Use the following data points in a 2D space and find the decision boundary that separates the two classes (−1 and +1).

Point	Coordinates (x1,x2)	Label (y)
Point 1	(1,1)	-1
Point 2	(2,2)	-1
Point 3	(4,4)	1
Point 4	(5,5)	1

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$y\left(w^T x + b\right) = 1$$

For simplicity, Let you assume Points 2 and 3 are the support vectors. Use the Lagrangian multiplier method to compute w (weight vector) and b (bias) for a support vector machine (SVM) whereas the alpha value is 0.5 for support vectors and 0 for others. Classify a new point, $(x1, x2) = (3, 3)$ using the decision boundary equation.

4. A retail company is designing a machine learning model to predict whether a customer will purchase a product (1 = Purchase, 0 = No Purchase) based on their browsing behaviour. The dataset has two features:
Time Spent on Website ($X1$): Continuous data representing the number of minutes spent on the site.
Number of Pages Visited ($X2$): An integer indicating how many pages the customer explored.
The company wants to use Support Vector Machines (SVMs) to classify customers. However, they are unsure whether to use a hard margin or a soft margin and how the choice would affect the model. Illustrate how the concepts of hard margin and soft margin SVMs work in these real-world scenarios with your plots and decision boundary stumps.

5. A healthcare organization is building a model to predict whether a patient has a disease (1 = Disease, 0 = No Disease) using the AdaBoost algorithm. The dataset consists of the following:

X1 (Age)	X2 (Cholesterol Level)	Y(Has Disease)
45	200	0
50	240	1
40	210	0
35	190	0
60	260	1
55	230	1

Initialize sample weights equally

In Iteration 1, the weak learner uses the decision rule

$X2 \leq 220$, which classifies: $Y = 1$ if $X2 \leq 220$. $Y = 0$ otherwise.

In Iteration 2, assume the next weak learner uses the decision rule

$X1 \geq 50$, which classifies: $Y = 1$ if $X1 \geq 50$. $Y = 0$ otherwise.

Compute the weighted error rate, learner weights and update the weights accordingly.

6. A company is monitoring a network of computers to predict the likelihood of a cyberattack. They have built a Bayesian Belief Network based on three interconnected factors:

Firewall Status (F): Whether the firewall is active (1 = Active, 0 = Inactive).

Suspicious Traffic Detected (T): Whether suspicious traffic has been detected on the network (1 = Detected, 0 = Not Detected).

Cyberattack Occurred (A): Whether a cyberattack actually occurred (1 = Yes, 0 = No).

The relationships between these variables are modeled as follows:

The firewall being active reduces the likelihood of suspicious traffic.

The presence of suspicious traffic increases the likelihood of a cyberattack.

Visualize the structure of the Bayesian Belief Network with conceptual representation.

Highlight the relevant probabilities with estimated (your own assumed data) conditional probability table.

Using the Bayesian Belief Network, calculate the probability of a cyberattack if the firewall is active and suspicious traffic is detected.

7. A research team is using a linear model as a dictionary learning model to reconstruct signals. A signal $Y \in R^{2\times1}$ is represented as a linear combination of dictionary atoms $D \in R^{2\times2}$ and sparse code coefficients $X \in R^{2\times1}$ .Estimate the sparse coefficients X that minimize the reconstruction error, subject to $L1$ and $L2$ regularization.

Signal: $Y = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ Dictionary: $D = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}_{2\times2}$

1×2

Coefficients (initial guess): $X = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Calculate the reconstruction error

Calculate the value of the objective function with L1 and L2 Loss.

8. Discuss the importance of integrating distributed systems with core components of a Big Data Platform, and how do they address the challenges posed by the 5Vs of Big Data?

9.a) Demonstrates how the Flajolet-Martin algorithm can effectively approximate the number of distinct elements in a large data stream with limited memory.
A streaming platform wants to calculate the approximate number of distinct users who watch videos on their platform daily.
The following stream of user IDs is observed over time:
Stream(X) =[101,202,303,101,404,505,202,606,303,707,101] Where X is the stream of user id's.
Use the defined hash function $h(x)$ = Binary Representation of (x mod 32).

OR

9.b) A real-time data monitoring system tracks the presence of a signal (1) or its absence (0) over time. Due to memory constraints, the system needs to estimate the number of 1's in the most recent stream of data using the DGIM (Datar-Gionis-Indyk-Motwani) algorithm.
Given the following stream: S=[1,0,1,1,0,1,0,1,1]
Process the stream step-by-step, constructing buckets as per the DGIM rules.
Clearly indicate how the buckets are modified when each new bit (1 or 0) enters-the stream.

10.a) Illustrate the flow of query execution in Hive, starting from the user submitting a query to the final output being generated. Highlight the interactions between the core components of the HIVE architecture with an example query.

OR

10.b) Imagine you're managing a large, distributed online ticket booking system. Your system has multiple servers handling user requests, processing payments, and updating seat availability. To ensure consistency, all servers must have the same information about which seats are booked and which are available.
Apply an idea of ZooKeeper as a critical coordination and synchronization service for managing distributed systems with the help of suitable diagram.

⇔⇔⇔ **H/E/TY** ⇔⇔⇔