Linear Classification

MA515: Foundations of Data Science

Arun Kumar

Department of Mathematics, IIT Ropar

August 14, 2025

Panoramic View of the Talk

1. Classification Problem
2. Linear Regression as Classifier
3. Logistic Regression
4. Linear Discriminant Analysis

Example (Loan approval)

- Given the properties x of a customer like age, income, liability, assets, job
- To predict $f(x)$, the loan approved or not.

Example (Graduate admission)

- Given the properties x of a student like grades, GRE score, university ranking, subject of interest
- To predict $f(x)$, student get admission or not.

Linear Classifier

- A Classifier partitions input space into decision regions
- An input/feature space which is linearly separable can be partitioned by a linear decision boundary.

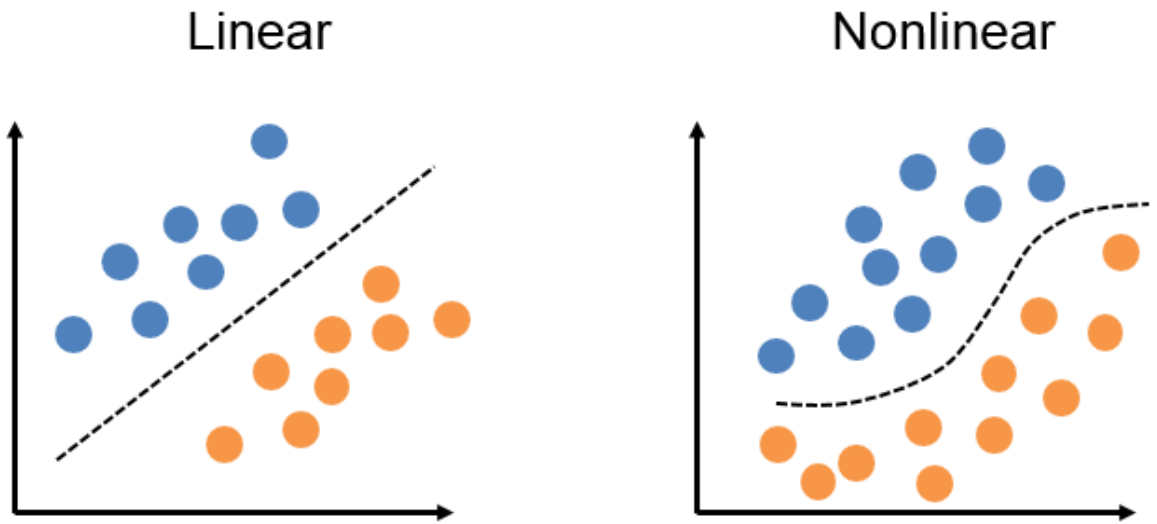


Figure: Linear vs Non-linear Classifier

Linear Regression as Classifier

Linear Regression as Classifier

- Suppose, we have two classes and 1 feature only.
- We substitute the labels with 0 and 1.
- Fit the SLR and if the predicted value is ≤ 0.5, we predict the label as 0 and if it is greater than 0.5, we predict the label as 1.

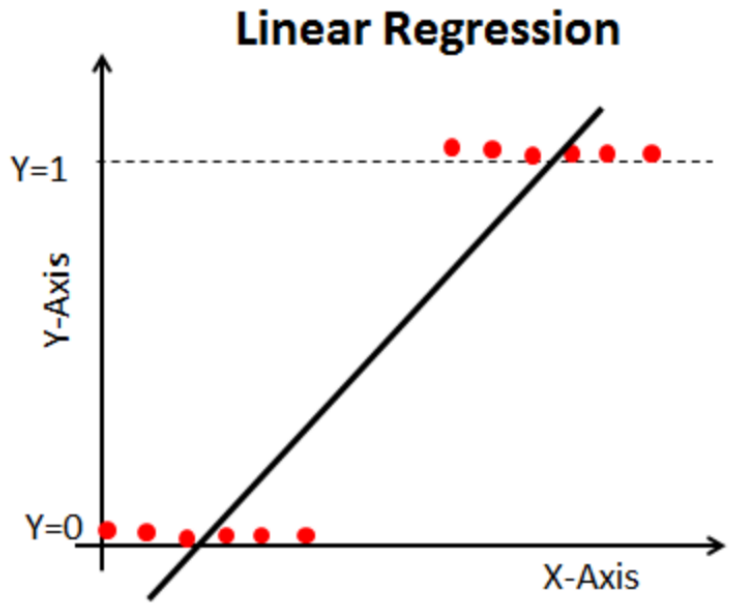


Figure: Linear Regression as Classifier

- Get the model $\hat{f} = X\hat{\beta}$, where

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

- For a new data point U, if the predicted value $U\hat{\beta} \leq 0.5$, we predict label 0 and if it is greater than 0.5, we predict label 1.
- Here U is a row vector $U = (u_1, u_2, \ldots, u_p)$.

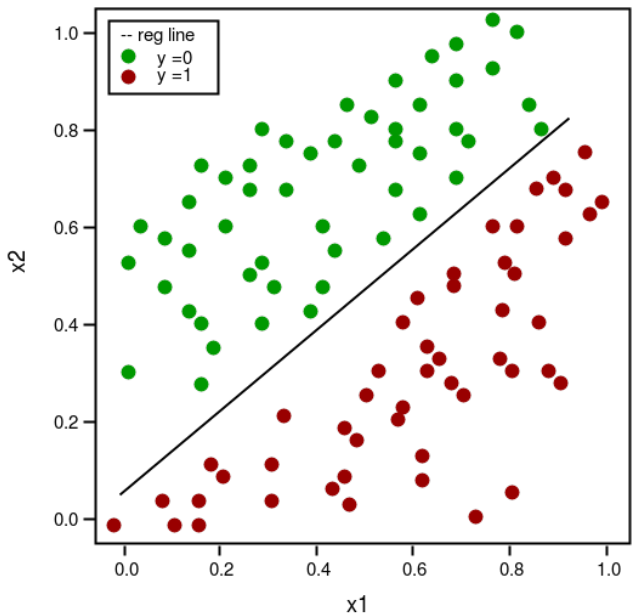


Figure: Linear regression as a classifier with 2 features

Sensitivity of Linear Regression

Linear regression is sensitive to imbalanced data. An Imbalanced Dataset refers to a situation where the number of instances across different classes in a classification problem is not evenly distributed.

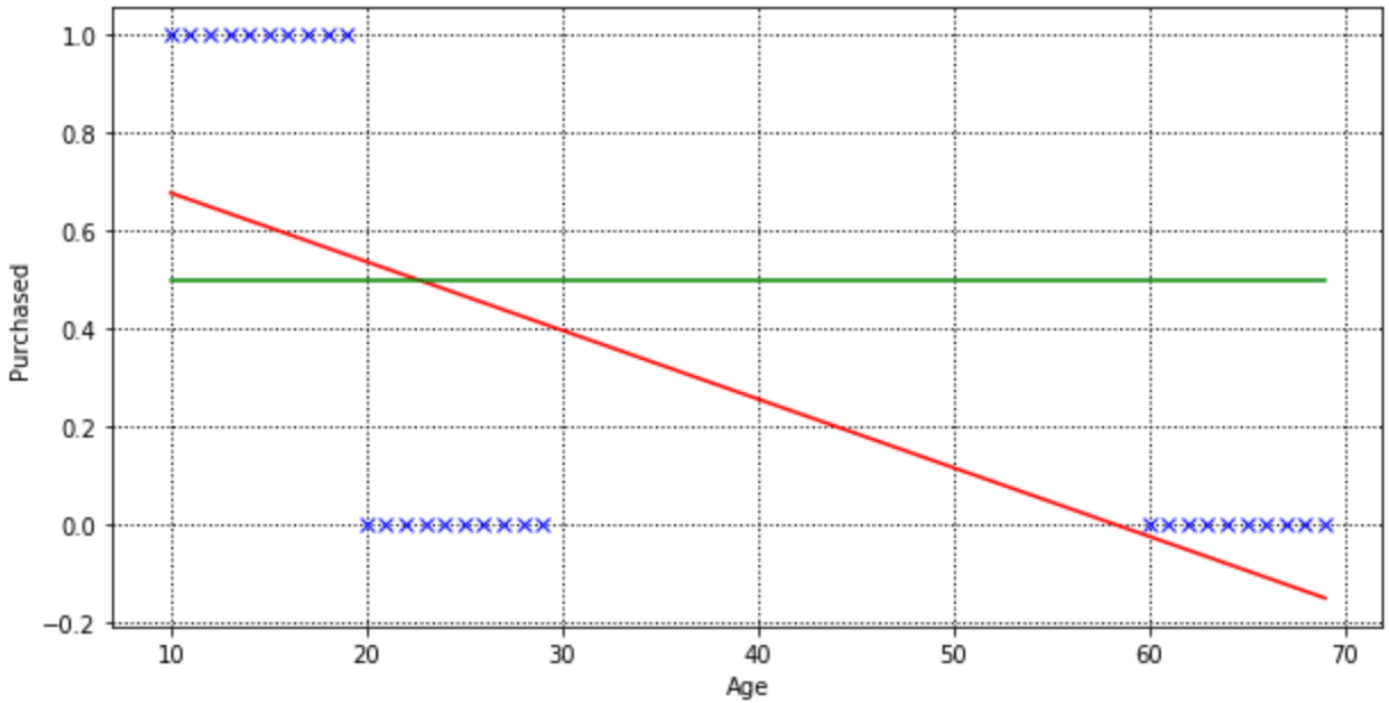


Figure: Linear regression sensitivity

Logistic Regression as Classifier

Definition

A sigmoid function is any mathematical function whose graph is *S*-shaped or sigmoid curve. A sigmoid function is a bounded, differentiable, real function and has a positive derivative at each point.

Example (Logistic function)

$$f(x) = \frac{1}{1 + e^{-x}}, \ x \in \mathbb{R}.$$

Example (Error function)

$$f(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt, \ x > 0.$$

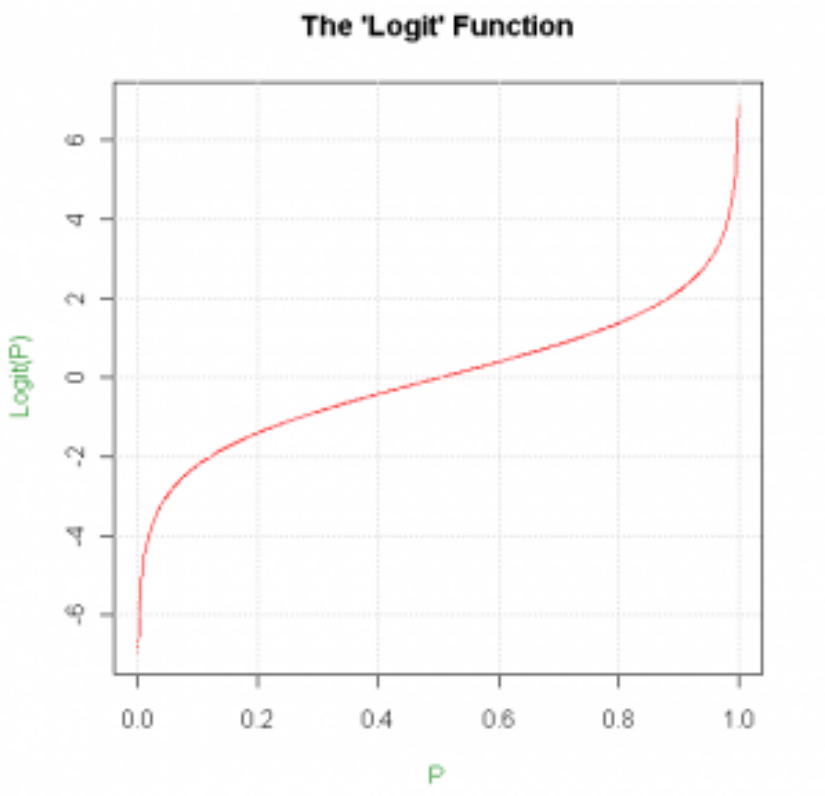
Example (Hyperbolic tangent or Tan Hyperbolic)

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \ x \in \mathbb{R}.$$

Definition (Logit Function)

The logit function takes p values in the range $(0, 1)$ and transform them to y values in the interval $(-\infty, \infty)$ and is defined by

$$logit(p) = \log\left(\frac{p}{1-p}\right).$$



Definition (Inverse-logit Function)

The inverse-logit (or logistic) function does the reverse, and takes x values along the real line and transform them to y values in the interval $(0, 1)$ and is defined by

$$logit^{-1}(x) = logistic(x) = \frac{e^x}{1 + e^x}.$$

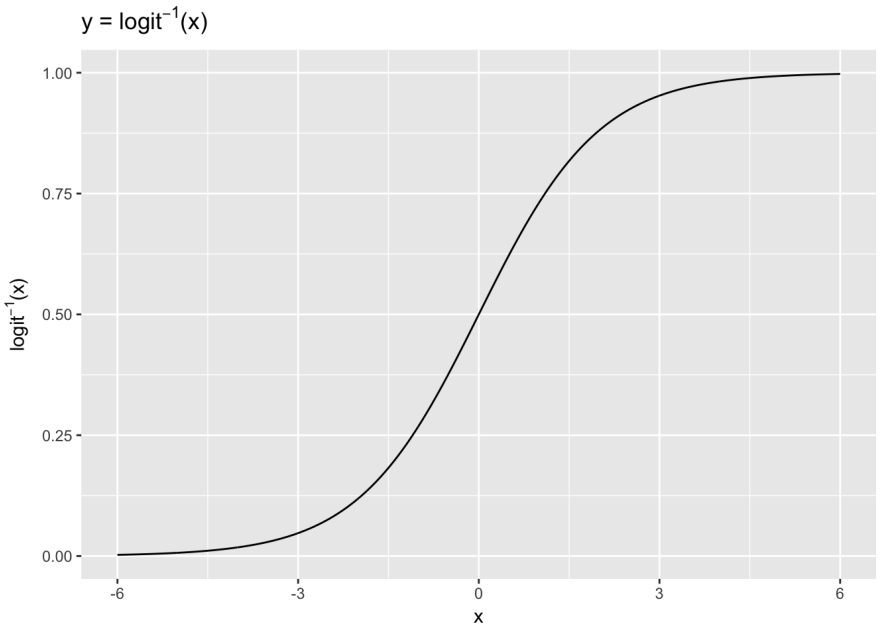


Figure: Logistic function

- In logistic regression rather than predicting 1 and 0, we predict the probability of the response variable.
- The logistic regression model is given by

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

or equivalently

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}.$$

- Here x is the independent variable and $p(x)$ is the probability of response variable taking value 1.

Estimating β_0 and β_1

- Our aim is to use the training data $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, to estimate the parameters β_0 and β_1.
- We can use the maximum likelihood estimation technique.
- The likelihood function is

$$L(\beta_0, \beta_1 | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}.$$

- The log-likelihood function is

$$\begin{aligned}
\log L(\beta_0, \beta_1 | x_1, x_2, \ldots, x_n) &= \sum_{i=1}^{n} [y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))] \\
&= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i \log\left(\frac{p(x_i)}{1 - p(x_i)}\right) \\
&= -\sum_{i=1}^{n} \log\left(1 + e^{\beta_0 + \beta_1 x_i}\right) + \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_i).
\end{aligned}$$

Estimating β_0 and β_1 contd...

- The log-likelihood function is

$$\log L(\beta_0, \beta_1 | x_1, x_2, \ldots, x_n) = -\sum_{i=1}^{n} \log\left(1 + e^{\beta_0 + \beta_1 x_i}\right) + \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i).$$

- If we take derivatives with respect to β_0 and β_1 of the log-likelihood function and put these to equal to 0 to get the critical points.
- We see that it is not possible to solve these equations in term of β_0 and β_1.
- We choose β_0 and β_1, which maximizes the log-likelihood function.
- We need some **iterative method** and generally **Newton's method** or **stochastic gradient method** is used.

Linear Discriminant Analysis

Example

Suppose you have 4 coins. Three of them are fair coins but the fourth is a weighted coin which has 80% chance of landing heads up. You take a coin randomly out of your pocket and toss it. Suppose it lands heads up. What is the probability that the coin is the weighted coin ?

Prior: $\pi(\text{fair}) = 0.75$, $\pi(\text{weighted}) = 0.25$.

Likelihood: $\mathbb{P}(\text{head}|\text{fair}) = 0.5$ and $\mathbb{P}(\text{head}|\text{weighted}) = 0.8$.

Posterior:

$$\mathbb{P}(\textit{fair}|\textit{head}) = \frac{\mathbb{P}(\textit{head}|\textit{fair})\mathbb{P}(\textit{fair})}{\mathbb{P}(\textit{head}|\textit{fair})\mathbb{P}(\textit{fair}) + \mathbb{P}(\textit{head}|\textit{weighted})\mathbb{P}(\textit{weighted})} = 0.652.$$

$$\mathbb{P}(\textit{weighted}|\textit{head}) = 0.348.$$

Bayes' theorem - different representations

Discrete parameter discrete data

$$p_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)}$$

Discrete parameter continuous data

$$p_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{f_X(x)}$$

Continuous parameter discrete data

$$f_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{p_X(x)}$$

Continuous parameter continuous data

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{f_X(x)}$$

Example

Find the posterior distribution of probability of heads denoted by θ on a coin. Suppose x heads are observed in n tosses.

Solution: Suppose we assume a uniform prior for Θ i.e. $f_\Theta(\theta) = 1,\ 0 < \theta < 1$. Also

$$p_{X|\Theta}(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

We then have

$$\begin{aligned} f_{\Theta|X}(\theta|x) &= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{\int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta} \\ &= \frac{(n+1)!\theta^x(1-\theta)^{n-x}}{x!(n-x)!}. \end{aligned}$$

- Suppose that, we wish to classify an observation into one of K classes, where $K \geq 2$.
- Let $\pi_k = \mathbb{P}(Y = k)$ represent the **prior probability** that a randomly chosen observation comes from the kth class.
- Let

$$f_k(x) = \mathbb{P}(X = x | Y = k),$$

  denote the density function of X for an observation that come from the class k.
- In other words, $f_k(x)$ represent the **likelihood** that an observation in the kth class has $X \approx x$.

- Using Bayes' theorem

$$\mathbb{P}(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}.$$

- To get $\mathbb{P}(Y = k|X = x)$, we put the estimate of π_k and $f_k(x)$.
- These estimates are observed from the training data.

- We assume that $p = 1$, i.e. we have one predictor only.
- Assume further that $f_k(x)$ is Gaussian i.e.

$$f_k(x) = \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}},$$

  where μ_k and σ_k^2 are the mean and variance parameters for the kth class.
- Let us further assume that the variance among K classes is same i.e. $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2 = \sigma^2$.
- Thus

$$\mathbb{P}(Y = k|X = x) = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \pi_k}{\sum_{l=1}^{K} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_l)^2}{2\sigma^2}} \pi_l}.$$

- The Bayes classifier assign an observation to the class for which $\mathbb{P}(Y = k|X = x)$ is higher.
- Since the denominator is same, hence the classifier assign $X = x$ to the class where numerator is higher i.e.

$$g_k(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \pi_k$$

is highest.

- Equivalently, $X = x$ is assigned to the class for which

$$\log g_k(x) = \log(\pi_k) - \log(\sigma\sqrt{2\pi}) - \frac{(x-\mu_k)^2}{2\sigma^2}$$

is highest.

- Alternatively, $X = x$ is assigned to the class for which

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

is largest.

- For simplicity assume that $k = 2$ and $\pi_1 = \pi_2$.
- The classifier assign an observation $X = x$ to the class 1 if

$$x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log(\pi_1) > x\frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log(\pi_2).$$

- Equivalently, $X = x$ is assigned to the class 1, if

$$x(\mu_1 - \mu_2) > \frac{1}{2}(\mu_1^2 - \mu_2^2) \implies x > \frac{\mu_1 + \mu_2}{2},$$

  provided $\mu_1 > \mu_2$.
- The Bayes' decision boundary in this case will be

$$x = \frac{\mu_1 + \mu_2}{2}.$$

Estimation of parameters in LDA

- We estimate the parameters from the training data.
- The mean for kth class is estimated by

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{\{i:y_i=k\}} x_i.$$

- The common variance is estimated by

$$\hat{\sigma}^2 = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{\{i:y_i=k\}} (x_i - \hat{\mu}_k)^2.$$

- Further, $\hat{\pi}_k = \frac{n_k}{N}$.
- Here N is the total number of observations and n_k counts the number of observations in the kth class.

LDA for Bivariate Binary Classification

Setup of LDA

- Goal: Classify an observation $\mathbf{x}$ into one of two classes $\mathcal{C}_1$ or $\mathcal{C}_2$
- Assume each class is normally distributed with:

$$\mathbf{x} \mid \mathcal{C}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma), \quad k = 1, 2$$

- **Key assumption:** Both classes share the same covariance matrix Σ
- LDA finds a **linear boundary** that best separates the classes

Bayes decision rule:

$$\text{Assign } \mathbf{x} \text{ to } \mathcal{C}_1 \text{ if } p(\mathcal{C}_1 \mid \mathbf{x}) > p(\mathcal{C}_2 \mid \mathbf{x})$$

Using Gaussian likelihoods with same Σ, the rule simplifies to:

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \Sigma^{-1} \boldsymbol{\mu}_k + \ln \pi_k$$

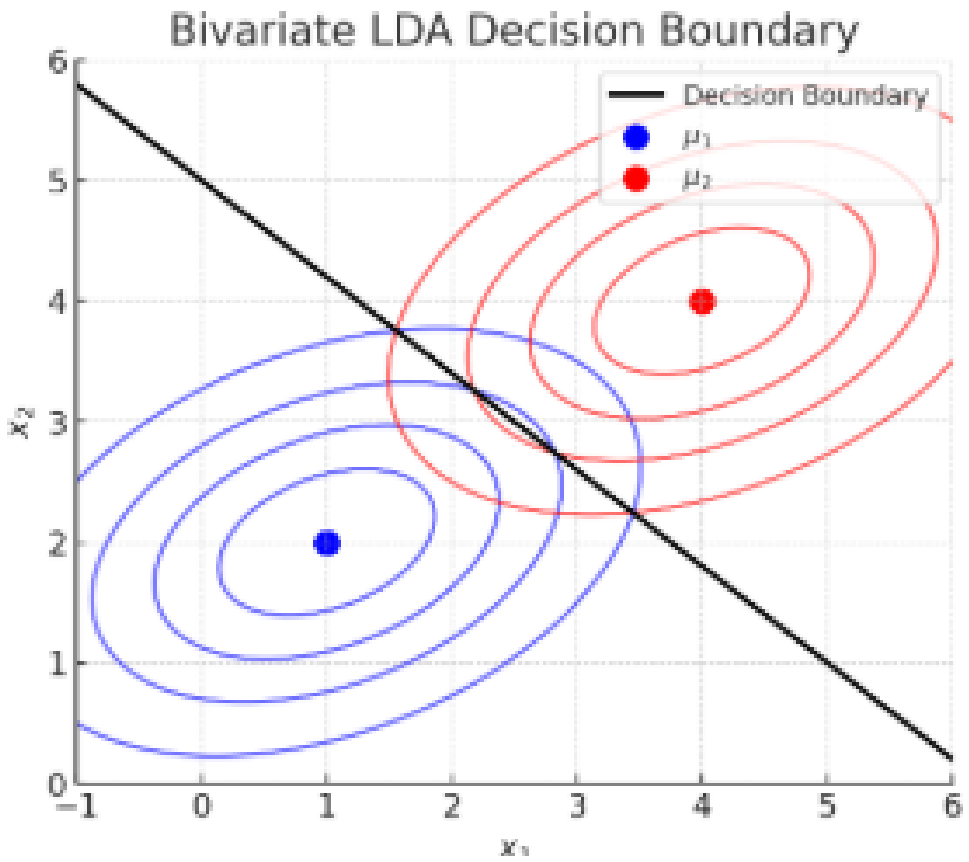
Decision Boundary Equation

Set $\delta_1(\mathbf{x}) = \delta_2(\mathbf{x})$:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1} \mathbf{x} = \frac{1}{2}\left(\boldsymbol{\mu}_1^\top \Sigma^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \Sigma^{-1} \boldsymbol{\mu}_2\right) - \ln \frac{\pi_1}{\pi_2}$$

- Linear equation in $\mathbf{x}$
- Normal vector: $\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

LDA in Action



- Ellipses: Contours of the bivariate normal distributions
- Solid line: Decision boundary
- Points: Sample observations

THANK YOU!