STA 35C: Statistical Data Science III

Lecture 11: Linear Discriminant Analysis

Dogyoon Song

Spring 2025, UC Davis

Announcement

Midterm 1 is in class on Fri, Apr 25 (12:10 pm - 1:00 pm)

- **Please plan to arrive early**: The exam will start at 12:10 pm and end at 1:00 pm sharp
- You may bring *one **hand-written** sheet of letter-sized paper (8.5 × 11 inches), double-sided* with formulas, brief notes, etc.
- **Calculator**: Simple (non-graphing) scientific calculators allowed
- **No textbooks** or other materials beyond the single cheat sheet
- **SDC accommodations:** Confirm scheduling with AES online

Resources for additional help & guidance

- Practice midterm posted on course webpage
- Discussion sections
- Office hours (Instructor: *Wed 4–6 pm*[1], TA: Thu 1–2 pm)
- Questions on Piazza

[1] Extended hours for this week only

Agenda

- Generative models for classification
  - Generative model approach
  - Linear discriminant analysis (LDA)
  - Overview of other models: QDA & Naive Bayes
- (If time permits) brief summary of what we have learned so far
  - Probability basics
  - Statistical learning
  - Regression
  - Classification

Discriminative vs. generative models

Discriminative: Directly model $\Pr(Y \mid X)$, e.g., logistic regression

Generative: Instead of modeling $\Pr(Y \mid X)$ directly, use Bayes' theorem:

$$\Pr(Y = k \mid X = x) = \frac{\Pr(Y = k,\ X = x)}{\Pr(X = x)} = \frac{\pi_k\, f_k(x)}{\sum_j \pi_j\, f_j(x)}$$

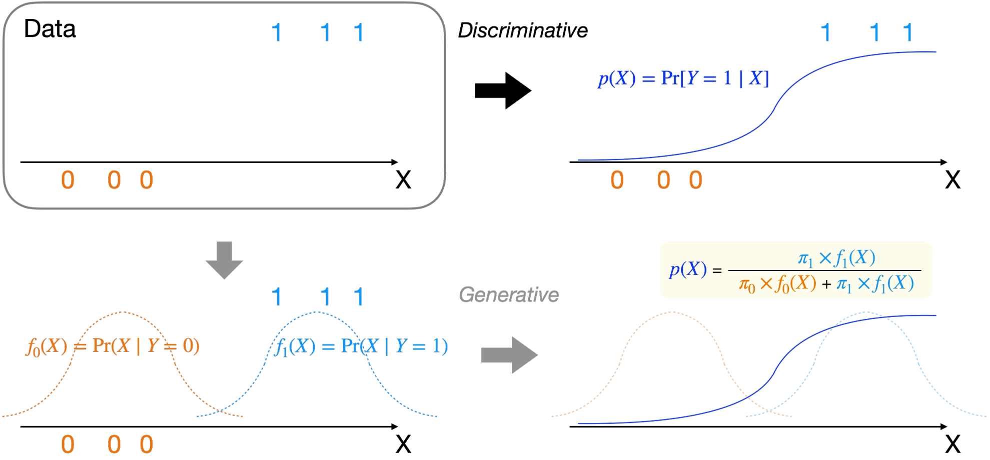
where

- $\pi_k := \Pr(Y = k)$: *Prior* probability that a randomly chosen observation is from the k-th class
- $f_k(X) := \Pr(X \mid Y = k)$[2]: Class-conditional *density* of X for observations from the k-th class

Key difference: Generative methods must model each $f_k(x)$, which can be demanding but can yield advantages if done correctly

[2] Strictly speaking, the equality holds only when X is discrete; if X is continuous, $f_k(x)$ gives density

Visualization of the workflow



$$p(X) = \Pr[Y = 1 \mid X]$$

$$f_0(X) = \Pr(X \mid Y = 0)$$

$$f_1(X) = \Pr(X \mid Y = 1)$$

$$p(X) = \frac{\pi_1 \times f_1(X)}{\pi_0 \times f_0(X) + \pi_1 \times f_1(X)}$$

Figure: A schematic contrast: discriminative approaches (**black**) directly learns $\Pr(Y|X)$, while generative (**gray**) models $\Pr(X|Y)$ and $\Pr(Y)$ first, then obtains $\Pr(Y|X)$ via Bayes.

LDA basics: The $p = 1$ case

Assumptions:

- For each class $k = 1, \ldots, K$, the predictor X is Gaussian with mean μ_k and **common** variance σ^2
- That is, $X \mid (Y = k) \sim \mathcal{N}(\mu_k, \sigma^2)$, so

$$f_k(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Big(-\frac{(x-\mu_k)^2}{2\sigma^2}\Big)$$

To compute $\Pr(Y = k \mid X = x)$, we estimate:

- $\pi_k = \Pr(Y = k)$
- $f_k(x) = \Pr(X = x \mid Y = k)$

and use

$$\Pr(Y = k \mid X = x) = \frac{\Pr(Y = k,\, X = x)}{\Pr(X = x)} = \frac{\pi_k\, f_k(x)}{\sum_{j=1}^{K} \pi_j\, f_j(x)}.$$

Question: But do we need to compute the entire $\Pr(Y = k \mid X = x)$?

Linear discriminant function

Our goal is **classification:** To classify a new x, we pick k maximizing

$$\Pr(Y = k \mid x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}$$

The denominator is constant for all $k \implies$ Comparing the numerators $\pi_k f_k(x)$ suffices

Taking log and rearranging terms:

$$\log\left(\pi_k f_k(x)\right) = \log \pi_k - \log\left(\sqrt{2\pi}\,\sigma\right) - \frac{(x-\mu_k)^2}{2\sigma^2}$$

$$= \underbrace{x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k}_{=:\text{Linear discriminant function}} + \underbrace{\left(-\log\left(\sqrt{2\pi}\,\sigma\right) - \tfrac{x^2}{2\sigma^2}\right)}_{\text{terms independent of } k}$$

LDA assumes common σ^2 across classes $k \implies$ "ignored" terms do not affect the choice

Parameter estimation in LDA

To compute the linear discriminant function

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

We need estimates of π_k, μ_k, σ: Given training data $\{(x_i, y_i)\}_{i=1}^n$,

$$\hat{\pi}_k = \frac{n_k}{n}, \quad \text{where } n_k = \#\{y_i = k\}$$

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

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

Extension to $p \geq 2$: The same idea applies, but we use p-dimensional mean vectors μ_k and $p \times p$ covariance matrix Σ, instead of scalars; see Lecture 10 (Slides 19–24)

LDA example

Suppose we have 7 observations as follows ($K = 2$, $p = 1$):

ID	x	Class
1	1	1
2	2	1
3	3	1
4	5	2
5	6	2
6	6	2
7	7	2

Task:

- **Classification:** given a new x, predict the class $y = k$ it likely belongs to
- For classification, we need to **compute discriminant functions** $\delta_1(x)$ and $\delta_2(x)$
- For this computation, we need to **estimate** π_1, π_2, μ_1, μ_2, and σ^2

LDA example: 1) Parameter estimation

1) Class priors: 3 observations belong to class 1, and 4 observations belong to class 2

$$\hat{\pi}_1 = \frac{3}{7}, \quad \hat{\pi}_2 = \frac{4}{7}$$

2) Class means: Sample mean for each class

$$\hat{\mu}_1 = \frac{1+2+3}{3} = 2, \quad \hat{\mu}_2 = \frac{5+6+6+7}{4} = 6$$

Common variance: Pooled sample covariance

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{2} \sum_{\substack{i: \\ y_i = k}} (x_i - \hat{\mu}_k)^2 = \frac{1}{5}(2+2) = 0.8$$

where (1) $n - K = 5$ as $n = 7$ and $K = 2$ and (2) the sum of squared deviations for each class is given by

- Class 1: $(1-2)^2 + (2-2)^2 + (3-2)^2 = 1 + 0 + 1 = 2$
- Class 2: $(5-6)^2 + (6-6)^2 + (6-6)^2 + (7-6)^2 = 1 + 0 + 0 + 1 = 2$

LDA example: 2) Computing discriminants

Discriminant functions: for $k = 1, 2$, we plug in estimates

$$\delta_k(x) = x\,\frac{\hat{\mu}_k}{\hat{\sigma}^2} \;-\; \frac{\hat{\mu}_k^2}{2\,\hat{\sigma}^2} \;+\; \log \hat{\pi}_k$$

Inserting

- $\hat{\pi}_1 = \frac{3}{7}, \hat{\pi}_2 = \frac{4}{7}$ and
- $\hat{\mu}_1 = 2$, $\hat{\mu}_2 = 6$, and $\hat{\sigma}^2 = 0.8$

yields

$$\begin{aligned} \delta_1(x) &= x\,\frac{\hat{\mu}_1}{0.8} \;-\; \frac{\hat{\mu}_1^2}{2 \times 0.8} \;+\; \log\left(\tfrac{3}{7}\right) \\ \delta_2(x) &= x\,\frac{\hat{\mu}_2}{0.8} \;-\; \frac{\hat{\mu}_2^2}{2 \times 0.8} \;+\; \log\left(\tfrac{4}{7}\right) \end{aligned}$$

LDA example: 3) Classification and decision boundary

Decision rule: For a given x, we predict

$$\hat{Y} = \begin{cases} 1 & \text{if } \delta_1(x) \geq \delta_2(x), \\ 2 & \text{if } \delta_1(x) < \delta_2(x) \end{cases}$$

Decision boundary: The decision rule above predicts $\hat{Y} = 1$ if and only if

$$\begin{aligned} \delta_1(x) \geq \delta_2(x) \quad & \iff \quad x\,\frac{\hat{\mu}_1}{0.8} - \frac{\hat{\mu}_1^2}{2 \times 0.8} + \log\left(\tfrac{3}{7}\right) \geq x\,\frac{\hat{\mu}_2}{0.8} - \frac{\hat{\mu}_2^2}{2 \times 0.8} + \log\left(\tfrac{4}{7}\right) \\ & \iff \quad \frac{\hat{\mu}_1 - \hat{\mu}_2}{0.8}\,x - \frac{\hat{\mu}_1^2 - \hat{\mu}_2^2}{2 \times 0.8} + \left(\log\left(\tfrac{3}{7}\right) - \log\left(\tfrac{4}{7}\right)\right) \geq 0 \end{aligned}$$

Here, $\hat{\mu}_1 - \hat{\mu}_2 < 0$, and hence, we can simplify this to

$$x - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} + \frac{0.8}{\hat{\mu}_1 - \hat{\mu}_2} \log\left(\tfrac{3}{4}\right) \leq 0 \quad \iff x \leq \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} - \frac{0.8}{\hat{\mu}_2 - \hat{\mu}_1} \log\left(\frac{4}{3}\right)$$

LDA visualization (1D)

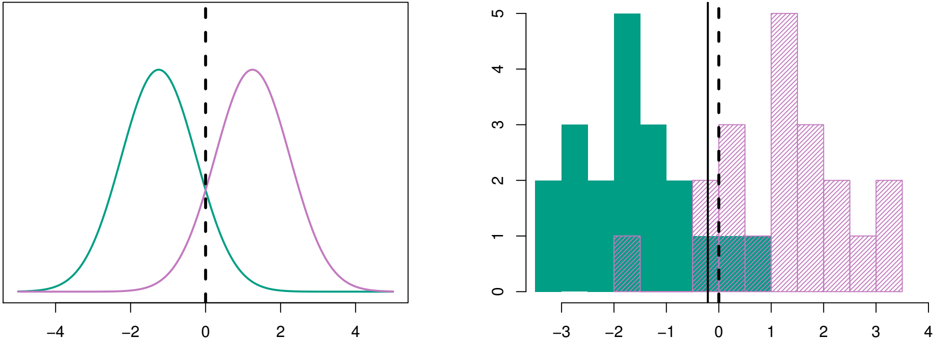


Figure: **(Left)** Two one-dimensional normal density functions. The dashed vertical line is the Bayes decision boundary. **Right)** Histograms of 20 observations from each class. The dashed vertical line again shows the Bayes decision boundary, while the solid vertical line represents the LDA decision boundary estimated from the training data [JWHT21, Figure 4.4].

LDA visualization (2D)

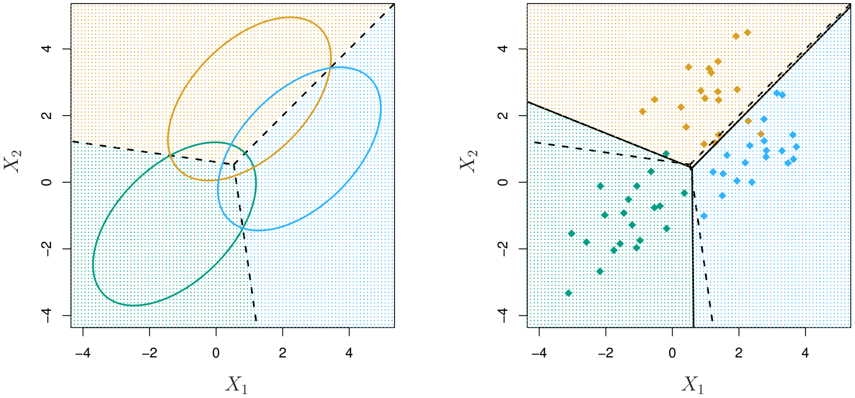


Figure: An illustration of LDA decision boundaries. Observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, a class-specific mean vector, and a common covariance matrix. **(Left)** Ellipses indicating the 95% probability region for each of the three classes, with dashed lines showing the Bayes decision boundaries. **(Right)** 20 observations from each class, and the corresponding LDA decision boundaries (solid black lines) [JWHT21, Figure 4.6].

Additional generative approaches

Different modeling assumptions on classwise density f_k lead to different methods.

Quadratic Discriminant Analysis (QDA)

- $X \mid Y = k$ is Gaussian with mean μ_k and *possibly different* Σ_k
- The discriminant function is no longer linear in x:

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k) \; - \; \frac{1}{2}\log|\Sigma_k| \; + \; \log \pi_k.$$

- QDA is more flexible but requires estimating more parameters

Naive Bayes

- For high-dimensional or discrete X, specifying $f_k(x)$ can be difficult
- *Naive* assumption: predictors X_j are conditionally independent given $Y = k$
  $\implies \; f_k(x) = \prod_{j=1}^{p} f_{k,j}(x_j)$
- Typically fast and often effective, but the independence assumption may not be valid

Visual comparison between LDA and QDA

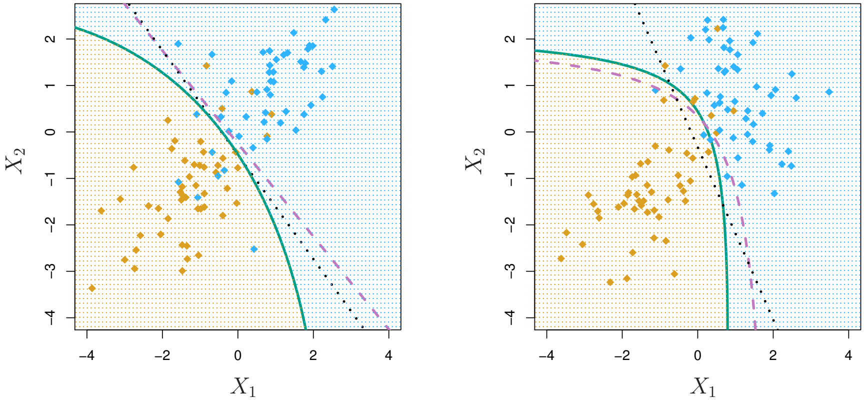


Figure: A comparison of LDA and QDA decision boundaries. The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) boundaries are shown. LDA always produces a linear boundary, whereas QDA can be curved [JWHT21, Figure 4.6].

Wrap-up

Recap of classification:

- Classification problem: setup, objectives
- Logistic regression
  - Model & interpretation of regression coefficients
  - Parameter estimation via MLE
  - Extensions: multiple predictors, multinomial
  - Decision boundary
- Linear Discriminant Analysis
  - Classification via a generative model
  - Discriminant function
- Classification error
  - Confusion matrix
  - Choice of decision threshold

References

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.
Springer, New York, NY, 2nd edition, 2021.