STA 35C: Statistical Data Science III

Lecture 10: Generative Models for Classification

Dogyoon Song

Spring 2025, UC Davis

Agenda

Last time: Logistic regression & classification assessment

Today:

- **Generative vs. discriminative models**
  - Why generative modeling?
- **Linear discriminant analysis (LDA)**
  - Basics: $p = 1$ then general $p \geq 1$
  - A concrete example ($p = 2$)
  - Parameter estimation
- **Naive Bayes**
  - Conditional independence assumption
  - Parameter estimation & example usage

Discriminative vs. Generative Models

Discriminative (e.g. logistic regression):

- Directly model $\Pr(Y \mid X)$, e.g., using a linear function
- Find a decision boundary in X-space that separates classes

Generative (e.g. LDA, Naive Bayes):

- Instead of modeling $\Pr(Y \mid X)$ directly, model:
  - The *prior* probability $\pi_k := \Pr(Y = k)$ that a randomly chosen observation comes from the k-th class
  - The class-conditional *density function* $f_k(X) := \Pr(X \mid Y = k)$[1] of X for an observation that comes from the k-th class
- Then use Bayes' theorem to compute the *posterior probability*:

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

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

Visualization of the workflow

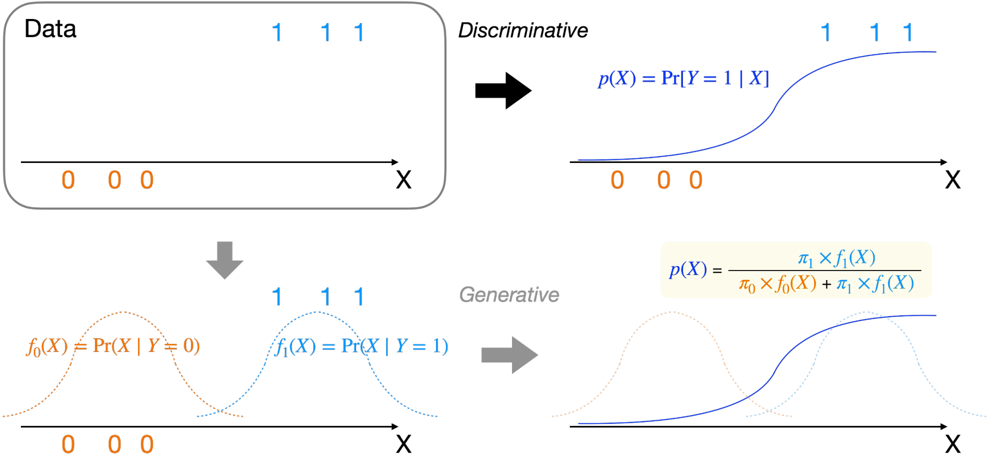

Figure: A schematic contrast: discriminative approaches (**black**) directly learns $\Pr(Y|X)$, while generative (**gray**) models $\Pr(X|Y)$ and $\Pr(Y)$ first, then obtains $\Pr(Y|X)$ via Bayes.

Contrasting the two approaches

Both aim to estimate $\Pr(Y \mid X)$, but:

Discriminative workflow:

- Postulate a functional form for $\Pr(Y = 1 \mid X)$
- Fit parameters from data
- Directly output $p(x) = \Pr(Y = 1|x)$

Generative workflow:

- Postulate each class distribution $f_k(x)$
  - Key challenge: specifying X's distribution per class
- Estimate $\pi_k = P(Y = k)$ (often just the proportion in class k)
- Compute $p(x) = \Pr(Y = k \mid x)$ via Bayes' theorem

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

Why Generative Models?

Upsides:

- **Well-separated classes**: discriminative approaches (e.g., logistic regression) may become unstable, while generative can be more robust
- **If model assumption is correct**: fewer data are needed for good performance
- **K-class extension**: straightforward via Bayes

Downsides:

- Must specify $f_k(x)$: can be difficult in high dimensions ($p \gg 1$)
- If assumptions fail, performance may degrade

Pop-up Quiz #1: Generative vs. discriminative

Question: Which statement best describes a key advantage of a generative model (like LDA) over a discriminative one (like logistic regression)?

A) Generative models need *no* distributional assumptions on X.
B) Discriminative models cannot be extended to $K > 2$ classes.
C) If the assumed $f_k(x)$ is correct, generative models can be data-efficient.
D) Generative models ignore class priors π_k.

LDA Basics: The $p = 1$ Case

Assumptions:

- $Y \in \{1, \ldots, K\}$ classes, and $\pi_k = \Pr[Y = k]$
- $X \mid (Y = k) \sim \mathcal{N}(\mu_k, \sigma^2)$, with same σ^2 for all k

- Then the class-conditional density is

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

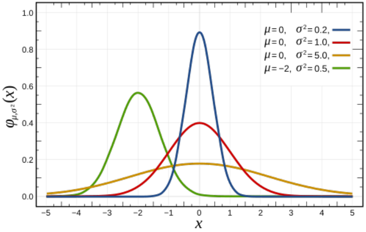

Figure: PDF of 1D Gaussian distribution (Image from Wikipedia[a]).

[a] https://en.wikipedia.org/wiki/Normal_distribution

Decision boundary for $p = 1$

By Bayes' theorem:

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

Bayes classifier: choose k maximizing $\Pr(Y = k \mid x)$

- We compare $\log\left(\pi_k f_k(x)\right)$ to find maximizing k

Linear discriminant function: When σ^2 is common across classes, the *quadratic* terms cancel, leaving a *linear* function:

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

We classify to the k with largest $\delta_k(x)$; the boundary between k and j is *linear* in x

From $p = 1$ to $p \geq 1$

General assumption:

- $X \in \mathbb{R}^p$ and $X \mid (Y = k) \sim \mathcal{N}(\mu_k, \Sigma)$
- Common covariance Σ, distinct μ_k
- $\pi_k = P(Y = k)$

Class-conditional density:

$$f_k(x) = \frac{1}{(2\pi)^{p/2} \, |\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x - \mu_k)^\top \Sigma^{-1}(x - \mu_k)\Big).$$

Discriminant function:

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

Boundary between classes k and j is linear in x.

Visualization of a multivariate Gaussian

Goal: Illustrate shape of $\mathcal{N}(\mu_k, \Sigma)$ in 2D

- Elliptical contours, reflecting Σ
- If Σ is the same for both classes, the ratio of densities is linear in x

e.g. if Σ has correlation terms, the ellipses tilt

Parameter Estimation in LDA

Given training data $\{(x_i, y_i)\}_{i=1}^n$:

- $\hat{\pi}_k = \frac{n_k}{n}$, $n_k = \#\{y_i = k\}$
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i:\, y_i=k} x_i$
- $\hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^K \sum_{i:\, y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top$

Then

$$\hat{\delta}_k(x) = x^\top \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^\top \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k,$$

and predict $\arg\max_k \hat{\delta}_k(x)$.

Concrete Example ($p = 2$, $K = 2$)

Scenario: Suppose $K = 2$ classes, $X \in \mathbb{R}^2$. We gather 8 total points:

User	X_1	X_2	Class
1	1.2	2.5	1
2	1.8	2.9	1
3	2.2	3.2	1
4	3.0	4.0	1
5	3.5	4.2	2
6	4.0	5.0	2
7	4.3	5.2	2
8	4.5	5.6	2

- We'll estimate π_1, π_2, μ_1, μ_2, and a *common* Σ.
- Then see how $\delta_1(x)$ vs. $\delta_2(x)$ forms a linear boundary in $\mathbb{R}^2$.

Concrete Example: Parameter estimation

Class priors:

$$\hat{\pi}_1 = \frac{4}{8}, \quad \hat{\pi}_2 = \frac{4}{8}.$$

Means:

$$\hat{\mu}_1 = \begin{bmatrix} \bar{x}_{1,1} \\ \bar{x}_{1,2} \end{bmatrix}, \quad \hat{\mu}_2 = \begin{bmatrix} \bar{x}_{2,1} \\ \bar{x}_{2,2} \end{bmatrix}.$$

Covariance:

$$\hat{\Sigma} = \frac{1}{8-2} \sum_{k=1}^{2} \sum_{i \in \text{class } k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top.$$

Compute numerically (in practice, one might use R).

Concrete Example: Decision boundary

Discriminant functions:

$$\hat{\delta}_1(x) = x^\top \hat{\Sigma}^{-1}\hat{\mu}_1 - \frac{1}{2}\hat{\mu}_1^\top \hat{\Sigma}^{-1}\hat{\mu}_1 + \log \hat{\pi}_1,$$

$$\hat{\delta}_2(x) = x^\top \hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_2^\top \hat{\Sigma}^{-1}\hat{\mu}_2 + \log \hat{\pi}_2.$$

The boundary is where $\hat{\delta}_1(x) = \hat{\delta}_2(x)$, which rearranges to a linear equation in x_1, x_2.

Hence:

$$\{\mathbf{x} : \hat{\delta}_1(\mathbf{x}) = \hat{\delta}_2(\mathbf{x})\} \quad \Longleftrightarrow \quad (\text{some linear function of } x_1, x_2) = 0.$$

A straight line in $\mathbb{R}^2$ dividing class 1 and class 2.

Extension to Quadratic discriminant analysis (QDA)

- If each class k has *its own* covariance Σ_k, then the log-ratio remains *quadratic* in x
- This yields *quadratic* decision boundaries
- QDA is more flexible but requires estimating more parameters

Pop-up quiz #2: LDA boundaries

Question: In LDA with $p = 2$ and $K = 2$ classes, why is the decision boundary *always* linear?

A) Each class has its own covariance matrix, forcing a hyperplane boundary.
B) We assume the same Σ, so the quadratic parts cancel in the log ratio.
C) $p = 2$ is too small to allow curved boundaries.
D) LDA only applies to data that are linear in X.

Naive Bayes: Another Generative Approach

Motivation:

- For high-dimensional or discrete X, specifying $f_k(x)$ is difficult
- *Naive* assumption: features X_j are conditionally independent given $Y = k$
- Then $f_k(x) = \prod_{j=1}^{p} f_{k,j}(x_j)$

Result:

$$\Pr(Y = k \mid x) \;\propto\; \hat{\pi}_k \prod_{j=1}^{p} \hat{f}_{k,j}(x_j).$$

- Popular in text classification (bag-of-words)
- Effective if feature independence is not *too* violated

Parameter Estimation in Naive Bayes

Discrete features:

- Estimate $\hat{P}(X_j = a \mid Y = k)$ from training frequencies
- e.g. text classification: count how often word w_j appears in each class k

Continuous features:

- Often assume $X_j \mid (Y = k)$ is Gaussian $\rightarrow$ estimate $\hat{\mu}_{k,j}, \hat{\sigma}_{k,j}$

Putting it all together:

$$\Pr(Y = k \mid x) \propto \hat{\pi}_k \prod_{j=1}^{p} \hat{f}_{k,j}(x_j).$$

Predict: $\arg\max_k \Pr(Y = k \mid x)$.

Wrap-up

Generative models:

- We model $P(X \mid Y)$ & $P(Y)$, then use Bayes to get $P(Y \mid X)$
- If assumptions hold, can be data-efficient

LDA:

- Gaussian class-conditional with common Σ
- Linear boundaries
- Detailed example: $p = 1$ and $p = 2$

Naive Bayes:

- Factorizes $f_k(x)$ under conditional independence
- Reduces complexity, helpful in high-dimensional or discrete settings

References