

STA 35C: Statistical Data Science III

Lecture 10: Generative Models for Classification

Dogyoon Song

Fall 2026, UC Davis

Announcements

Homework 3 is due **Thu, Oct 15, 11:59 PM PT**

- Please submit on time and follow the submission instructions
- You can collaborate on Homework, but all submitted work must be your own (also, don't forget to list all collaborators)

Lectures on **Wed, Oct 21** and **Fri, Oct 23** will be held via Zoom; details and the Zoom link will be posted on Canvas

Agenda

Last time:

- Extensions of logistic regression
- Assessing classification performance

Today:

- More on classification performance and the ROC curve
- Generative vs. discriminative classification
- Linear discriminant analysis (LDA)

Recap: Logistic regression and decision boundary

Logistic regression: Linear model for log-odds with more predictors

$$\log \left(\frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \quad p(x) = \Pr(Y = 1 \mid X = x)$$

- The log-odds (=logit) is modeled using a linear function in $X_1, \dots, X_p$
- Multinomial ($K \geq 3$): Choose one class as a baseline, and model the $(K - 1)$ log-odds relative to that baseline

Decision rule ($K = 2, 0 < p^* < 1$): Predict $\hat{y} = 1$ if and only if

$$\begin{aligned} \hat{p}(x) \geq p^* &\iff \log \left(\frac{\hat{p}(x)}{1 - \hat{p}(x)} \right) \geq \log \left(\frac{p^*}{1 - p^*} \right) \\ &\iff \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \geq \log \left(\frac{p^*}{1 - p^*} \right) \end{aligned}$$

- With two predictors and at least one nonzero slope, the decision boundary is a line:

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = \log \left(\frac{p^*}{1 - p^*} \right)$$

Recap: Confusion matrix, false positive, and false negative

Confusion matrix: A performance evaluation table summarizing outcomes

		<i>True default status</i>		
		No	Yes	Total
<i>Predicted default status</i>	No	9432	138	9570
	Yes	235	195	430
Total		9667	333	10000

Figure: An example confusion matrix for the `Default` dataset [JWHT21, Table 4.5].

- **Error rate:** Among all observations, how many are misclassified?
- **False positive rate (FPR):** Among actual negatives, how many are falsely claimed positive?
- **False negative rate (FNR):** Among actual positives, how many are missed?

$$\text{Error rate} = \frac{FP + FN}{TP + TN + FP + FN} = \frac{235 + 138}{10000},$$

$$\text{FPR} = \frac{FP}{TN + FP} = \frac{235}{9667} \approx 0.024,$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{195 + 9432}{10000},$$

$$\text{FNR} = \frac{FN}{TP + FN} = \frac{138}{333} \approx 0.414.$$

Recap: The ROC curve and selecting threshold p^*

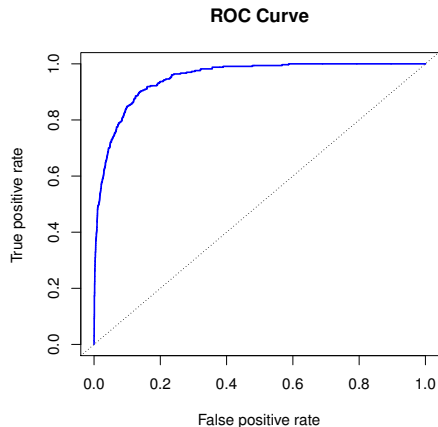


Figure: An example ROC curve, with AUC [JWHT21, Figure 4.8].

The ROC curve and AUC

- Plot TPR vs. FPR as p^* moves $1 \rightarrow 0$
 - $\text{TPR} = \frac{TP}{P} = \frac{TP}{TP+FN} = 1 - \text{FNR}$
 - $\text{FPR} = \frac{FP}{N} = \frac{FP}{TN+FP}$
- Curves closer to the upper-left corner indicate better classification performance
- **Area under curve (AUC):** Reflects overall discriminative power across thresholds
 - Perfect classifier: $\text{AUC} = 1$
 - Random guess: $\text{AUC} = 0.5$

How to choose p^* ?

- Choose p^* based on the relative costs of false positives and false negatives
- e.g., maximize TPR with FPR constrained

Pop-up quiz: Choosing p^*

Scenario: A hypothetical screening model outputs $\hat{p}(x) = \Pr(Y = 1 \mid x)$, where $Y = 1$ means “disease present.” Missing a true case is much more costly than a false alarm.

Question: Which threshold choice is most sensible?

- A) Use a larger threshold such as $p^* = 0.9$, to reduce false positives as much as possible.
- B) Use a smaller threshold such as $p^* = 0.2$, to catch more true positives even if false positives increase.
- C) Always use $p^* = 0.5$; changing the threshold only changes the reported probabilities, not the predictions.
- D) The threshold does not matter if the ROC curve has high AUC.

Answer: B. When false negatives are especially costly, we usually lower the threshold to increase sensitivity, accepting more false positives.

Discriminative vs. generative models

Discriminative (e.g. logistic regression): Model the posterior directly

$$\Pr(Y = k \mid X = x)$$

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

- The *prior* $\pi_k := \Pr(Y = k)$ that a randomly chosen observation is from the k -th class
- The class-conditional *PMF or density* $f_k(X) := \Pr(X \mid Y = k)$ ¹ of X within class k

Then use Bayes' theorem to compute the *posterior probability*:

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

Classification rule (equal misclassification costs):

$$\hat{y}(x) = \arg \max_k \Pr(Y = k \mid X = x) \quad \Longleftrightarrow \quad \hat{y}(x) = \arg \max_k \pi_k f_k(x).$$

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

Schematic visualization: Discriminative vs. generative

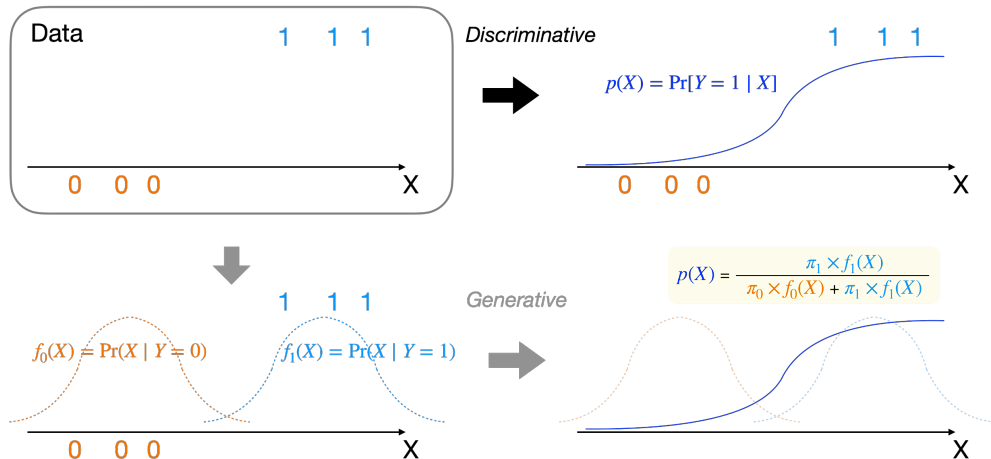


Figure: A schematic contrast: discriminative approaches (**black**) directly model $\Pr(Y | X)$, while generative approaches (**gray**) model $\Pr(X | Y)$ and $\Pr(Y)$ first, then obtain $\Pr(Y | X)$ via Bayes' rule.

Contrasting the two approaches

Both aim to estimate $\Pr(Y | X)$, but they get there differently

Discriminative

- **Procedure:** specify a model for $\Pr(Y = k | X = x)$ (or the log-odds) and fit
- **Strengths:** fewer assumptions on X ; often strong for direct prediction
- **Limitations:** does not model classwise predictor distributions; can be less data-efficient when a simple generative model is correct

Generative

- **Procedure:** model $\pi_k = \Pr(Y = k)$ and $f_k(x) = p(x | Y = k)$, apply Bayes' rule
- **Strengths:** can be data-efficient if the assumptions are good; naturally incorporates priors and class structure
- **Limitations:** stronger distributional assumptions; performance can suffer if $f_k(x)$ is misspecified

Key difference: Discriminative methods focus on the decision boundary; generative methods model how each class generates X

Pop-up quiz: Generative vs. discriminative

Question: Which quantity must a generative classifier such as LDA model explicitly, beyond the class prior π_k ?

- A) The class-conditional distribution $f_k(x) = p(x \mid Y = k)$
- B) The R^2 of the classifier
- C) Only the decision threshold p^*
- D) The posterior $\Pr(Y = k \mid X = x)$ directly, with no assumptions on X

Answer: A. Generative classifiers model π_k and $f_k(x)$, then obtain the posterior by Bayes' rule.

Linear discriminant analysis (LDA)

Suppose the response has two classes, $Y \in \{0, 1\}$, and one predictor $X \in \mathbb{R}$.

LDA assumptions:

- $\pi_0 = \Pr(Y = 0)$, $\pi_1 = \Pr(Y = 1)$ are the class priors, with $\pi_0 + \pi_1 = 1$
- $X \mid (Y = 0) \sim \mathcal{N}(\mu_0, \sigma^2)$
- $X \mid (Y = 1) \sim \mathcal{N}(\mu_1, \sigma^2)$
- The two classes may have different means, but **share the same positive variance** σ^2

So the class-conditional densities are

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma^2}\right), \quad k \in \{0, 1\}.$$

Using Bayes' rule,

$$\Pr(Y = 1 \mid X = x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}.$$

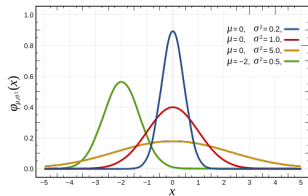


Figure: PDF of 1D Gaussian distribution (Image from [Wikipedia](#)^a).

LDA in 1D: Prediction and decision boundary (1/2)

To classify a new observation x , compare the two posterior scores:

$$\Pr(Y = 1 \mid X = x) \quad \text{vs.} \quad \Pr(Y = 0 \mid X = x).$$

Since the denominator $(\pi_0 f_0(x) + \pi_1 f_1(x))$ is the same, we predict

$$\hat{y} = 1 \quad \text{if and only if} \quad \pi_1 f_1(x) \geq \pi_0 f_0(x).$$

Taking logs of $\pi_k f_k(x)$

$$\begin{aligned} \log(\pi_k f_k(x)) &= \log \pi_k - \log(\sqrt{2\pi}\sigma) - \frac{(x-\mu_k)^2}{2\sigma^2} \\ &= \underbrace{x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k}_{\text{=:Linear discriminant function}} - \underbrace{\log(\sqrt{2\pi}\sigma) - \frac{x^2}{2\sigma^2}}_{\text{we can ignore these}} \end{aligned}$$

LDA in 1D: Prediction and decision boundary (2/2)

Ignoring constants in $\log(\pi_k f_k(x))$, we define the **linear discriminant function**

$$\delta_k(x) := x \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k, \quad k \in \{0, 1\}.$$

We predict class 1 when $\delta_1(x) \geq \delta_0(x)$

The decision boundary in 1D is a single cutoff point:

$$x^* = \frac{\mu_0 + \mu_1}{2} - \frac{\sigma^2}{\mu_1 - \mu_0} \log\left(\frac{\pi_1}{\pi_0}\right) \quad (\text{when } \mu_1 > \mu_0).$$

- If $\pi_0 = \pi_1$, then $x^* = \frac{\mu_0 + \mu_1}{2}$, the midpoint of the two means
- If $\pi_1 > \pi_0$, the cutoff shifts toward class 0, so class 1 is predicted more often

LDA visualization (1D)

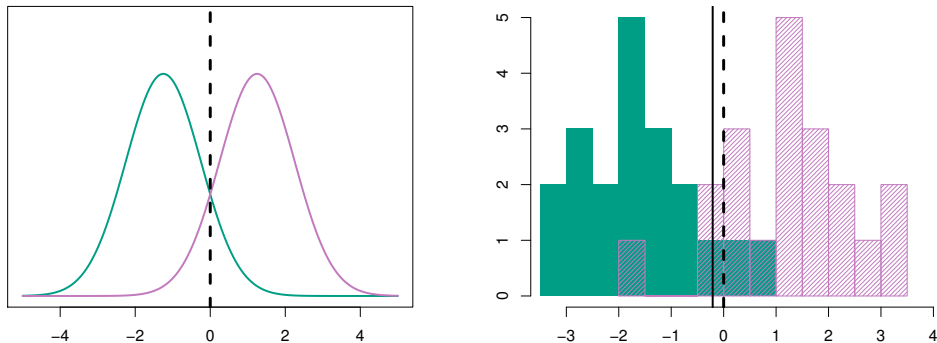


Figure: **(Left)** Two one-dimensional normal density functions. The dashed vertical line is the Bayes decision boundary. **(Right)** Histograms of 20 observations from each class. The dashed vertical line again shows the Bayes decision boundary, while the solid vertical line represents the LDA decision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

LDA in 1D: Estimating the LDA parameters

Given independent training data $\{(x_i, y_i)\}_{i=1}^n$ with both classes present, estimate:

Class priors:

$$\hat{\pi}_k = \frac{n_k}{n}, \quad n_k = \#\{i : y_i = k\}, \quad k \in \{0, 1\}.$$

Class means:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i, \quad k \in \{0, 1\}.$$

Common variance (pooled estimate):

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{k \in \{0,1\}} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2.$$

Then plug $\hat{\pi}_k, \hat{\mu}_k, \hat{\sigma}^2$ into the cutoff formula or the posterior formula.

Illustration of LDA: A worked example

Example

Suppose we have the following training dataset ($K = 2$, $p = 1$):

ID	1	2	3	4	5	6	7
Predictor x	1	2	3	4	5	6	7
Class y	0	0	0	1	1	1	1

In this example, we will:

- Estimate $\hat{\pi}_0, \hat{\pi}_1, \hat{\mu}_0, \hat{\mu}_1, \hat{\sigma}^2$
- Form the fitted discriminant functions $\delta_0(x)$ and $\delta_1(x)$
- Compute the decision boundary and classify new observations with x

Illustration of LDA: 1) Parameter estimation

Example

Class priors: 3 observations belong to class 0, and 4 observations belong to class 1

$$\hat{\pi}_0 = \frac{3}{7}, \quad \hat{\pi}_1 = \frac{4}{7}$$

Class means: Sample mean for each class

$$\hat{\mu}_0 = \frac{1 + 2 + 3}{3} = 2, \quad \hat{\mu}_1 = \frac{4 + 5 + 6 + 7}{4} = 5.5$$

Common variance (pooled): Pooled sample variance estimator

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=0}^1 \sum_{\substack{i: \\ y_i=k}} (x_i - \hat{\mu}_k)^2 = \frac{1}{5} (2 + 5) = 1.4$$

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 0: $(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2 = 1 + 0 + 1 = 2$
- Class 1: $(4 - 5.5)^2 + (5 - 5.5)^2 + (6 - 5.5)^2 + (7 - 5.5)^2 = 2.25 + 0.25 + 0.25 + 2.25 = 5$

Illustration of LDA: 2) Discriminants and classification

Example

Recall the fitted discriminant functions are

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

Plugging in $\hat{\mu}_0 = 2$, $\hat{\mu}_1 = 5.5$, $\hat{\sigma}^2 = 1.4$, $\hat{\pi}_0 = 3/7$, $\hat{\pi}_1 = 4/7$,

$$\delta_0(x) = x \frac{2}{1.4} - \frac{2^2}{2 \times 1.4} + \log\left(\frac{3}{7}\right) = \frac{10}{7}x - \frac{10}{7} + \log\left(\frac{3}{7}\right) \approx 1.429x - 2.276,$$

$$\delta_1(x) = x \frac{5.5}{1.4} - \frac{5.5^2}{2 \times 1.4} + \log\left(\frac{4}{7}\right) = \frac{55}{14}x - \frac{605}{56} + \log\left(\frac{4}{7}\right) \approx 3.929x - 11.363.$$

Predict class 1 when $\delta_1(x) \geq \delta_0(x)$, i.e.

$$3.929x - 11.363 \geq 1.429x - 2.276 \iff x \geq 3.635$$

The estimated cutoff is $x^* \approx 3.635$: $x = 3.5$ is predicted as class 0, while $x = 4.5$ is as class 1.

Pop-up quiz: LDA in 1D

Suppose

$$X \mid (Y = 0) \sim \mathcal{N}(2, \sigma^2), \quad X \mid (Y = 1) \sim \mathcal{N}(6, \sigma^2),$$

and now class 1 becomes more common than class 0, so $\pi_1 > \pi_0$.

Question: Compared with the equal-prior case, what happens to the LDA cutoff?

- A) It stays at $x = 4$
- B) It moves left of 4, so class 1 is predicted more often
- C) It moves right of 4, so class 0 is predicted more often
- D) It cannot be determined without knowing σ^2

Answer: B. Since $x^* = \frac{\mu_0 + \mu_1}{2} - \frac{\sigma^2}{\mu_1 - \mu_0} \log(\pi_1 / \pi_0)$, larger π_1 shifts the cutoff left and makes class 1 easier to predict.

Wrap-up

Recapping logistic regression & classification assessment

- Classification performance depends on the threshold p^* , not just on the fitted probabilities.
- Confusion matrices, false positive/negative rates, ROC curves, and AUC reveal different aspects of classifier performance.

Generative models and LDA:

- Discriminative methods model $\Pr(Y | X)$ directly, while generative methods model $\Pr(Y)$ and $P(X | Y)$ and then use Bayes' rule.
- LDA is a generative classifier that assumes Gaussian class-conditional distributions with a common variance/covariance structure.
- Under LDA, prediction is based on comparing linear discriminant scores, which leads to linear decision boundaries.

Suggested reading: [JWHT21, Ch. 4.4.1; review classification assessment in Ch. 4.4.2]

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.