

Introduction to Machine Learning

– Prof. Balaraman Ravindran | IIT Madras

Problem Solving Session (Week-3)

Shreya Bansal

PMRF PhD Scholar
IIT Ropar

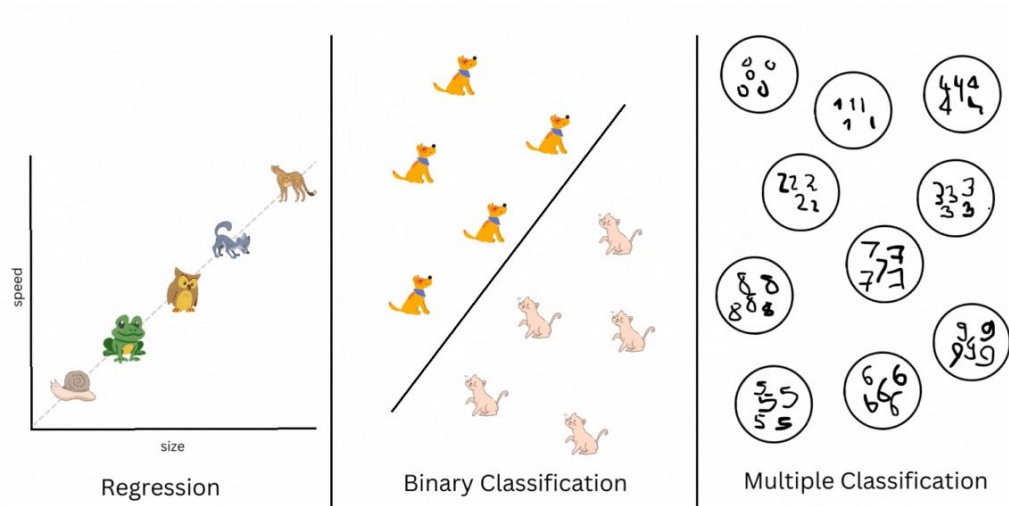
Week-3 Contents

— — —

1. Linear Classification
2. Logistic Regression
3. Linear Discriminant Analysis
4. Tutorial on Weka

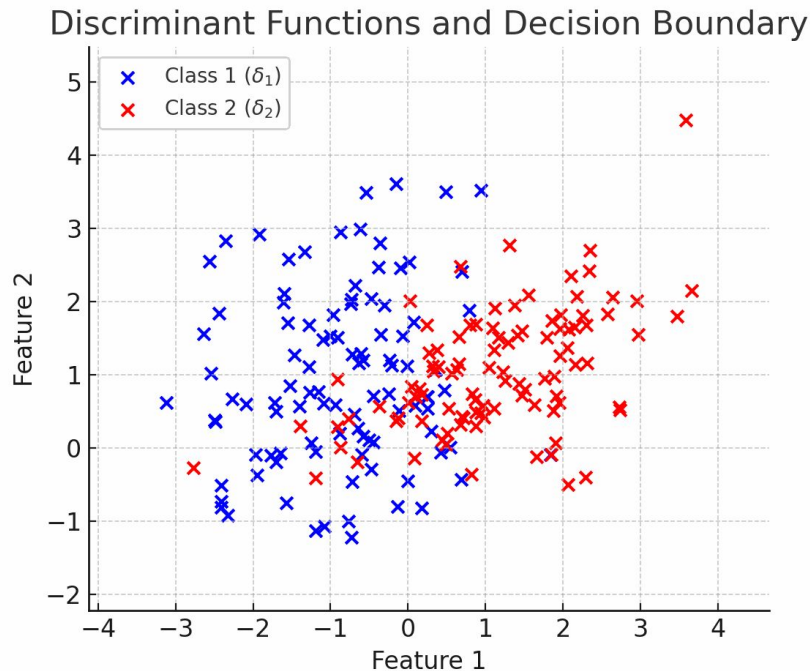
Linear Classification

- Linear classification separates data using a hyperplane.
- Two main approaches:
 - Modeling a discriminant function
 - Directly modeling the hyperplane



Discriminant Function

- A function is assigned to each class
- The class with the highest function value determines classification
- Goal: Learn the discriminant functions δ_1
- Example: Two-class problem with δ_1 and δ_2
- The boundary is where $\delta_1 = \delta_2$



Conditions for a Linear Separating Surface

— — —

- Models the relationship between dependent and independent variables by fitting a linear equation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

where:

- \mathbf{Y} is the output (dependent variable)
- \mathbf{X} is the input (independent variables)
- $\boldsymbol{\beta}$ are the model parameters
- ε is the error term

Approaches for Linear Classification

Discriminant Function-Based Approaches

- Linear regression (indicator variables)
- Logistic regression
- Linear Discriminant Analysis (LDA)

Hyperplane-Based Approaches

- Perceptron
- Finding the optimal hyperplane

Problem Setup

— — —

- **Classes and Input Representation:**

We have K classes: $G=\{1,2,...,K\}$

Input X belongs to R^p , meaning it is a p -dimensional vector.

- **One-of-K Encoding for Classes:**

Each class label is represented as a one-hot vector Y of size K .

Example for $K=3$

Class 1: $Y=[1,0,0]$

Class 2: $Y=[0,1,0]$

Class 3: $Y=[0,0,1]$

Problem Setup

— — —

- Linear Regression Model:

The function $f(X)$ is computed as: $f(X) = X\beta$

where β is a $p \times K$ weight matrix.

- Prediction Rule: The final class label is determined using the argmax function: $\hat{y} = \arg \max_k f_k(X)$
where $f_k(X)$ is the score for class k .

Finding Optimal β

— — —

- To minimize the error function, take its derivative w.r.t. β and set it to zero:

$$\partial/\partial\beta (\mathbf{Y}^T \mathbf{Y} - 2\beta^T \mathbf{X}^T \mathbf{Y} + \beta^T \mathbf{X}^T \mathbf{X} \beta) = 0$$

- Solving for β gives:

$$\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Sample Question: Regression for classification

x1	x2	y	One hot Y
1	2	1	[1,0,0]
2	1	1	[1,0,0]
1	1	1	[1,0,0]
4	2	2	[0,1,0]
3	1	2	[0,1,0]
4	3	2	[0,1,0]
6	5	3	[0,0,1]
7	5	3	[0,0,1]
6	6	3	[0,0,1]

Sample Question: Regression for classification

X (9X3)		
1	1	2
1	2	1
1	1	1
1	4	2
1	3	1
1	4	3
1	6	5
1	7	5
1	6	6

Step-1: Define
X and Y

Y (9X3)		
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

Sample Question: Regression for classification

X (9X3)		
1	1	2
1	2	1
1	1	1
1	4	2
1	3	1
1	4	3
1	6	5
1	7	5
1	6	6

X^T (3 X 9)	1	1	1	1	1	1	1	1	1
	1	2	1	4	3	4	6	7	6
	2	1	1	2	1	3	5	5	6



a (3 x 3)	9	34	27
	34	168	136
	26	129	111



a^{-1} (3 x 3)	0.473	-0.124	0.037
	-0.102	0.127	-0.131
	0.007	-0.118	0.152

Step-2:
Calculate $a^{-1} = (X^T X)^{-1}$

Sample Question: Regression for classification

Y (9X3)		
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

X^T (3 X9)	1	1	1	1	1	1	1	1	1
	1	2	1	4	3	4	6	7	6
	2	1	1	2	1	3	5	5	6



b (3 x 3)	3	3	3
	4	11	19
	4	6	16

Step-3: Calculate $b = X^T Y$

Sample Question: Regression for classification

a⁻¹ (3 x 3)	0.473	-0.124	0.037
	-0.102	0.127	-0.131
	0.007	-0.118	0.152

b (3 x 3)	3	3	3
	4	11	19
	4	6	16

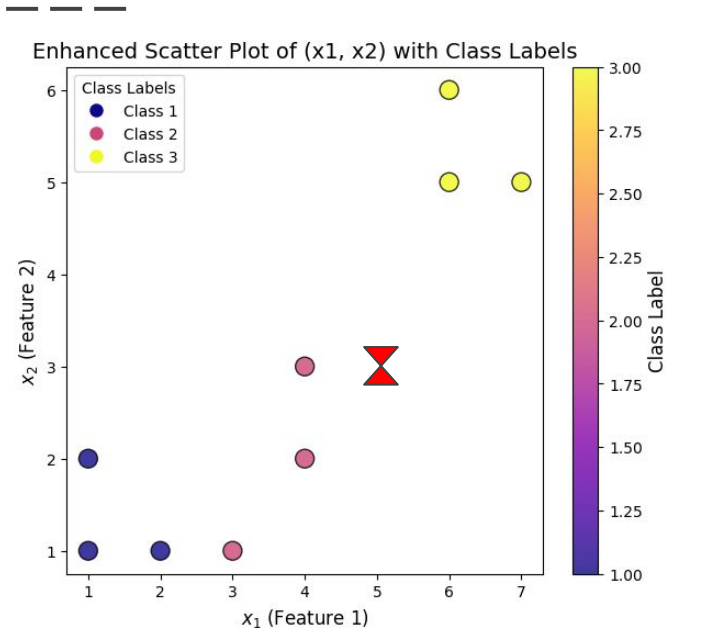


β (3 x 3)	1.071	0.277	-0.345
	-0.322	0.305	0.011
	0.157	-0.365	0.211

$$\begin{aligned}y_1 &= 1.071 - 0.322 x_1 + 0.157 x_2 \\y_2 &= 0.277 + 0.305 x_1 - 0.365 x_2 \\y_3 &= -0.345 + 0.011 x_1 + 0.211 x_2\end{aligned}$$

Step-4: Calculate $\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{a}^{-1} \mathbf{b}$

Sample Question: Regression for classification



X'		
1	5	3

β (3 x 3)	1.071	0.277	-0.345
	-0.322	0.305	0.011
	0.157	-0.365	0.211

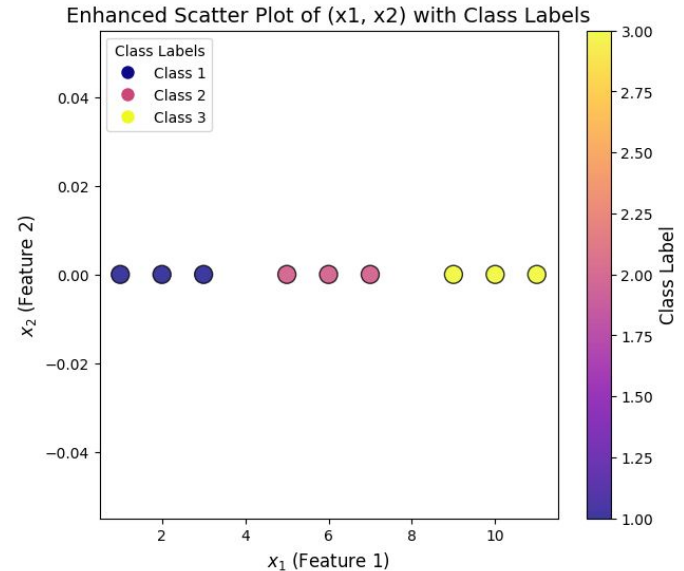
Y'		
-0.068	0.707	0.343
Max= 0.707 (class-2)		

Step-4: Check for point (5,3)

Pitfalls of Linear Regression for Classification

— — —

x1	y	One hot Y
1	1	[1,0,0]
2	1	[1,0,0]
3	1	[1,0,0]
5	2	[0,1,0]
6	2	[0,1,0]
7	2	[0,1,0]
9	3	[0,0,1]
10	3	[0,0,1]
11	3	[0,0,1]



Pitfalls of Linear Regression for Classification

X (9X2)	
1	1
1	2
1	3
1	5
1	6
1	7
1	9
1	10
1	11

Step-1: Define
X and Y

Y (9X3)		
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

Pitfalls of Linear Regression for Classification

X (9X2)	
1	1
1	2
1	3
1	5
1	6
1	7
1	9
1	10
1	11

X^T	1	1	1	1	1	1	1	1	1
(2 X 9)	1	2	3	5	6	7	9	10	11



a	9	54
(2 x 2)	54	426



a^{-1}	0.464	-0.058
(2 x 2)	-0.058	0.009

Step-2:
Calculate $a^{-1} = (X^T X)^{-1}$

Pitfalls of Linear Regression for Classification

Y (9X3)		
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

X^T (2 X9)	1	1	1	1	1	1	1	1	1
	1	2	3	5	6	7	9	10	11



\mathbf{b} (2 x 3)	3	3	3
	6	18	30

Step-3: Calculate $\mathbf{b} = \mathbf{X}^T \mathbf{Y}$

Pitfalls of Linear Regression for Classification

a ⁻¹ (2 x 2)	0.464	-0.058
	-0.058	0.009

β (2 x 3)	1.044	0.348	-0.348
	-0.12	-0.012	0.096

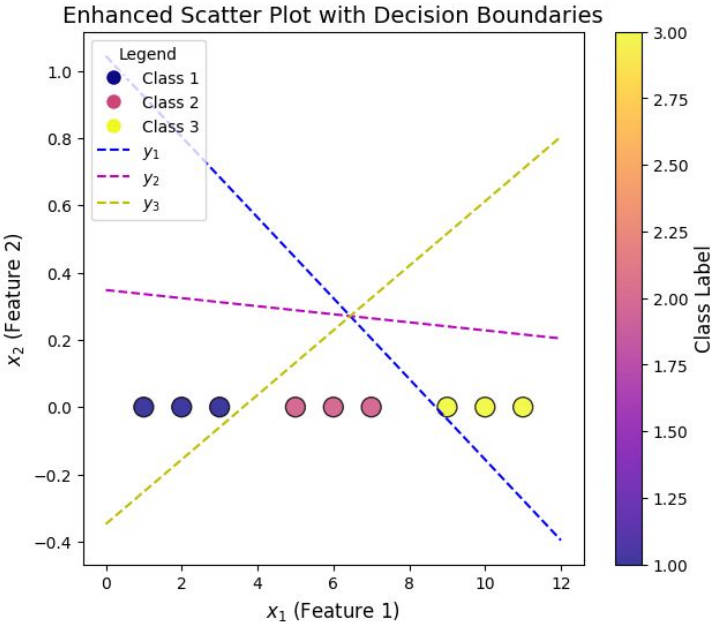
b (2 x 3)	3	3	3
	6	18	30



$$\begin{aligned}y_1 &= 1.044 - 0.12 x_1 \\y_2 &= 0.348 - 0.012 x_1 \\y_3 &= -0.348 + 0.096 x_1\end{aligned}$$


Step-4: Calculate $\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{a}^{-1} \mathbf{b}$

Pitfalls of Linear Regression for Classification



X'	
1	6

β (2 x 3)	1.044	0.348	-0.348
	-0.12	-0.012	0.096

Y'		
1.764	0.348	0.228
Max= 1.764 (class-1) 		

Step-4: Check for point (6)

Logistic Regression

— — —

- Logistic regression is a classification algorithm.
- It models the probability of an outcome.
- Commonly used for binary classification problems.
- Logistic regression transforms a regression problem into a classification problem.



Logistic Regression

— — —

- We aim to model $P(Y=1|X)$, the probability that the output is 1 given X .
- Instead of modeling probability directly, we apply a transformation.
- The logit function transforms probability into log-odds:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta X$$

Logistic Regression

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta X$$

- Probability of success: $P(X)$
- Probability of failure: $1 - P(X)$
- Odds: Ratio of success to failure probabilities.
- Logit function: Log of odds

Sigmoid Function

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta X$$

- **Applying the inverse of the logit function gives: (solve above)**

$$P(X) = \left(\frac{1}{1 + \exp^{-(\beta_0 + \beta X)}} \right)$$

- **This function ensures probabilities remain between 0 and 1.**
- **Decision boundary: $P(X) > 0.5 \rightarrow$ Predict class 1, else predict class 0**

Logistic Regression for Multiple Classes

$$P(y = k \mid X) = \frac{e^{X\beta_k}}{\sum_{j=1}^K e^{X\beta_j}}$$

- Multinomial logistic regression extends logistic regression to multiple classes.
- Uses the softmax function to compute probabilities.
- Requires estimating parameters for K-1 classes.
- One class is chosen as a reference with coefficients set to zero.

Maximum Likelihood Estimation (MLE)

— — —

- Logistic regression uses maximum likelihood estimation (MLE) to estimate parameters.

Given training data , $D = \{(\mathbf{x}_1, g_1), (\mathbf{x}_2, g_2), \dots, (\mathbf{x}_n, g_n)\}$ likelihood is:

$$\mathcal{L}(\beta) = \prod_{i=1}^N (P(X_i)^{g_i} (1 - P(X_i))^{1-g_i})$$

- Log-likelihood is maximized to find optimal .

Log-Likelihood Function

— — —

$$\mathcal{L}(\beta) = \prod_{i=1}^N (P(X_i)^{g_i} (1 - P(X_i))^{1-g_i})$$

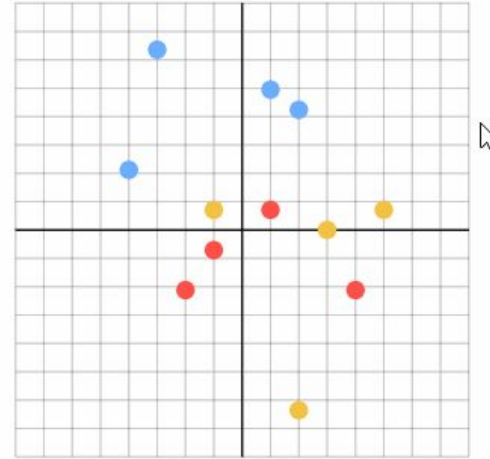
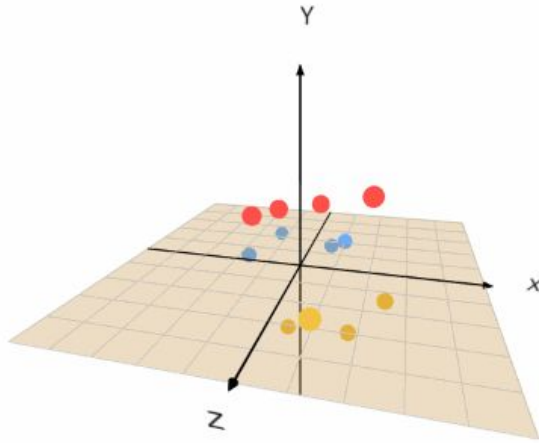
- Taking the logarithm of the likelihood function:

$$\log \mathcal{L}(\beta) = \sum_{i=1}^N [g_i \log P(X_i) + (1 - g_i) \log(1 - P(X_i))]$$

- Optimization involves taking derivatives and solving .
- Unlike linear regression, there is no closed-form solution, so numerical methods (e.g., gradient descent/ Newton Raphson) are used.

Linear Discriminant Analysis (LDA)

— — —



Bayes' Theorem & Class Probabilities

- We aim to compute: $P(Y=k|X)$, Using Bayes' rule:

$$\{P(X|Y=k)P(Y=k)\} / P(X)$$

- Prior Probability: $P(Y=k)$
- Likelihood: $P(X|Y=k)$ (Class conditional density, denoted as $f_k(X)$)
- Marginal Probability: $P(X)$ (Summing over all class probabilities)

Gaussian Assumption in LDA & QDA

— — —

- **LDA Assumption:** follows a single multivariate Gaussian distribution. (covariance =1)
- **QDA (Quadratic Discriminant Analysis):** Also assumes Gaussian densities but with different covariance matrices per class.

$$f_k(X) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (X - \mu_k)^T \Sigma^{-1} (X - \mu_k)\right)$$



Mixture Distributions in Classification

- Complex distributions can be modeled as a mixture of multiple Gaussians.
- Example: Two overlapping class distributions modeled using multiple Gaussians.
- Mixture models improve accuracy in complex scenarios.

See. Gaussian Mixture Models

Nonparametric Methods

— — —

- **Parametric Approach: Fixed number of Gaussians per class.**
- **Nonparametric Approach: Allows for flexible, data-driven complexity.**
- **Example: Adding Gaussians dynamically based on data distribution.**
- **Trade-offs:**
 -  **More flexible than fixed models**
 -  **Risk of overfitting**

Naive Bayes Assumption

— — —

- Factorizes the class-conditional density:

$$P(X|Y=k) = P(X_1|Y=k) * P(X_2|Y=k) * * P(X_n|Y=k)$$

- Assumption: Features are conditionally independent given the class.
- Why "Naive"?
- Strong assumption: Real-world data usually has correlated features.
- Despite assumption, works well in practice (e.g., text classification).

Assumption of Same Covariance Matrix

— — —

- 1D Example:

Class 1: Mean of Gaussian at μ_1

Class 2: Mean of Gaussian at μ_2

Same covariance structure for both classes, meaning we can only shift, not change the shape of the Gaussian.

- 2D Example: The covariance for each class results in ellipses that have the same shape.

Visualizing the Covariance Assumption

- **Shape Consistency:** Both classes must have similar shapes in terms of variance.
- **Impact on Decision Boundaries:** If the covariance is the same, the decision boundary between two classes is linear.

Parameter Estimation in LDA

- **Estimating Parameters:** The parameters to estimate include:
- μ (Mean): For each class.
- Σ (Covariance): The shared covariance matrix.
- π (Prior Probability): The proportion of data points in each class.
- **Estimate from Data:** These parameters are derived from training data.

Classifying with LDA

— — —

- **Decision Rule:**
- **Assign data to class k if its probability of belonging to class k is higher than that of any other class.**
- **Multiple Classes:** If there are K classes, you need to make K-1 comparisons to find the correct class.

$$P(Y=k|X) = \{P(X|Y=k)P(Y=k)\} / P(X)$$

$$f_k(X) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X - \mu_k)^T \Sigma^{-1}(X - \mu_k)\right)$$

Simplification of Terms

- Log of Likelihood Ratio:
- The terms involving Σ cancel out because the covariance is the same for all classes.
- This simplification reduces complexity, focusing on means and priors.

Discriminant Function in LDA

— — —

- **Discriminant Function:** The discriminant function assigns a class based on which function value is higher.
- **Maximizing Between-Class Variance:** LDA maximizes the separation between class means while minimizing within-class variance.

LDA as Feature Selection

- **Feature Selection:** LDA can be seen as a feature selection method that finds directions maximizing class separation.
- **Comparison with PCA:**
- **PCA maximizes total variance.**
- **LDA maximizes variance between classes while minimizing variance within classes.**

Discriminant Function in LDA

— — —

- **Discriminant Function:** The discriminant function assigns a class based on which function value is higher.
- **Maximizing Between-Class Variance:** LDA maximizes the separation between class means while minimizing within-class variance.

Derivation and Explanation of the Fisher Criterion for Linear Discriminant Analysis

— — —

- The objective of Linear Discriminant Analysis (LDA) is to find a linear projection of the data that maximizes the separation between different classes while minimizing the spread (variance) within each class.
- This process involves maximizing the between-class variance and minimizing the within-class variance.

Between-Class Variance

— — —

The goal is to maximize the variance between the class means after projecting the data onto a linear subspace defined by the vector w .

Notation: m_1, m_2 : Class means for classes 1 and 2, respectively.

μ_k : Mean for class k (for $k=1,2$).

w : The projection vector.

$w^T x$: The projection of a point x onto the direction w .

Between-Class Variance

— — —

Projection of class means: For each class, we have the means projected onto the direction w :

Projected mean for class 1: $w^T m_1$

Projected mean for class 2: $w^T m_2$

To maximize the distance between the two classes, we want to maximize the quantity $|w^T m_1 - w^T m_2|$

This can be simplified to: $w^T(m_1 - m_2)$

Where

m_1 and m_2 are the class means, and w is the direction vector.

Between-Class Variance

— — —

Normalization of w :

To prevent the projection from becoming unbounded (since scaling w can make $w^T(m_1 - m_2)$ arbitrarily large), we impose a constraint that $\|w\| = 1$, meaning the length of the vector w is normalized to 1.

Within-Class Variance:

— — —

The within-class variance is the variance of the data points within each class with respect to their respective class means. We need to minimize this quantity to ensure that the data within each class is tightly clustered after projection.

The total within-class variance is the sum of the variances within each class.

For class 1, this is given by: $S_1 = \sum_{x \in C_1} (x - m_1)(x - m_1)^T$

Similarly, for class 2: $S_2 = \sum_{x \in C_2} (x - m_2)(x - m_2)^T$

Thus, the total within-class variance is: $S_w = S_1 + S_2$

This term captures the spread of the data points within each class. After projecting onto the direction w , the goal is to minimize the variance of the projected data points.

Fisher Criterion

— — —

The objective is to find the projection vector w that maximizes the between-class variance while minimizing the within-class variance. This leads to the Fisher criterion:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Where:

$S_B = (m_1 - m_2)(m_1 - m_2)^T$: Between-class scatter matrix.

$S_W = S_1 + S_2$: Within-class scatter matrix.

The criterion $J(w)$ is the ratio of between-class variance to within-class variance. We aim to maximize this ratio, which will give us the optimal projection vector w .

Maximizing the Fisher Criterion:

— — —

To find the optimal w , we take the derivative of $J(w)$ with respect to w and set it equal to zero:

This results in the generalized eigenvalue problem:

$$S_W^{-1}S_B w = \lambda w$$

Where λ is the eigenvalue and w is the eigenvector that gives the optimal direction of projection.

Generalization to k Classes:

— — —

When dealing with more than two classes, the procedure generalizes by computing the total between-class variance and within-class variance for all k classes.

Between-Class Variance: The variance of the projected class means for all k classes.

Within-Class Variance: The variance within each class, summed over all classes.

The Fisher criterion for k classes then becomes:

Where S_B is the between-class scatter matrix for k classes, and S_W is the within-class scatter matrix for k classes.

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Sample Question: Fisher Criterion

x_1	x_2	y
2	3	c1
3	3	c1
4	2	c1
7	8	c2
8	7	c2
9	9	c2

x_1	x_2	y
2	3	c1
3	3	c1
4	2	c1
3	2.67	m_1

x_1	x_2	y
7	8	c2
8	7	c2
9	9	c2
8	8	m_2

Step-1: Calculate m_1, m_2

Sample Question: Fisher Criterion

$(m_2 - m_1)$
$8 - 3 = 5$
$8 - 2.67 = 5.33$

$(m_2 - m_1)^T$	5	5.33
-----------------	---	------



25	26.65
26.65	28.41

Step-2: Calculate Between-Class Covariance (B)

$$S_B = (m_2 - m_1)(m_2 - m_1)^T$$

Sample Question: Fisher Criterion

$x_1 - \mu_1$	$x_2 - \mu_1$
2-3=-1	3-2.67=0.33
3-3=0	3-2.67=0.33
4-3=1	2-2.67=-0.67



$(x_1 - \mu_1)$	-1	0	1
	0.33	0.33	-0.67

$(x_1 - \mu_1)^T$	-1	0	1
	0.33	0.33	-0.67



2	-1
-1	0.67



4	0
0	2.67

Step-3: Calculate Within-Class Covariance

$$S_W = \sum_{i=1}^n (x_i - \mu_k)(x_i - \mu_k)^T$$

$x_1 - \mu_2$	$x_2 - \mu_2$
7-8=-1	8-8=0
8-8=0	7-8=-1
9-8=1	9-8=1



$(x_2 - \mu_2)^T$	-1	0
	0	-1
	1	1



2	1
1	2

$(x_2 - \mu_2)$	-1	0	1
	0	-1	1

Sample Question: Fisher Criterion

— — —

S_W	
4	0
0	2.67



S_W^{-1}	
0.25	0
0	0.37



$S_W^{-1}S_B$	6.25	6.66
	9.86	10.51

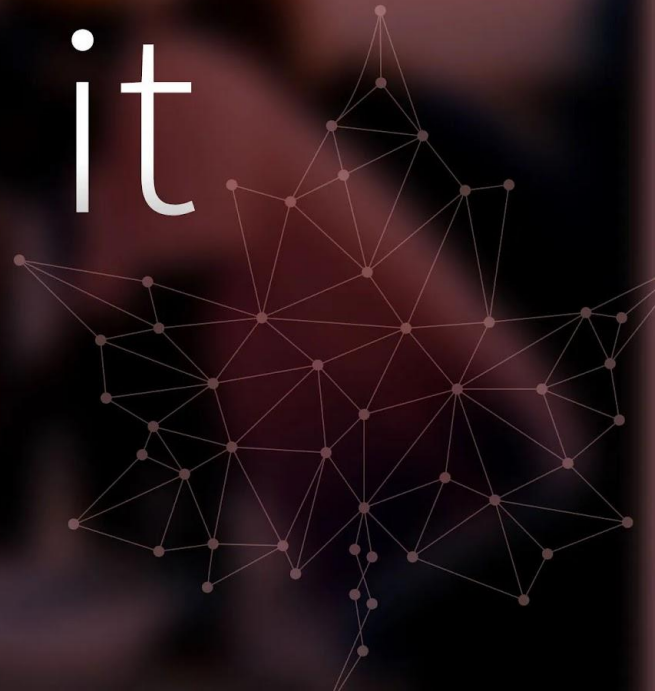
S_B	
25	26.65
26.65	28.41

$\lambda = 16.75, 0.01$
Solve for eigen vectors

Step-2: Solve eigen vector for
 $S_W^{-1}S_B w = \lambda w$

Assignment-3 (Cs-101- 2024) (Week-3)

Let's ^{SOLVE} = it



Question-1

— — —

01:00

For a two-class problem using discriminant functions (δ_k - discriminant function for class k), where is the separating hyperplane located?

- a) Where $\delta_1 > \delta_2$
- b) Where $\delta_1 < \delta_2$
- c) Where $\delta_1 = \delta_2$
- d) Where $\delta_1 + \delta_2 = 1$

Question-1- Correct answer

For a two-class problem using discriminant functions (δ_k - discriminant function for class k), where is the separating hyperplane located?

- a) Where $\delta_1 > \delta_2$ — class-1
- b) Where $\delta_1 < \delta_2$ — class-2
- c) Where $\delta_1 = \delta_2$ — boundary
- d) Where $\delta_1 + \delta_2 = 1$ — incorrect

Correct options: (c)

Question-2

03:00

Given the following dataset consisting of two classes, A and B , calculate the prior probability of each class.

- a) $P(A) = 0.5, P(B)=0.5$
- b) $P(A) = 0.625, P(B)=0.375$
- c) $P(A) = 0.375, P(B)=0.625$
- d) $P(A) = 0.6, P(B)=0.4$

Feature-1	Class
2.3	A
1.8	A
3.2	A
1.7	B
3.0	A
2.1	A
1.9	B
2.4	B

Question-2 - Correct answer

Given the following dataset consisting of two classes, A and B , calculate the prior probability of each class.

- a) $P(A) = 0.5, P(B)=0.5$
- b) $P(A) = 0.625, P(B)=0.375$ — $P(A)= \frac{5}{8}$ and $P(B)= \frac{3}{8}$
- c) $P(A) = 0.375, P(B)=0.625$
- d) $P(A) = 0.6, P(B)=0.4$

Correct options: (b)

Question-3

01:00

In a 3-class classification problem using linear regression, the output vectors for three data points are $[0.8, 0.3, -0.1]$, $[0.2, 0.6, 0.2]$, and $[-0.1, 0.4, 0.7]$.

To which classes would these points be assigned?

- a) 1,2,1
- b) 1,2,2
- c) 1,3,2
- d) 1,2,3

Question-3- Correct answer

— — —

In a 3-class classification problem using linear regression, the output vectors for three data points are $[0.8, 0.3, -0.1]$, $[0.2, 0.6, 0.2]$, and $[-0.1, 0.4, 0.7]$.

To which classes would these points be assigned?

- a) 1,2,1
- b) 1,2,2
- c) 1,3,2
- d) 1,2,3 - (Max at each class - $[0.8, 0.6, 0.7]$)

Correct options: (d)

Question-4

— — —

01:00

If you have a 5-class classification problem and want to avoid masking using polynomial regression, what is the minimum degree of the polynomial you should use?

- a) 3
- b) 4
- c) 5
- d) 6

Question-4 - Correct answer

— — —

If you have a 5-class classification problem and want to avoid masking using polynomial regression, what is the minimum degree of the polynomial you should use?

- a) 3
- b) 4
- c) 5
- d) 6

Correct options: (b)

Question-5

— — —

01:00

Consider a logistic regression model where the predicted probability for a given data point is 0.4. If the actual label for this data point is 1, what is the contribution of this data point to the log-likelihood?

- a) -1.3219
- b) -0.9163
- c) +1.3219
- d) +0.9163

Question-5 - Correct answer

Consider a logistic regression model where the predicted probability for a given data point is 0.4. If the actual label for this data point is 1, what is the contribution of this data point to the log-likelihood?

- a) -1.3219
- b) -0.9163 — $[1 \cdot \log(0.4) + (1-1) \cdot \log(1-0.4)]$
- c) +1.3219
- d) +0.9163

Correct options: (b)

Question-6

— — —

01:00

What additional assumption does LDA make about the covariance matrix in comparison to the basic assumption of Gaussian class conditional density?

- a) The covariance matrix is diagonal
- b) The covariance matrix is identical
- c) The covariance matrix is the same for all classes
- d) The covariance matrix is different for each class

Question-6 - Correct answer

What additional assumption does LDA make about the covariance matrix in comparison to the basic assumption of Gaussian class conditional density?

- a) The covariance matrix is diagonal
- b) The covariance matrix is identical
- c) The covariance matrix is the same for all classes
- d) The covariance matrix is different for each class

Correct options: (c)

Question-7

— — —

03:00

What is the shape of the decision boundary in LDA?

- a) Quadratic
- b) Linear
- c) Circle
- d) Can not be determined

Question-7 - Correct answer

What is the shape of the decision boundary in LDA?

- a) Quadratic
- b) Linear
- c) Circle
- d) Can not be determined

Correct options: (b)

Question-8

— — —

01:00

For two classes C_1 and C_2 with within-class variances $\sigma_{w1}^2=1$ and $\sigma_{w2}^2=4$ respectively, if the projected means are $\mu'_1=1$ and $\mu'_2=3$, what is the Fisher criterion $J(w)$?

- a) 0.5
- b) 0.8
- c) 1.25
- d) 1.5

Question-8- Correct answer

For two classes C_1 and C_2 with within-class variances $\sigma_{w1}^2=1$ and $\sigma_{w2}^2=4$ respectively, if the projected means are $\mu'_1=1$ and $\mu'_2=3$, what is the Fisher criterion $J(w)$?

- a) 0.5
- b) 0.8**
- c) 1.25
- d) 1.5

$$\begin{aligned} J(w) &= (\text{Between Class}) / (\text{within class}) \\ &= ((1-3)^2) / (1+4) \\ &= 4/5 = 0.8 \end{aligned}$$

Correct options: (b)

Question-9

— — —

01:00

Given two classes C_1 and C_2 with means $\mu_1 = [2 \ 3]^T$ and $\mu_2 = [5 \ 7]^T$ respectively, what is the direction vector w for LDA when the within-class covariance matrix S_w is the identity matrix I ?

- a) $[4 \ 3]^T$
- b) $[5 \ 7]^T$
- c) $[0.7 \ 0.7]^T$
- d) $[0.6 \ 0.8]^T$

Question-9- Correct answer

Given two classes C_1 and C_2 with means $\mu_1 = [2 \ 3]^T$ and $\mu_2 = [5 \ 7]^T$ respectively, what is the direction vector w for LDA when the within-class covariance matrix S_w is the identity matrix I ?

- a) $[4 \ 3]^T$
- b) $[5 \ 7]^T$
- c) $[0.7 \ 0.7]^T$
- d) $[0.6 \ 0.8]^T$

$$S_w = I$$

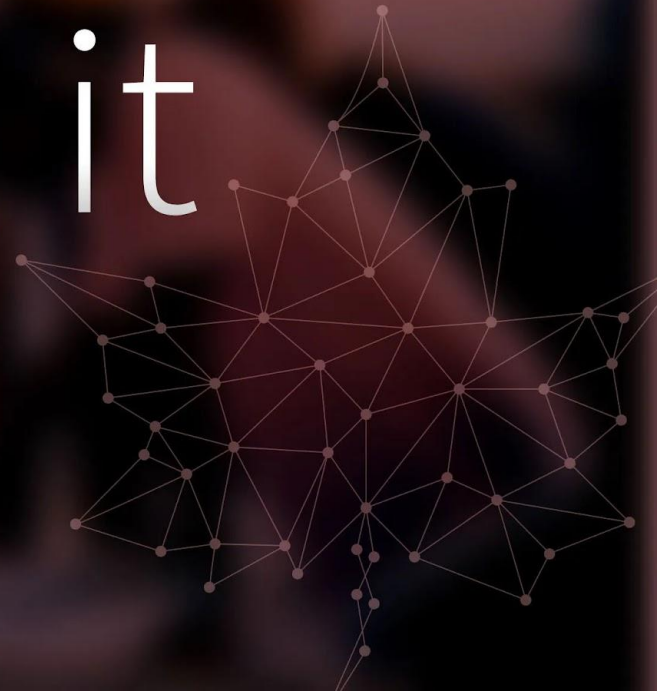
S_B	
9	12
12	16

Solve largest eigen vector for $S_w^{-1} S_B$ and normalize it

Correct options: (d)

Assignment-3 (Cs-46- 2025) (Week-3)

Let's ^{SOLVE} = it



Question-1

— — —

01:00

Which of the following statement(s) about decision boundaries and discriminant functions of classifiers is/are true?

- a) In a binary classification problem, all points x on the decision boundary satisfy $\delta_1(x) = \delta_2(x)$
- b) In a three-class classification problem, all points on the decision boundary satisfy $\delta_1(x) = \delta_2(x) = \delta_3(x)$.
- c) In a three-class classification problem, all points on the decision boundary satisfy at least one of $\delta_1(x) = \delta_2(x)$, $\delta_2(x) = \delta_3(x)$ or $\delta_3(x) = \delta_1(x)$.
- d) If x does not lie on the decision boundary then all points lying in a sufficiently small neighbourhood around x belong to the same class.

Question-1- Correct answer

— — —

Which of the following statement(s) about decision boundaries and discriminant functions of classifiers is/are true?

- a) In a binary classification problem, all points x on the decision boundary satisfy $\delta_1(x) = \delta_2(x)$
- b) In a three-class classification problem, all points on the decision boundary satisfy $\delta_1(x) = \delta_2(x) = \delta_3(x)$.
- c) In a three-class classification problem, all points on the decision boundary satisfy at least one of $\delta_1(x) = \delta_2(x)$, $\delta_2(x) = \delta_3(x)$ or $\delta_3(x) = \delta_1(x)$.
- d) If x does not lie on the decision boundary then all points lying in a sufficiently small neighbourhood around x belong to the same class.

Correct options: (a)(c)(d)

Question-2

03:00

The following table gives the binary ground truth labels y_i for four input points x_i (not given). We have a logistic regression model with some parameter values that computes the probability $p_1(x_i)$ that the label is 1. Compute the likelihood of observing the data given these model parameters

- a) 0.072
- b) 0.144
- c) 0.288
- d) 0.002

y_i	$p_1(x_i)$
1	0.8
0	0.5
0	0.2
1	0.9

Question-2 - Correct answer

— — —

The following table gives the binary ground truth labels y_i for four input points x_i (not given). We have a logistic regression model with some parameter values that computes the probability $p_1(x_i)$ that the label is 1. Compute the likelihood of observing the data given these model parameters

- a) 0.072
- b) 0.144
- c) 0.288 ($0.8 * 0.5 * 0.8 * 0.9$)
- d) 0.002

Correct options: (c)

Question-3

— — —

01:00

You train an LDA classifier on a dataset with 2 classes. The decision boundary is significantly different from the one obtained by logistic regression. What could be the reason?

- a) The underlying data distribution is Gaussian
- b) The two classes have equal covariance matrices
- c) The underlying data distribution is not Gaussian
- d) The two classes have unequal covariance matrices

Question-3- Correct answer

You train an LDA classifier on a dataset with 2 classes. The decision boundary is significantly different from the one obtained by logistic regression. What could be the reason?

- a) The underlying data distribution is Gaussian
- b) The two classes have equal covariance matrices
- c) The underlying data distribution is not Gaussian
- d) The two classes have unequal covariance matrices

Correct options: (c)(d)

Question-4

— — —

01:00

Which of the following statement(s) about logistic regression is/are true?

- a) It learns a model for the probability distribution of the data points in each class.
- b) The output of a linear model is transformed to the range (0, 1) by a sigmoid function.
- c) The parameters are learned by minimizing the mean-squared loss.
- d) The parameters are learned by maximizing the log-likelihood.

Question-4 - Correct answer

Which of the following statement(s) about logistic regression is/are true?

- a) It learns a model for the probability distribution of the data points in each class.
- b) The output of a linear model is transformed to the range (0, 1) by a sigmoid function.
- c) The parameters are learned by minimizing the mean-squared loss.
- d) The parameters are learned by maximizing the log-likelihood.

Correct options: (b)(d)

Question-5

01:00

Consider a modified form of logistic regression given below where k is a positive constant and β_0 and β_1 are parameters. So $p(x) =$

$$\log = \left(\frac{1-p(x)}{kp(x)} \right) = \beta_0 + \beta_1 x$$

a)

☐
$$\frac{e^{-\beta_0}}{ke^{-\beta_0} + e^{\beta_1 x}}$$

b)

☐
$$\frac{e^{-\beta_1 x}}{e^{-\beta_0} + e^{k\beta_1 x}}$$

c)

☐
$$\frac{e^{\beta_1 x}}{ke^{\beta_0} + e^{\beta_1 x}}$$

d)

☐
$$\frac{e^{-\beta_1 x}}{ke^{\beta_0} + e^{-\beta_1 x}}$$

Question-5 - Correct answer

Consider a modified form of logistic regression given below where k is a positive constant and β_0 and β_1 are parameters.

$$\log = \left(\frac{1-p(x)}{kp(x)} \right) = \beta_0 + \beta_1 x$$

a)

☐
$$\frac{e^{-\beta_0}}{ke^{-\beta_0} + e^{\beta_1 x}}$$

b)

☐
$$\frac{e^{-\beta_1 x}}{e^{-\beta_0} + e^{k\beta_1 x}}$$

c)

☐
$$\frac{e^{\beta_1 x}}{ke^{\beta_0} + e^{\beta_1 x}}$$

d)

☐
$$\frac{e^{-\beta_1 x}}{ke^{\beta_0} + e^{-\beta_1 x}}$$

Correct options: (d)

Question-6

03:00

Consider a Bayesian classifier for a 5-class classification problem. The following tables give the class-conditioned density $f_k(\mathbf{x})$ for class $k \in \{1, 2, \dots, 5\}$ at some point \mathbf{x} in the input space.

- a) The predicted label at \mathbf{x} will always be class 4.
- b) If $2\pi_i \leq \pi_{i+1} \forall i \in \{1, \dots, 4\}$, the predicted class must be class 4
- c) If $\pi_i \geq (3/2)\pi_{i+1} \forall i \in \{1, \dots, 4\}$, the predicted class must be class 1
- d) The predicted label at \mathbf{x} can never be class 5

k	$f_k(\mathbf{x})$
1	0.15
2	0.20
3	0.05
4	0.50
5	0.01

Question-6 - Correct answer

Consider a Bayesian classifier for a 5-class classification problem. The following tables give the class-conditioned density $f_k(x)$ for class $k \in \{1, 2, \dots, 5\}$ at some point x in the input space.

- a) The predicted label at x will always be class 4.
- b) If $2_{\pi_i} \leq \pi_{i+1} \forall i \in \{1, \dots, 4\}$, the predicted class must be class 4
- c) If $\pi_i \geq (3/2)\pi_{i+1} \forall i \in \{1, \dots, 4\}$, the predicted class must be class 1
- d) The predicted label at x can never be class 5

Correct options: (b)(c)

Question-7

— — —

03:00

Which of the following statement(s) about a two-class LDA classification model is/are true?

- a) On the decision boundary, the prior probabilities corresponding to both classes must be equal.
- b) On the decision boundary, the posterior probabilities corresponding to both classes must be equal.
- c) On the decision boundary, class-conditioned probability densities corresponding to both classes must be equal.
- d) On the decision boundary, the class-conditioned probability densities corresponding to both classes may or may not be equal.

Question-7 – Correct answer

— — —

Which of the following statement(s) about a two-class LDA classification model is/are true?

- a) On the decision boundary, the prior probabilities corresponding to both classes must be equal.
- b) On the decision boundary, the posterior probabilities corresponding to both classes must be equal.
- c) On the decision boundary, class-conditioned probability densities corresponding to both classes must be equal.
- d) On the decision boundary, the class-conditioned probability densities corresponding to both classes may or may not be equal.

Correct options: (b)(d)

Question-8

01:00

Consider the following two datasets and two LDA classifier models trained respectively on these datasets. **Dataset A: 200 samples of class 0; 50 samples of class 1. Dataset B: 200 samples of class 0 (same as Dataset A); 100 samples of class 1 created by repeating twice the class 1 samples from Dataset A.** Let the classifier decision boundary learnt be of the form $w^T x + b = 0$ where, w is the slope and b is the intercept. Which of the given statement is true?

- a) The learned decision boundary will be the same for both models.
- b) The two models will have the same slope but different intercepts.
- c) The two models will have different slopes but the same intercept.
- d) The two models may have different slopes and different intercepts

Question-8- Correct answer

Consider the following two datasets and two LDA classifier models trained respectively on these datasets. Dataset A: 200 samples of class 0; 50 samples of class 1. Dataset B: 200 samples of class 0 (same as Dataset A); 100 samples of class 1 created by repeating twice the class 1 samples from Dataset A. Let the classifier decision boundary learnt be of the form $w^T x + b = 0$ where, w is the slope and b is the intercept. Which of the given statement is true?

- a) The learned decision boundary will be the same for both models.
- b) The two models will have the same slope but different intercepts.
- c) The two models will have different slopes but the same intercept.
- d) The two models may have different slopes and different intercepts

Correct options: (b)

Question-9

01:00

Which of the following statement(s) about LDA is/are true?

- a) It minimizes the inter-class variance relative to the intra-class variance.
- b) It maximizes the inter-class variance relative to the intra-class variance.
- c) Maximizing the Fisher information results in the same direction of the separating hyperplane as the one obtained by equating the posterior probabilities of classes.
- d) Maximizing the Fisher information results in a different direction of the separating hyperplane from the one obtained by equating the posterior probabilities of classes.

Question-9- Correct answer

Which of the following statement(s) about LDA is/are true?

- a) It minimizes the inter-class variance relative to the intra-class variance.
- b) It maximizes the inter-class variance relative to the intra-class variance.
- c) Maximizing the Fisher information results in the same direction of the separating hyperplane as the one obtained by equating the posterior probabilities of classes.
- d) Maximizing the Fisher information results in a different direction of the separating hyperplane from the one obtained by equating the posterior probabilities of classes.

Correct options: (b)(c)

Question-10

— — —

01:00

Which of the following statement(s) regarding logistic regression and LDA is/are true for a binary classification problem?

- a) For any classification dataset, both algorithms learn the same decision boundary.
- b) Adding a few outliers to the dataset is likely to cause a larger change in the decision boundary of LDA compared to that of logistic regression.
- c) Adding a few outliers to the dataset is likely to cause a similar change in the decision boundaries of both classifiers.
- d) If the intra-class distributions deviate significantly from the Gaussian distribution, logistic regression is likely to perform better than LDA.

Question-10- Correct answer

— — —

Which of the following statement(s) regarding logistic regression and LDA is/are true for a binary classification problem?

- a) For any classification dataset, both algorithms learn the same decision boundary.
- b) Adding a few outliers to the dataset is likely to cause a larger change in the decision boundary of LDA compared to that of logistic regression.
- c) Adding a few outliers to the dataset is likely to cause a similar change in the decision boundaries of both classifiers.
- d) If the intra-class distributions deviate significantly from the Gaussian distribution, logistic regression is likely to perform better than LDA.

Correct options: (b) (d)



THANK YOU

Suggestions and Feedback



Next Session:

**Sunday: 17-Aug-2025
3:00 – 5:00 PM**