

DS 4400

Machine Learning and Data Mining I Spring 2024

David Liu

Khoury College of Computer Science
Northeastern University

February 16 2024

Outline

- Announcements on homework, final project, and midterm.
- Generative classifiers
 - Difference from discriminative classifiers
- Linear Discriminant Analysis (LDA)
 - Training and inference
 - Why LDA is a linear classifier
 - Comparison with Logistic Regression

Announcements

- Please submit Homework 2. Deadline is tonight at 11:59pm
- Reminder that you have five late days to use over the course of the semester for homework assignments.
- Please tag pages to questions in Gradescope.
- Homework 1 grades will be released this afternoon. Re-grade requests feature is enabled on Gradescope.
 - Questions 1 and 2: Dhanush
 - Question 3: Jai
 - Questions 4 and 5: Caleb

Announcements - Midterm

Midterm is next Friday February 23.

Please come to class a few minutes early so that we can give as much time as possible for the exam.

You are allowed a one-page cheat sheet and a calculator.

Announcements – Final Project

- Final project resources released on Canvas
 - Document with example datasets
 - Examples of past submissions
- Project proposal due on Friday March 1 before Spring Break.

Announcements – Tuesday's Lecture

Next Tuesday February 20, we will have a lecture on the Ethics of AI / the societal impacts of AI.

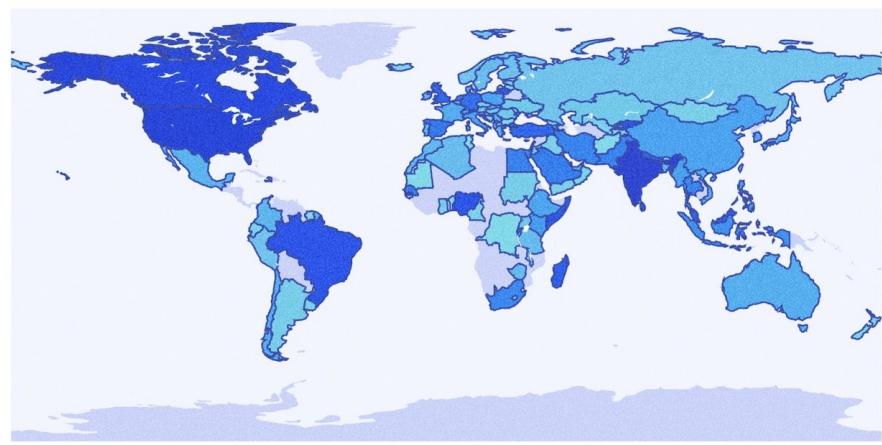
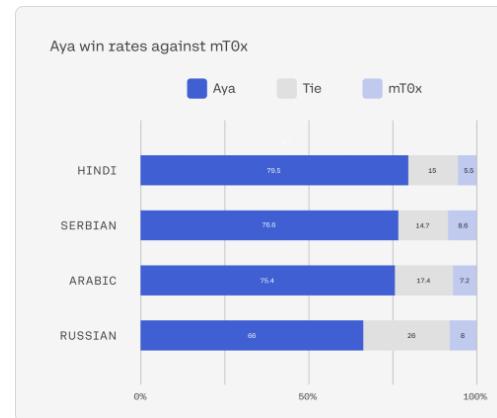
Please bring a laptop to the call as there will be an interactive component led by my labmate Samantha Dies.

News

Cohere for AI launches an open-source LLM, Aya, supporting 101 languages



Image by DALL-E 3 for VentureBeat



Generative vs Discriminative

- **Generative model**
 - Given X and Y, learns the joint probability $P(X, Y)$
 - Can generate more examples from distribution
 - Examples: LDA, Naïve Bayes, language models (GPT-2, GPT-3, BERT)
- **Discriminative model**
 - Given X and Y, learns a decision function for classification

LDA

- Classify to one of k classes
- Logistic regression computes directly
 - $P[Y = 1|X = x]$
- LDA uses Bayes Theorem to estimate it

LDA

- Classify to one of k classes
- Logistic regression computes directly
 - $P[Y = 1|X = x]$
 - Assume sigmoid function
- LDA uses Bayes Theorem to estimate it
 - $P[Y = k|X = x] = \frac{P[X = x|Y = k]P[Y=k]}{P[X=x]}$
 - Let $\pi_k = P[Y = k]$ be the prior probability of class k and $f_k(x) = P[X = x|Y = k]$

LDA

Assume $f_k(x)$ is Gaussian!

Unidimensional case (d=1)

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

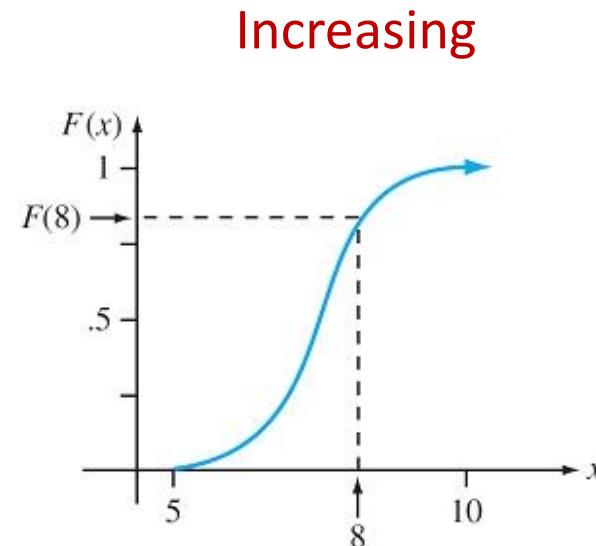
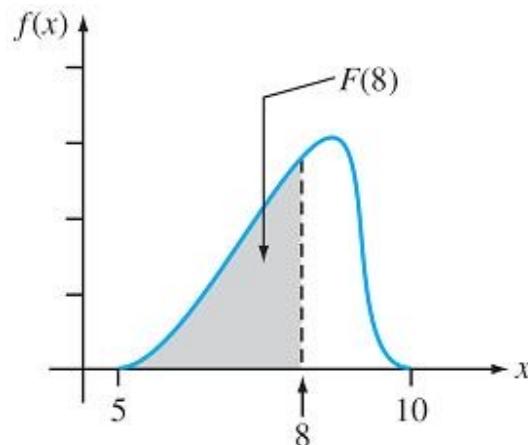
Continuous Random Variables

- $X: U \rightarrow V$ is continuous RV if it takes infinite number of values
- The **cumulative distribution function CDF** $F: R \rightarrow \{0,1\}$ for X is defined for every value x by:

$$F(x) = \Pr(X \leq x)$$

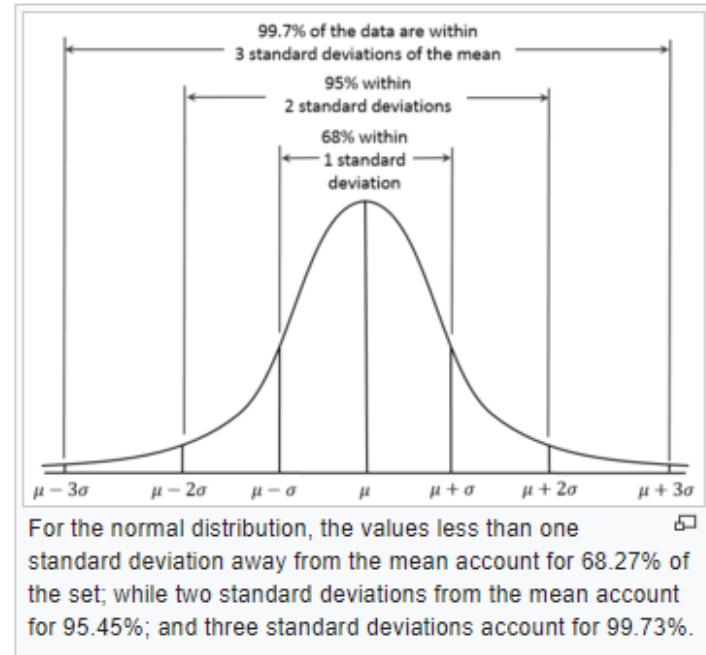
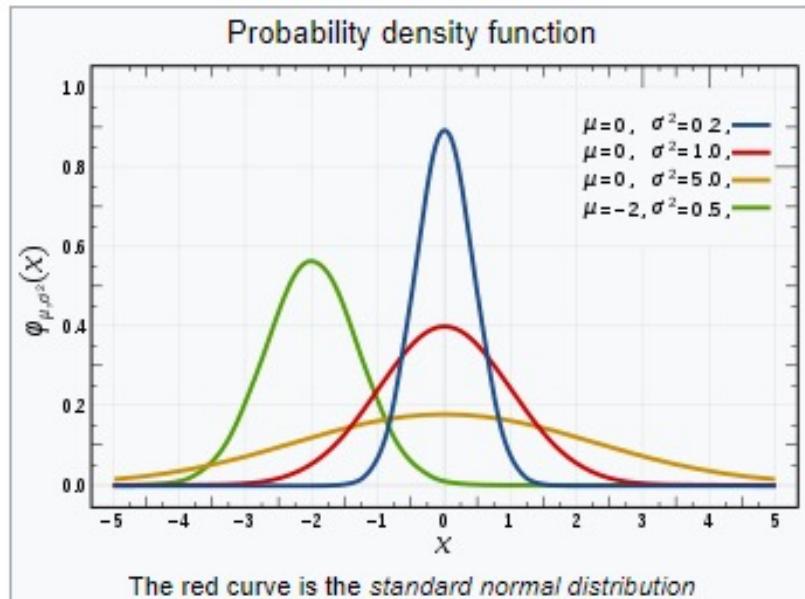
- The **probability distribution function PDF** $f(x)$ for X is

$$f(x) = dF(x)/dx$$



Gaussian Distribution

Normal Distribution



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

LDA

Assume $f_k(x)$ is Gaussian!

Unidimensional case (d=1)

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

LDA

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

Assume $f_k(x)$ is Gaussian!

Unidimensional case (d=1)

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}.$$

Assumption: $\sigma_1 = \dots \sigma_k = \sigma$

LDA Training and Testing

Given training data $(x_i, y_i), i = 1, \dots, n, y_i \in \{1, \dots, K\}$

1. Estimate sample mean and variance

$$\begin{aligned}\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_i \\ \hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2\end{aligned}$$

2. Estimate prior

$$\hat{\pi}_k = n_k/n.$$

Given testing point x , predict k that maximizes:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}.$$

Overview of LDA

A whiteboard with handwritten notes about Linear Discriminant Analysis (LDA). The notes include a data table, training equations, and prediction steps.

Data:

X	Y
5'2"	0
5'7"	1
6'2"	1
5'10"	0

1 → male Don't Know: Data μ_k, σ, π_k

Training:

$$L(X, Y) = \prod_{i=1}^N P(X_i, Y_i | \mu, \sigma)$$

Prediction:

We have: $\hat{\mu}_k, \hat{\sigma}, \hat{\pi}_k$

New X

For each X_i :
Predict Y_i

17

Description of White Board Image

In this class, we will focus on using LDA for *prediction*.

The estimands $\widehat{\mu}_k, \widehat{\sigma}_k^2, \widehat{\pi}_k$ are derived via maximum likelihood estimation, the “training” phase labeled on the white board.

For prediction, we use these estimands to determine which class has the highest probability, as shown in the bottom red box on slide 16.

LDA decision boundary

LDA decision boundary

Pick class k to maximize

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

Example: $k = 2, \pi_1 = \pi_2$

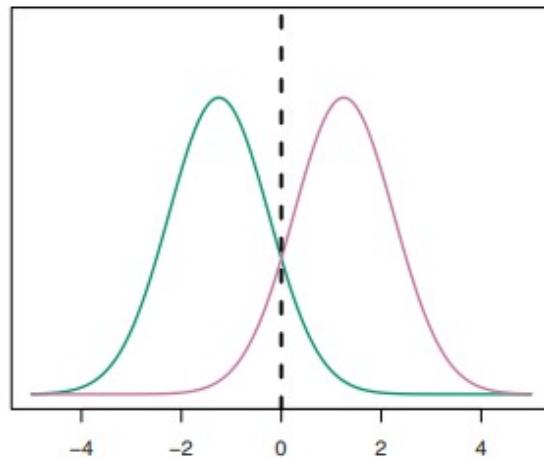
LDA decision boundary

Pick class k to maximize

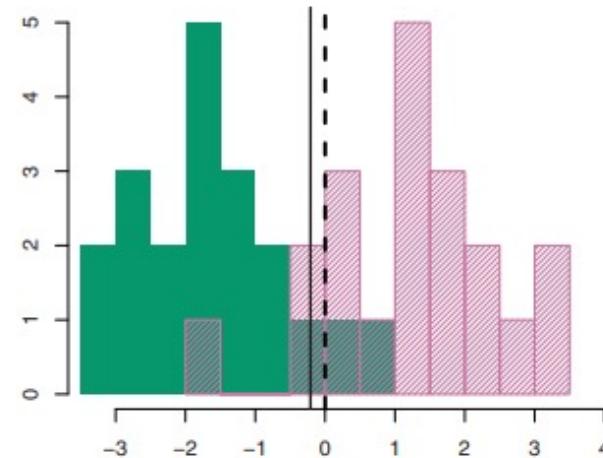
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Example: $k = 2, \pi_1 = \pi_2$

Classify as class 1 if $x > \frac{\mu_1 + \mu_2}{2}$



True decision boundary



Estimated decision boundary

LDA Training and Testing

Given training data $(x_i, y_i), i = 1, \dots, n, y_i \in \{1, \dots, K\}$

1. Estimate mean and variance

$$\begin{aligned}\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_i \\ \hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2\end{aligned}$$

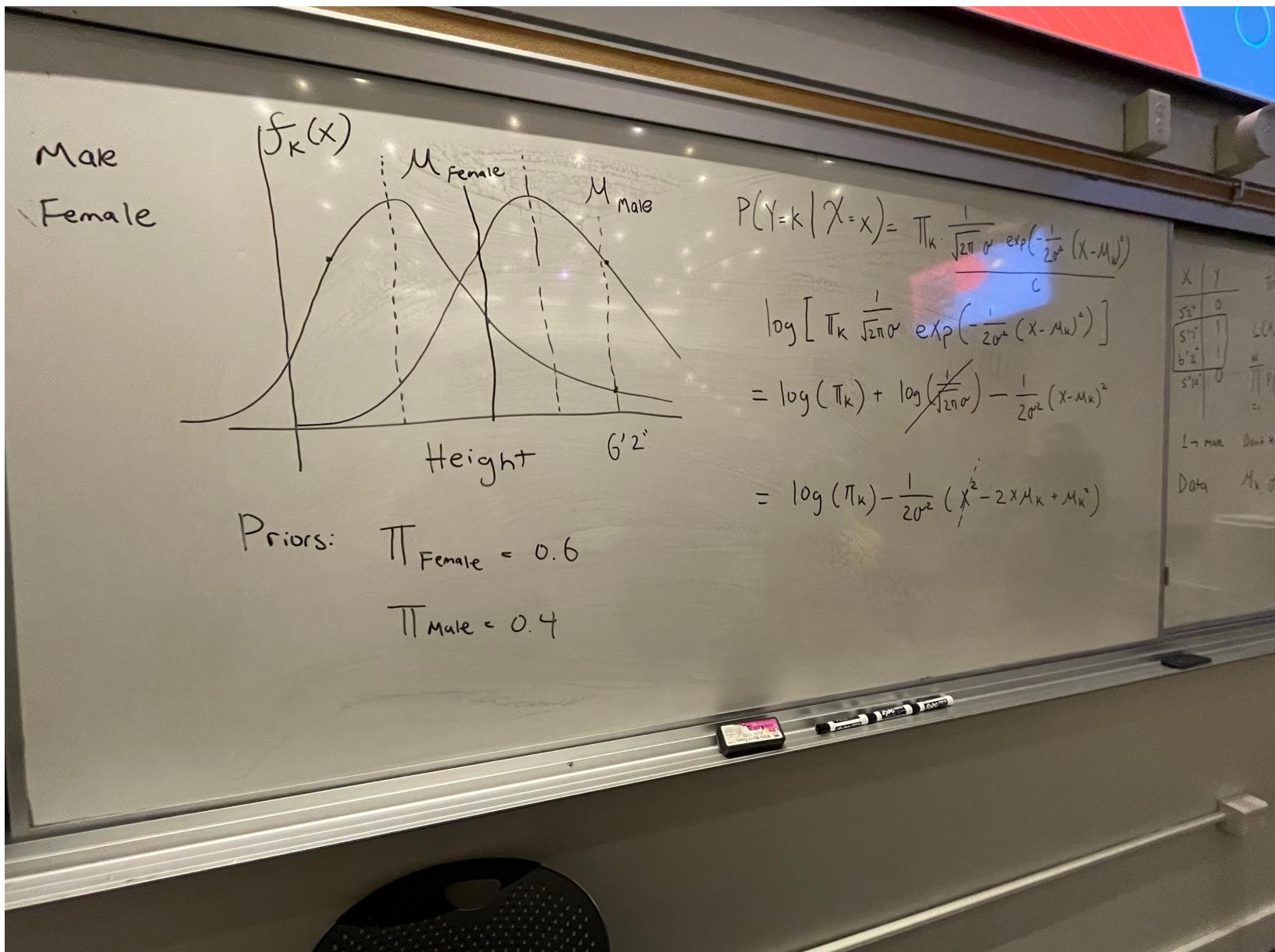
2. Estimate prior

$$\hat{\pi}_k = n_k/n.$$

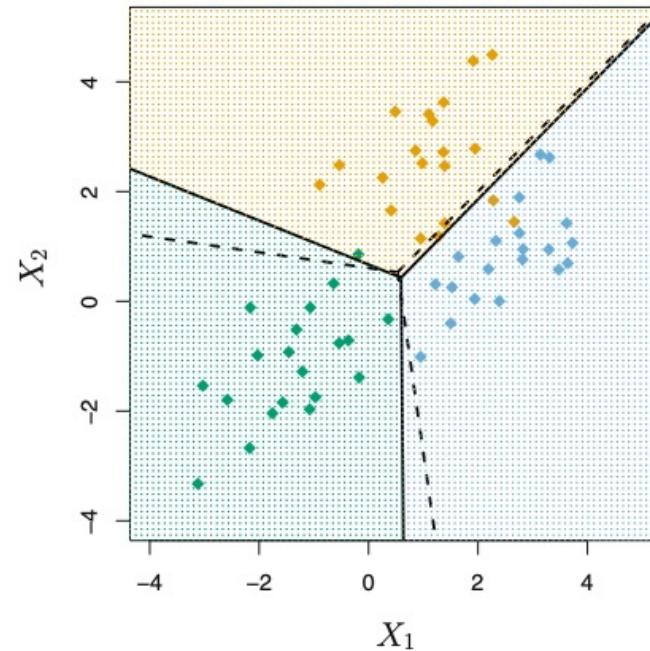
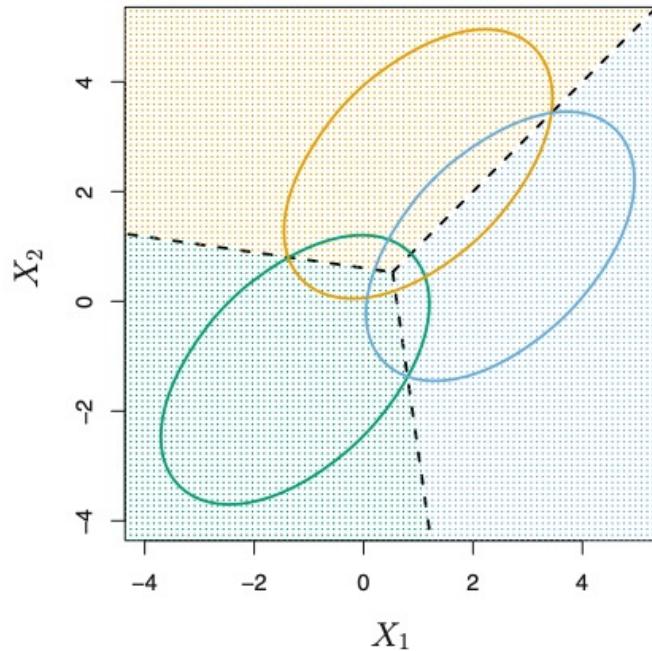
Given testing point x , predict k that maximizes:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

LDA Decision Boundary Derivation



Multi-Dimensional LDA

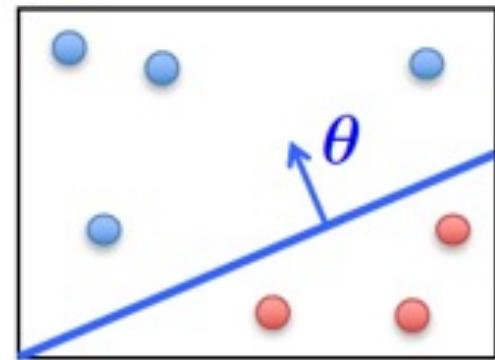


- LDA can be extended to multi-dimensional data
- Assumption that $f_k(x)$ is a multi-variate Gaussian

Linear models

- Logistic regression

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$



- LDA

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

LDA vs Logistic Regression

LDA vs Logistic Regression

- Logistic regression computes directly $\Pr[Y = 1|X = x]$ by assuming sigmoid function
 - Uses Maximum Likelihood Estimation
 - **Discriminative Model**
- LDA uses Bayes Theorem to estimate it
 - Estimates mean, co-variance, and prior from training data
 - **Generative model**
 - Assumes Gaussian distribution for $f_k(x) = \Pr[X = x|Y = k]$
- Which one is better?
 - LDA can be sensitive to outliers
 - LDA works well for Gaussian distribution
 - Logistic regression is more complex to solve, but more expressive