STA 35C: Statistical Data Science III

Lecture 10: Generative Models for Classification

Dogyoon Song

Spring 2026, UC Davis

Announcements

Homework 3 is due tomorrow (Tue, Apr 21, 11:59 PM)

- 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)

Midterm 1 is in class on Fri, Apr 24

- You may bring *one **hand-written** letter-sized (8.5 × 11 inches), double-sided sheet of paper* with formulas and brief notes
- **Calculator**: Simple (non-graphing) calculators only
- **No textbooks** or other materials beyond the single cheat sheet
- **SDC accommodations:** Confirm scheduling with AES online
- A practice midterm is available on the course webpage

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

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

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

Decision boundary (when $K = 2$): We predict $\hat{y} = 1$ if and only if

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

- When there are two predictors ($p = 2$), 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

		True default status		
		No	Yes	Total
Predicted default status	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?

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

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

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

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

Recap: The ROC curve and selecting threshold p^*

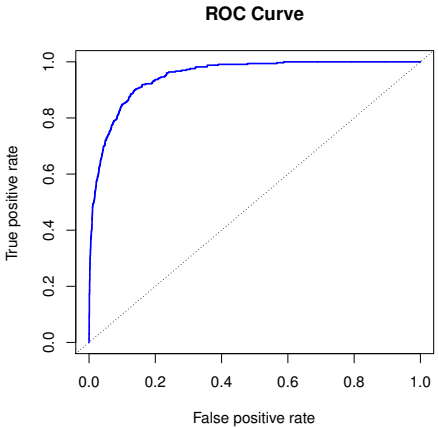

Figure: An example ROC curve, with AUC [JWHT21, Figure 4.8].

The ROC curve and AUC

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

- Directly model $\Pr(Y = k \mid X = x)$
- 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* $\pi_k \coloneqq \Pr(Y = k)$ that a randomly chosen observation is from the k-th class
  - The class-conditional *PMF/density* $f_k(X) \coloneqq \Pr(X \mid Y = k)$[1] of X within class k
- Then use Bayes' theorem to compute the *posterior probability*:

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

- The resulting classification rule is $\hat{y}(x) = \arg\max_k \pi_k f_k(x)$

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

Contrasting the two approaches

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

Discriminative workflow:

- *Postulate a functional form* for $\Pr(Y = k \mid X = x)$ (or the log-odds)
- Fit parameters from data
- Directly output posterior class probabilities $p(x) = \Pr(Y = k | X = x)$

Generative workflow:

- *Postulate each class distribution* $f_k(x)$
  - Key challenge: specifying X's distribution per class
- Estimate $\pi_k = P(Y = k)$ (often by 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 is more restrictive but can be more data-efficient when the assumptions are good

Visualization of the workflow

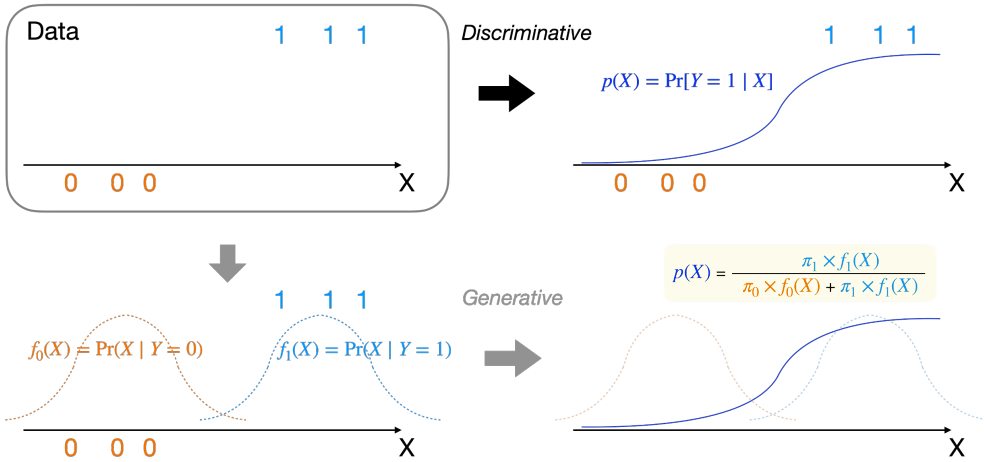

Figure: A schematic contrast: discriminative approaches (**black**) directly model $\Pr(Y \mid X)$, while generative approaches (**gray**) model $\Pr(X \mid Y)$ and $\Pr(Y)$ first, then obtain $\Pr(Y \mid X)$ via Bayes' rule.

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
- $X \mid (Y = 0) \sim \mathcal{N}(\mu_0, \sigma^2)$
- $X \mid (Y = 1) \sim \mathcal{N}(\mu_1, \sigma^2)$
- The two classes may have different means, but share the same variance σ^2

So the class-conditional densities are

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

Using Bayes' rule,

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

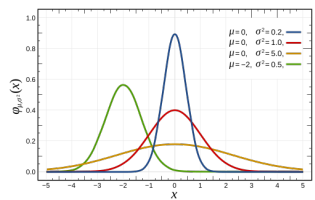

Figure: PDF of 1D Gaussian distribution (Image from Wikipedia[a]).

LDA in 1D: Prediction and decision boundary

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 is the same, we predict

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

Taking logs of $\pi_k \, f_k(x)$ and ignoring constants, we define the linear discriminant function

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

We predict class 1 when $\delta_1(x) \geq \delta_0(x)$, and the decision boundary is a single cutoff point:

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

- If $\pi_0 = \pi_1$, then $x^* = \dfrac{\mu_0 + \mu_1}{2}$, 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)

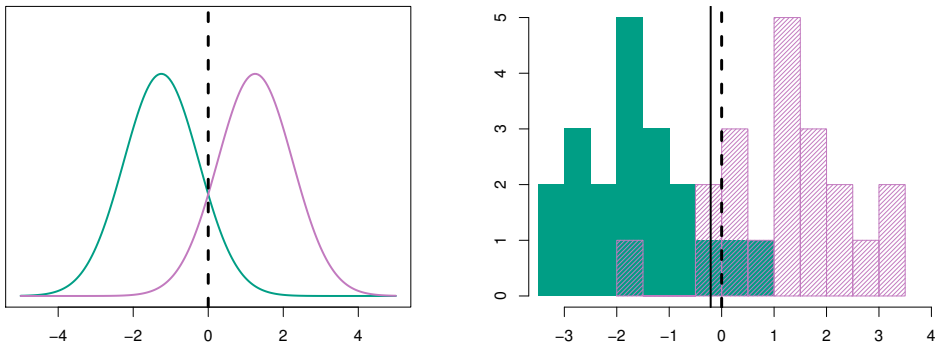

Figure: **(Left)** Two one-dimensional normal density functions. The dashed vertical line is the Bayes decision boundary. **Right)** Histograms of 20 observations from each class. The dashed vertical line again shows the Bayes decision boundary, while the solid vertical line represents the LDA decision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

LDA in 1D: Estimating the LDA parameters

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

Class priors:

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

Class means:

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

Common variance:

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

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 discriminant functions $\delta_0(x)$ and $\delta_1(x)$
- Compute the decision boundary and classify a new x

Illustration of LDA: 1) Parameter estimation

Example

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

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

Class means: Sample mean for each class

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

Common variance (pooled): Pooled sample covariance

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=0}^{1} \sum_{\substack{i: \\ y_i = k}} (x_i - \hat{\mu}_k)^2 = \frac{1}{5}\big(2+2\big) = 0.8$$

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

Illustration of LDA: 2) Discriminants and classification

Example

Recall the 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 $\hat{\mu}_0 = 2, \hat{\mu}_1 = 6, \hat{\sigma}^2 = 0.8, \hat{\pi}_0 = 3/7, \hat{\pi}_1 = 4/7$,

$$\delta_0(x) = x\,\frac{\hat{\mu}_1}{0.8} \;-\; \frac{\hat{\mu}_1^2}{2 \times 0.8} \;+\; \log\Big(\frac{3}{7}\Big) \approx 2.5x - 3.347,$$

$$\delta_1(x) = x\,\frac{\hat{\mu}_2}{0.8} \;-\; \frac{\hat{\mu}_2^2}{2 \times 0.8} \;+\; \log\Big(\frac{4}{7}\Big) \approx 7.5x - 23.060$$

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

$$7.5x - 23.060 \geq 2.5x - 3.347 \quad \Longleftrightarrow \quad x \geq 3.94.$$

So the estimated LDA cutoff is $x^* \approx 3.94$: for example, $x = 3.5$ is classified as class 0, while $x = 4.5$ is classified as class 1.

Pop-up quiz: LDA in 1D

Suppose

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

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 $\Pr(Y \mid X)$ directly, while generative methods model $\Pr(Y)$ and $P(X \mid Y)$ 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.

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.