

# MACHINE LEARNING (ECOMP722/IT722) - COMPREHENSIVE QUESTION BANK

## UNIT 1: Introduction & Getting to Know Data

### Machine Learning Basics & Fundamentals

**Q1.** What is Machine Learning? Explain with the help of examples its applications in regression problems. **(January 2023)** [8 marks]

**Q2.** Compare and contrast supervised and unsupervised learning techniques with suitable examples. **(January 2023 OE)** [7 marks]

**Q3.** Consider a classification problem to decide if credit card is to be issued to a person based on whether he has a job. Identify the (i) Experience (ii) Performance (iii) Task with respect to Machine Learning terminology by providing suitable justification. **(January 2023)** [4 marks]

### Overfitting and Underfitting

**Q4.** Define overfitting and underfitting. Provide examples to illustrate their effects on model performance. **(November 2024)** [6 marks]

**Q5.** Explain the Bias-Variance trade-off in machine learning. How does model complexity impact this trade-off? **(November 2021)** [6 marks]

**Q6.** What is the Bias-Variance tradeoff in machine learning and how does it affect model performance? **(November 2024)** [5 marks]

### Data Attributes and Types

**Q7.** Explain the difference between nominal, ordinal, and continuous attributes. Provide examples for each. **(November 2024)** [6 marks]

### Data Preprocessing & Statistical Measures

**Q8.** [PROBLEM QUESTION - Refer to November 2024 Paper] Suppose the following height values (in cm) are given for analysis (in ascending order): 150, 152, 155, 157, 160, 160, 162, 164, 165, 165, 168, 170, 172, 175, 175, 178, 180, 182, 185, 188, 190, 193, 195, 200, 205

- a) Calculate the mean and median of the height values. How do these measures of central tendency differ in this dataset?
- b) Determine the mode(s) of the dataset. Would you classify the dataset as unimodal, bimodal, or multimodal based on the mode(s)?
- c) Estimate the first quartile (Q1) and third quartile (Q3). How do these quartiles help in understanding the distribution of the data?
- d) Provide the five-number summary (minimum, Q1, median, Q3, maximum) for the dataset.

- e) Create a boxplot for the given height values and describe any outliers.
- f) Calculate the midrange of the dataset. What does this value tell you about the central tendency of the dataset? **(November 2024)** [8 marks]

**Q9.** [PROBLEM QUESTION - Refer to November 2021 Paper] A dataset contains several missing values on different features. Describe a technique for handling missing values and demonstrate how you would apply one of them to the following dataset with Feature A, Feature B, Feature C with missing values. **(November 2021)** [8 marks]

**Q10.** Consider the two objects represented by attribute values (1,6,2,5,3) and (3,5,2,6,7):

- 1. Compute the Euclidean distance between two objects
- 2. Compute the Manhattan distance between two objects
- 3. Compute the Minkowski distance between two objects, using  $q=3$  **(January 2023)** [6 marks]

**Q11.** [PROBLEM QUESTION - Refer to November 2021 Paper] Consider the following data points representing two objects. Compute the Manhattan distance and Euclidean distance between the two objects. **(November 2021)** [6 marks]

**Q12.** Explain the following measures:

- 1. Euclidean distance
- 2. Manhattan distance
- 3. Minkowski distance
- 4. Cosine Similarity **(December 2023)** [6 marks]

**Q13.** What is data normalization and why is it important in machine learning? Briefly describe two normalization techniques. **(November 2021)** [8 marks]

---

## UNIT 2: Basic Classification Methods & Model Evaluation

### Decision Tree Induction

**Q14.** [PROBLEM QUESTION - Refer to January 2023 Paper] Construct the ID3 decision tree for the following training data: Balloon dataset with Color, Size, Act, Age attributes. **(January 2023)** [10 marks]

**Q15.** [PROBLEM QUESTION - Refer to May 2024 Paper] What is Machine Learning? Construct the ID3 decision tree for the Sunlight attribute given: Height, Weight, Location, Sunlight data. **(May 2024)** [8 marks]

**Q16.** [PROBLEM QUESTION - Refer to December 2023 Paper] The following dataset is supposed to be a part of a transportation study regarding mode choice to select Bus, Car or Train in a city. Construct a decision tree with ID3 for the target attribute "TRANSPORTATION MODE": Gender, Car ownership, Travel cost, Income, Transportation mode. **(December 2023)** [10 marks]

**Q17.** [PROBLEM QUESTION - Refer to May 2024 Paper] Transportation study dataset - similar to Q16. **(May 2024)** [10 marks]

**Q18.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] Consider the training examples for a binary classification problem with Price, Maintenance, Capacity, Airbag, Profitable attributes:

- a. Considering "profitable" as the binary valued attribute to be predicted, which attribute would be selected as the root node in a decision tree with multi-way splits using the information gain measure?
- b. For the same data set, suppose we decide to construct a decision tree using binary split and the Gini index impurity measure, which feature and split point combination would be the best to use as the root node? **(January 2023 OE)** [10 marks]

**Q19.** [PROBLEM QUESTION - Refer to November 2021 Paper] Apply the ID3 algorithm to write the rulebook for the given dataset: Humidity, Family, Play dataset. **(November 2021)** [8 marks]

### Naive Bayesian Classification

**Q20.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] Using Naive Bayesian Classification for the following training data, predict whether the unknown sample with [Age:7, Vaccination:1, Tumor Size: Medium, Tumor Site: Shoulder] belongs to malignant tumor or not. Dataset includes: Age, Vaccination, Tumor Size, Tumor Site, Malignant. **(January 2023 OE)** [6 marks]

**Q21.** What are Bayesian Belief Networks, and how are they used in classification? **(December 2023)** [6 marks]

**Q22.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] A Bayesian belief network for detecting heart disease and heartburn in patients with conditional probability tables:

- a. Determine the likelihood of a person to have heart diseases.
- b. If the person has high blood pressure, calculate the probability that they have heart disease.
- c. Given that the person having high blood pressure exercises regularly and eats a healthy diet, what is the posterior probability that the person has a heart disease? **(January 2023 OE)** [8 marks]

**Q23.** Explain how Naive Bayes Classification works. Illustrate your answer with an example. **(November 2021)** [5 marks]

### Rule-Based Classification

**Q24.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] Suppose a rule-based classifier produces the following rule set for Car Value with Mileage, Air Conditioner, Engine:

- R1: (Mileage: High)→(Car Value: Low)

- R2: (Mileage: Low)→(Car Value: High)
- R3: (Air Conditioner: Working)^(Engine: Good)→Car Value: High
- R4: (Air Conditioner: Working)^(Engine: Bad)→Car Value: Low
- R5: (Air Conditioner: Broken)→Car Value: Low
  - a. Are the rules mutually exclusive?
  - b. Is the rule set exhaustive?
  - c. Is ordering needed for this set of rules?
  - d. Do you need a default class for the rule set?
  - e. Find the coverage and accuracy for each of the rules. **(January 2023 OE)** [7 marks]

**Q25.** Explain the different types of associative classification used in machine learning. **(January 2023, May 2024)** [6 marks]

### Support Vector Machines (SVM)

**Q26.** Describe the working principle of the K-Nearest Neighbor (KNN) algorithm. Explain the working principle of the Support Vector Machine (SVM) algorithm. **(December 2023)** [7 marks each]

**Q27.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] With the help of relevant sketches and mathematical equations, explain how Support Vector Machines determine the maximum marginal hyperplane in case of linearly separable data. Consider positively labeled data points:  $\{[3,1], [3,-1], [6,1], [6,-1]\}$  and negatively labeled data points:  $\{[1,0], [0,1], [0,-1], [-1,0]\}$

- a. Verify if the data points are linearly separable.
- b. Using SVM algorithm create a line or a hyperplane which separates the data into classes. **(January 2023 OE)** [10 marks]

**Q28.** [PROBLEM QUESTION - Refer to May 2024 Paper] Consider the following positively labeled data points:  $\{(3,1),(3,-1),(6,1),(6,-1)\}$  and negatively labeled data points:  $\{(1,0),(0,1),(0,-1),(-1,0)\}$ . Using SVM create a hyperplane that converts data into classes. **(May 2024)** [10 marks]

**Q29.** [PROBLEM QUESTION - Refer to November 2024 Paper] Construct a Support Vector Machine (SVM) classifier with positively labeled data points:  $\{(3,1),(3,-1),(6,1),(6,-1)\}$  and negatively labeled data points:  $\{(1,0),(0,1),(0,-1),(-1,0)\}$ . Find the equation of separating hyperplane, identify support vectors, and verify classification. **(November 2024)** [8 marks]

**Q30.** Explain the Support Vector Machine (SVM) algorithm. **(January 2023)** [8 marks]

### K-Nearest Neighbor (KNN)

**Q31.** [PROBLEM QUESTION - Refer to November 2024 Paper] Myntra's customer dataset includes height, weight, and T-shirt sizes ('M' or 'L'). Predict T-shirt size for Anurag (Height: 161 cm, Weight: 61 kg) using KNN with  $k=3$ . Dataset provided with Height(cm), Weight(kg), T-Shirt Size for 10 customers. **(November 2024)** [8 marks]

**Q32.** Write an algorithm for k-nearest neighbor classification given k, the nearest number of neighbors, and n, the number of attributes describing each tuple. In k-nearest neighbors' classifiers, how is the distance computed:

- a. For attributes that are not numeric, but categorical?
- b. In case of missing values? **(January 2023 OE)** [6 marks]

**Q33.** [PROBLEM QUESTION - Refer to November 2021 Paper] You are tasked with classifying using k-NN. Given distances between test point and 5 nearest neighbours with their class labels. If  $k=3$ , classify the test point based on majority voting principle. **(November 2021)** [7 marks]

### Model Evaluation Metrics

**Q34.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] Environmental scientists solve two-class classification for genetic variant prediction. 500 samples, predicted 350 with variant, 150 without. Actual: 305 with, 195 without. TP=200. Complete confusion matrix and calculate accuracy, misclassification rate, sensitivity, specificity, precision, recall and F1-score. **(January 2023 OE)** [6 marks]

**Q35.** [PROBLEM QUESTION - Refer to November 2024 Paper] International Airlines decided 2 classifiers (A and B) to predict if flight from Pakistan arrives on time. Tested on 500 flights. Confusion matrix provided for both classifiers.

- (i) Which is preferable in terms of F1 score?
- (ii) Which is preferable in terms of accuracy? **(November 2024)** [7 marks]

**Q36.** [PROBLEM QUESTION - Refer to December 2023 Paper] Calculate accuracy of Practo Health Tech's patient diagnosis model using confusion matrix. Calculate precision and recall for Class C. Explain implications of false positive and false negative values. Interpret true positive value. **(December 2023)** [4+4 marks]

**Q37.** Suppose I have 10,000 emails in my mailbox out of which 200 are spams. The spam detection system detects 150 mails as spams, out of which 50 are actually spams. What is the precision and recall of my spam detection system? **(May 2024)** [4 marks]

**Q38.** Explain the key metrics used to evaluate classifier performance. Discuss precision, recall, F1-score, and accuracy with examples. **(November 2021)** [6 marks]

### Cross-Validation

**Q39.** What is k-fold cross-validation? Explain how it works and why it is preferred over the holdout method in certain situations. **(November 2024)** [7 marks]

**Q40.** Differentiate between holdout and cross validation. Why is cross-validation preferred over hold-out? How does stratified cross validation work? **(January 2023 OE)** [4 marks]

**Q41.** [PROBLEM QUESTION] A dataset is split into training and test sets using 10-fold cross-validation. Accuracy scores for each fold: 85%, 78%, 90%, 82%, 88%, 80%, 92%, 85%, 84%, 83%. Calculate mean accuracy and standard deviation. **(November 2021)** [7 marks]

**Q42.** Explain the following:

- i. Cross validation
- ii. Bootstrap (**May 2024**) [4 marks]

**Q43.** Describe the Holdout method and Cross-Validation technique for evaluating machine learning models. How do they differ in terms of performance estimation? (**November 2021**) [5 marks]

### **Ensemble Methods**

**Q44.** Explain the concept of bagging and boosting? State with suitable justification why it may improve the accuracy of decision tree induction. (**December 2023, January 2023**) [8 marks each]

**Q45.** What makes Random Forests robust? Explain how the Random Forest algorithm works. (**November 2021**) [5 marks]

**Q46.** Explain in brief the techniques used to improve classification accuracy. (**May 2024, November 2024**) [4-7 marks]

### **Lazy vs Eager Learners**

**Q47.** Explain the difference between Eager Learner and Lazy Learner classification algorithms. Give examples of each. (**November 2024**) [5 marks]

**Q48.** Distinguish between Lazy learners & Eager learners? (**January 2023, May 2024, November 2021**) [4 marks each]

### **ROC Curves**

**Q49.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] Data tuples sorted by decreasing probability values returned by classifier. For each tuple, compute TP, FP, TN, FN. Compute TPR and FPR for ROC curve. Table with 10 tuples, class (P/N), and probability values provided. (**January 2023 OE**) [6 marks]

**Q50.** Consider a binary classification problem where performance is evaluated using ROC curve. Draw and explain ROC curve, and illustrate how AUC can be interpreted in terms of classifier performance. (**November 2021**) [5 marks]

### **Logistic Regression**

**Q51.** Write a note on logistic regression. (**May 2024**) [4 marks]

**Q52.** Distinguish between logistic regression and linear regression? (**January 2023**) [4 marks]

**Q53.** Explain how logistic regression can be used for binary classification. What is the role of the logistic function in this context? (**November 2024**) [7 marks]

### **Other Classification Topics**

**Q54.** Outline methods for addressing the class imbalance problem. Suppose a bank wants to develop a classifier that guards against fraudulent credit card transactions. Illustrate how you can induce a quality classifier based

on large set of non-fraudulent examples and very small set of fraudulent cases. **(January 2023 OE)** [6 marks]

**Q55.** Define (i) Prior Probability (ii) Conditional Probability (iii) Posterior Probability **(May 2024)** [3 marks]

---

### UNIT 3: Cluster Analysis

#### Clustering Basics

**Q56.** What is clustering. Discuss the major requirements of Clustering Analysis. **(May 2024)** [6 marks]

**Q57.** Explain what are clusters and write requirements for cluster analysis. **(November 2021)** [7 marks]

**Q58.** Compare the properties of a good clustering algorithm. **(November 2021)** [7 marks]

**Q59.** Explain the basic requirements for performing cluster analysis. **(November 2024)** [6 marks]

#### Partitioning Methods - K-Means

**Q60.** [PROBLEM QUESTION - Refer to January 2023 Paper] Consider clustering points into 2 clusters. Distance function is Euclidean. Points A1-A8 with x,y coordinates. Initial cluster centers A1, A4, A7. Use K-Means algorithm. **(January 2023)** [12 marks]

**Q61.** [PROBLEM QUESTION - Refer to May 2024 Paper] Use K-means clustering to divide data into two clusters. Euclidean distance. Points A1-A8 with x,y coordinates provided. **(May 2024)** [11 marks]

**Q62.** [PROBLEM QUESTION - Refer to November 2024 Paper] Cluster points (x,y) into three clusters: A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9). Euclidean distance. Initially assign A1, B1, C1 as centers. Use k-means to show:

- (i) Three cluster centers after first round
- (ii) Final three clusters **(November 2024)** [10 marks]

**Q63.** [PROBLEM QUESTION - Refer to November 2024 Paper] Given data points: x1(4,5), x2(2,2), x3(4,5), x4(4,7), x5(3,1), x6(9,2), x7(6,2), x8(7,4), x9(8,4), x10(7,3). Perform:

- (i) Apply CURE algorithm with Euclidean distance, threshold  $\epsilon=1.5$ , Maximum 2
- (ii) Create C# tree
- (iii) Explain structure of C# tree after clustering **(November 2024)** [8 marks]

**Q64.** [PROBLEM QUESTION] Cluster the following data points using K-Means algorithm with  $k=3$ . **(November 2024)** [8 marks]

**Q65.** [PROBLEM QUESTION - Refer to November 2024 Paper] Consider points P1, P2, P3, P4, P5. Use distance matrix given below to perform hierarchical clustering using single link. Show results by drawing dendrogram. The dendrogram should clearly show the order in which points are merged. Distance matrix provided. **(November 2024)** [7 marks]

## K-Medoids

**Q66.** [PROBLEM QUESTION - Refer to December 2023 Paper] Apply K-Medoid algorithm and cluster dataset for  $k=2$ . X-coordinate and Y-coordinate data table provided. **(December 2023)** [12 marks]

**Q67.** [PROBLEM QUESTION] Consider dataset clustered using k-medoids: P1(1,1), P2(3,3), P3(10,10), P4(11,11). Calculate medoids after one iteration assuming P1 and P3 are initial medoids. **(November 2021)** [5 marks]

## Hierarchical Clustering - AGNES

**Q68.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] Apply Agglomerative Nesting (AGNES) algorithm. Draw dendrogram using single link and complete link approach. Distance matrix for items A,B,C,D,E provided. **(January 2023 OE)** [6 marks]

**Q69.** [PROBLEM QUESTION - Refer to May 2024 Paper] Use distance matrix to perform hierarchical clustering using single link and divisive hierarchical clustering technique. Distance matrix provided. **(May 2024)** [8 marks]

**Q70.** Explain the concept of hierarchical clustering and discuss the various types of hierarchical clustering schemes. **(January 2023)** [8 marks]

**Q71.** Discuss why there are more agglomerative methods than divisive methods? **(January 2023 OE)** [8 marks]

**Q72.** Differentiate between partitioning and hierarchical clustering methods. **(December 2023)** [8 marks]

**Q73.** Write short notes on the following Hierarchical Clustering Algorithms:

- 1. Chameleon Clustering
- 2. Probabilistic Clustering **(January 2023 OE)** [8 marks]

**Q74.** Explain Chameleon - Hierarchical based clustering algorithm. **(January 2023, May 2024)** [8 marks]

## Density-Based Clustering - DBSCAN

**Q75.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] Discuss need for density-based clustering. Given 12 two-dimensional dataset points with x,y coordinates, apply DBSCAN ( $\epsilon=1.9$ , minimum points=4) to decide core, border, outlier points. **(January 2023 OE)** [12 marks]

**Q76.** [PROBLEM QUESTION - Refer to December 2023 Paper] Given 8 data points with (X,Y) coordinates. Apply DBSCAN (Epsilon=0.5, MinPts=3):

- i. Identify and label clusters
- ii. Provide description of clusters **(December 2023)** [12 marks]

**Q77.** [PROBLEM QUESTION - Refer to November 2021 Paper] Given data points, apply DBSCAN (Eps=3, minPts=4). Identify Core points, border points, noise points. Data points P1-P5 with coordinates provided.



**(November 2021)** [17 marks]

**Q78.** Explain the following terms with reference to DBSCAN:

- (i) Core points
- (ii) Noise points
- (iii) Border points **(November 2024)** [6 marks]

**Q79.** Differentiate between DBSCAN & OPTICS density based clustering algorithm. **(January 2023, January 2023 OE)** [4 marks each]

### **Grid-Based Clustering**

**Q80.** Elaborate the following grid-based clustering methods:

- i. STING: Statistical Information Grid
- ii. CLIQUE: Clustering in Quest **(January 2023 OE)** [8 marks]

**Q81.** Explain any one grid based clustering Technique. **(May 2024)** [4 marks]

**Q82.** Explain the following clustering algorithm: (i) Chameleon 2. STING **(December 2023)** [6 marks]

### **Other Clustering Methods**

**Q83.** What are fuzzy clusters? Discuss situations in which fuzzy clustering techniques would be preferred over traditional clustering methods. **(January 2023 OE)** [6 marks]

**Q84.** Describe how the Expectation-Maximization (EM) algorithm works for probabilistic clustering. Explain the main steps involved. **(November 2021)** [6 marks]

**Q85.** While assessing the feasibility of clustering analysis, how is the number of clusters in a data set determined? **(January 2023 OE)** [6 marks]

### **High-Dimensional Data Clustering**

**Q86.** Explain the necessity of clustering high-dimensional data? **(January 2023, December 2023)** [4-6 marks]

**Q87.** With respect to clustering of high-dimensional data:

- 1. Why are traditional distance measures, used in low-dimensional cluster analysis, not effective on high-dimensional data?
  - 2. Discuss kinds of clusters that are meaningful on high-dimensional data. **(January 2023 OE)** [4 marks each]
-

## UNIT 4: Regression, Outlier Analysis & Dimensionality Reduction

### Linear Regression

**Q88.** Differentiate between:

- i. Simple and Multiple Regression
- ii. Linear and Non-Linear Regression. With help of mathematical expressions explain simple linear regression model. **(January 2023 OE)** [7+6 marks]

**Q89.** Explain the concept multiple linear regression for prediction. **(December 2023)** [6 marks]

**Q90.** Explain the difference between simple and multiple linear regression. Provide examples of scenarios where each would be applied. **(November 2021)** [7 marks]

**Q91.** [PROBLEM QUESTION - Refer to November 2021 Paper] Given dataset with three variables V1, V2, Y. Perform multiple linear regression to predict Y from X1 and X2, provide regression coefficients. **(November 2021)** [7 marks]

### Logistic Regression

**Q92.** Distinguish between logistic regression and linear regression? **(January 2023)** [4 marks]

**Q93.** Explain how logistic regression can be used for binary classification. What is the role of the logistic function in this context? **(November 2024)** [7 marks]

### Outlier Detection

**Q94.** What are outliers? Can noise objects be Outliers? With an example explain the collective outliers. **(January 2023 OE)** [4 marks]

**Q95.** What is an outlier? Briefly describe distance based outlier detection method? **(January 2023, May 2024)** [4 marks each]

**Q96.** What are the key challenges involved in outlier detection? Discuss how these challenges affect the performance of machine learning models. **(November 2024)** [7 marks]

**Q97.** Discuss the challenges of Outlier detection. **(January 2023 OE)** [4 marks]

**Q98.** Explain with the help of example the different types of outliers. **(January 2023)** [8 marks]

**Q99.** [PROBLEM QUESTION] Given dataset representing ages: [15, 24, 25, 33, 26, 31, 40, 51]

- (i) Use Z-score method to detect outlier. Assume threshold=2.
- (ii) Calculate Z-scores and identify outliers. **(November 2021)** [7 marks]

**Q100.** Define and discriminate between point outliers, contextual outliers, and collective outliers. Provide examples for each. **(November 2021)** [5 marks]

## Dimensionality Reduction - PCA

**Q101.** Explain the concept of Principal Component Analysis (PCA) and its use in dimensionality reduction. How does PCA transform the original variables? **(November 2024)** [6 marks]

**Q102.** Explain the procedure used in PCA to enable dimension reduction. **(January 2023)** [8 marks]

**Q103.** Using Principal Components Analysis (PCA), describe how dimensionality reduction is performed. What are main advantages of using PCA in high-dimensional data? **(November 2021)** [6 marks]

**Q104.** [PROBLEM QUESTION - Refer to November 2021 Paper] Given data matrix  $X$  with  $A_1, A_2$  values:

- (i) Perform PCA and calculate first principal component
- (ii) Explain how PCA helps in reducing dimensionality **(November 2021)** [8 marks]

**Q105.** Discuss the curse of dimensionality, model overfitting, and how reducing dimensions can improve model performance and interpretability. **(November 2024)** [4 marks]

## Linear Discriminant Analysis (LDA)

**Q106.** [PROBLEM QUESTION - Refer to January 2023 OE Paper] Discuss Curse of Dimensionality. How LDA helps in dimensionality reduction? Apply LDA to project feature into smaller subspace:  $X_1: [(4,1), (2,4), (2,3), (3,6), (4,4)]$ ,  $X_2: [(9,10), (6,8), (9,5), (8,7), (10,8)]$  **(January 2023 OE)** [10 marks]

**Q107.** [PROBLEM QUESTION - Refer to November 2024 Paper] Compute linear Discriminant projection for two-dimensional dataset:

- Class  $w_1: X_1 = \{(4,1), (2,4), (2,3), (3,6), (4,4)\}$
- Class  $w_2: X_2 = \{(9,10), (6,8), (9,5), (8,7), (10,8)\}$ 
  - (i) Calculate mean vectors  $m_1$  and  $m_2$
  - (ii) Compute within-class scatter matrix  $SW$
  - (iii) Find optimal linear discriminant direction  $w$
  - (iv) Project samples and interpret results **(November 2024)** [10 marks]

**Q108.** Explain how linear discriminant analysis (LDA) helps in reducing dimensionality of a data set. **(January 2023, December 2023)** [6 marks each]

## Similarity Measures

**Q109.** Explain how Geodesic Distance is used to measure similarity between two vertices in a graph. Does it suit to measure similarity between people in social network? **(January 2023 OE)** [6 marks]

---

## UNIT 5: Neural Networks (Additional/Advanced Topics)

### Artificial Neural Networks

**Q110.** [PROBLEM QUESTION - Refer to December 2023 Paper] Consider feedforward neural network with Input layer, Hidden layer, Output layer with initial weights and bias values. Target output=0.7, learning rate  $\eta=0.1$

- Perform one forward pass for input ( $x_1=0.6$ ,  $x_2=0.9$ ), calculate predicted output with calculations for each neuron
- Calculate error (E) using MSE (**December 2023**) [12 marks]

**Q111.** [PROBLEM QUESTION - Refer to May 2024 Paper] Consider feedforward neural network with weights provided. Network has good activation function. Perform forward pass, calculate error & backpropagate using error backpropagation algorithm. Actual output  $y=0.8$ , learning rate  $\eta=1$ . (**May 2024**) [10 marks]

**Q112.** Explain the importance of activation function and explain the sigmoid activation functions. (**January 2023, December 2023**) [8 marks]

---

## QUESTIONS REQUIRING REFERENCE TO ORIGINAL PAPERS

### (Questions with Detailed Data Tables/Diagrams/Matrices)

1. **Q8** - Height values complete statistical analysis with 27 data points
2. **Q9** - Missing values handling with Feature table
3. **Q11** - Distance calculation between two multi-dimensional objects
4. **Q14** - Balloon dataset with Color, Size, Act, Age attributes
5. **Q15** - Sunlight prediction with Height, Weight, Location attributes
6. **Q16 & Q17** - Transportation mode dataset with Gender, Car ownership, Travel cost, Income
7. **Q18** - Car profitability with Price, Maintenance, Capacity, Airbag - full data table
8. **Q19** - Humidity, Family, Play dataset for ID3
9. **Q20** - Tumor malignancy dataset with Age, Vaccination, Tumor Size/Site
10. **Q22** - Bayesian network diagram with conditional probability tables
11. **Q24** - Car value rules with complete data table showing records
12. **Q27, Q28, Q29** - SVM problems with 8 data points each