

DS 4400

Machine Learning and Data Mining I Spring 2022

Alina Oprea
Associate Professor
Khoury College of Computer Science
Northeastern University

February 28 2022

Announcements

- Please submit Homework 2 (will close site tonight)
- Homework 3 will be out later this week
- Project proposal is due on March 7
 - “Assignment” in Gradescope
- Midterm exam is on Wednesday, March 2
 - During class on Gradescope, 2:50-4:30pm
 - Topics: everything we covered in Lectures 1-10
 - What to bring: laptop, calculator, one cheat sheet
 - No access to lecture notes or Internet

Outline

- Generative classifiers
 - Difference from discriminative classifiers
- Linear Discriminant Analysis (LDA)
 - Training and inference
 - Why LDA is a linear classifier
 - Comparison with Logistic Regression
- Density Estimation and Naïve Bayes
 - Naïve Bayes assumption
 - Laplace smoothing

Generative vs Discriminative

- **Generative model (Module 4 on Canvas)**
 - 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
 - $P[Y = 1|X = x]$
 - 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

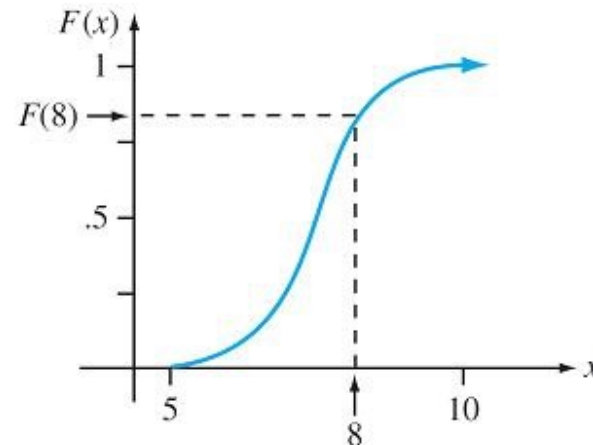
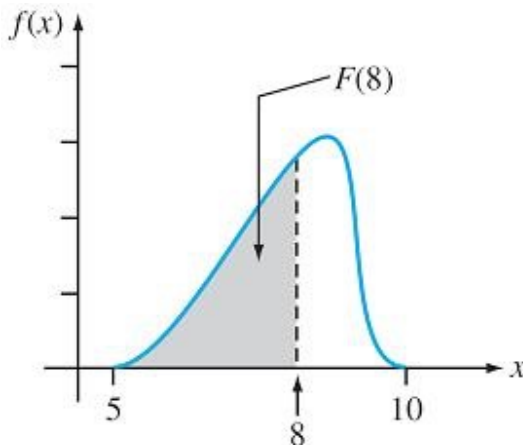
- $X:U \rightarrow V$ is continuous RV if it takes infinite number of values
- The **cumulative distribution function CDF** $F: R \rightarrow \{0,1\}$ for X is defined for every value x by:

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

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

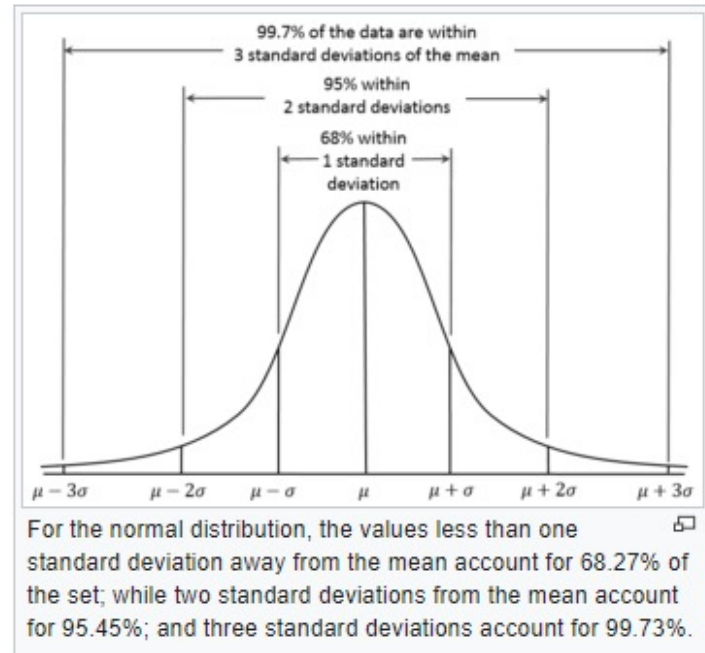
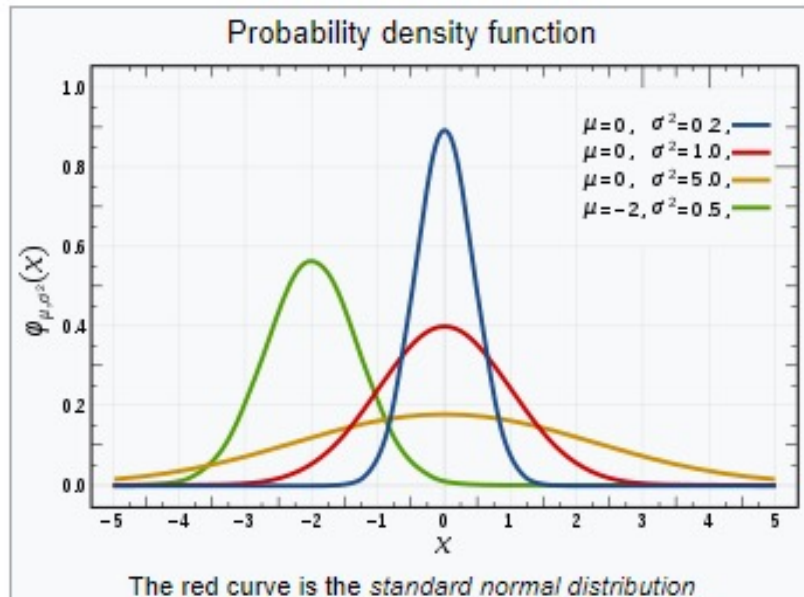
$$f(x) = dF(x)/dx$$

Increasing



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

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

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

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

2. Estimate prior

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

Given testing point x , predict k that maximizes:

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

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

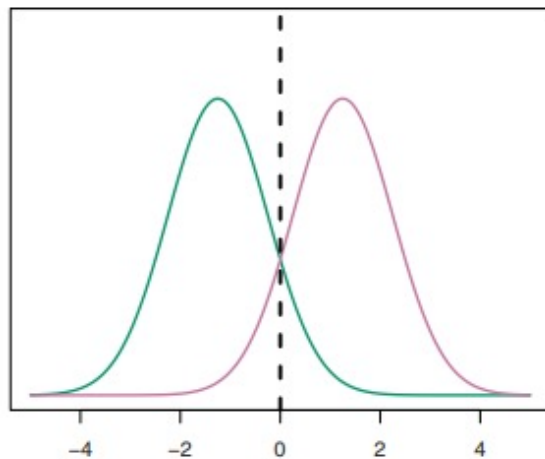
LDA decision boundary

Pick class k to maximize

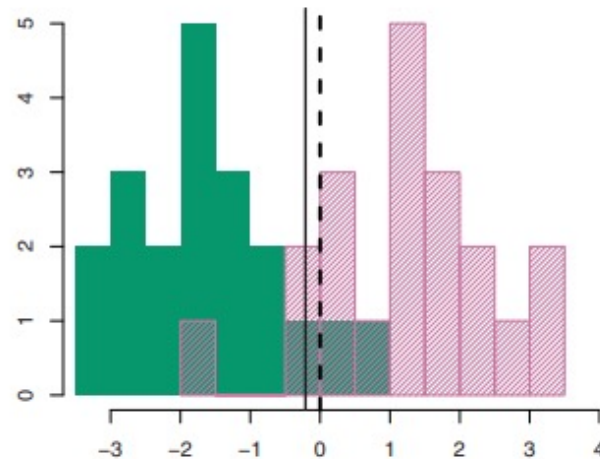
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

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

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

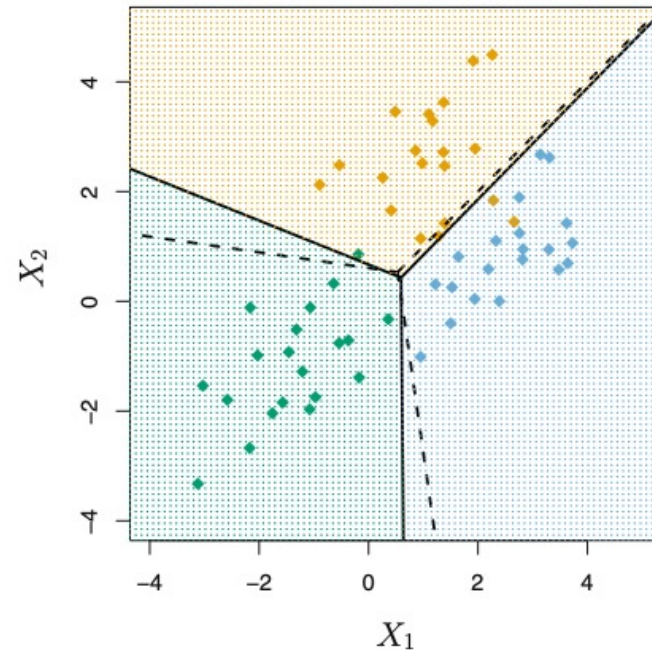
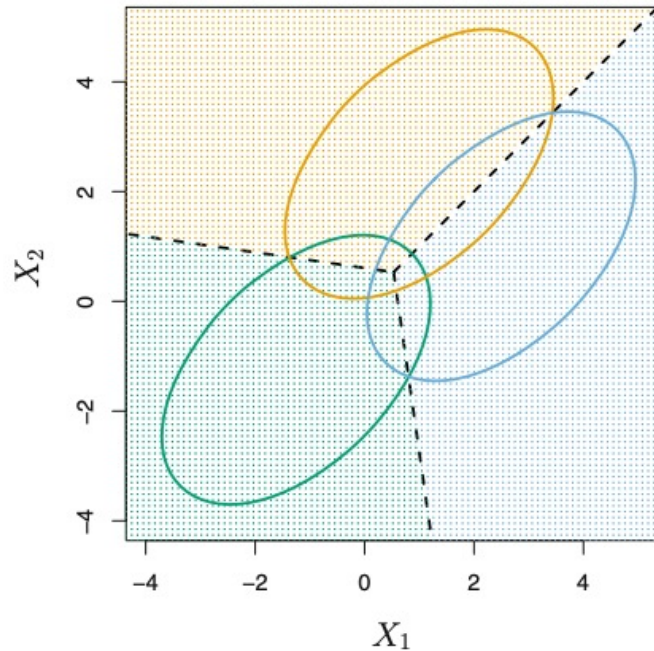
2. Estimate prior

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

Given testing point x , predict k that maximizes:

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

Multi-Dimensional LDA

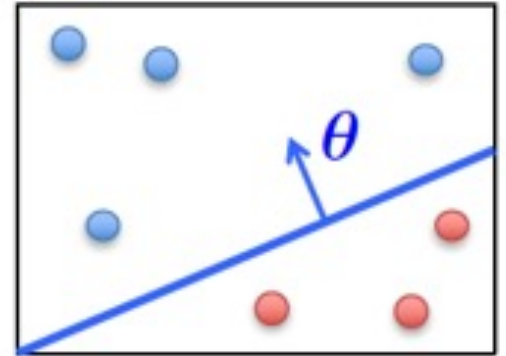


- LDA can be extended to multi-dimensional data
- Assumption that $f_k(x)$ is a multi-variate Gaussian

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

- 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