STA 35C: Statistical Data Science III

Lecture 10: Generative Models for Classification

Dogyoon Song

Spring 2025, UC Davis

Announcement

Homework 2 is due tomorrow (Tue, Apr 22) at 11:59 PM PT

- Please ensure your submission is properly formatted and submitted on time (See HW instructions & syllabus; there will be a separate announcement on Canvas)

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

- You may bring *one **hand-written** sheet of letter-sized paper (8.5 $\times$ 11 inches), double-sided* with formulas, brief notes, etc.
- **Calculator**: Simple (non-graphing) calculators only
- **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 4–5 pm, TA: Mon & Thu 1–2 pm)
- Questions on Piazza

Agenda

- **(Recap)** Logistic regression
  - From log-odds to (conditional) probabilities
  - Multinomial logistic regression ($K \geq 2$)
  - Decision boundary
- **(Recap)** Classification assessment
  - Error rates & Bayes classifier
  - Confusion matrix: False positives & false negatives
  - ROC curve
- Generative models for classification
  - Generative vs. discriminative models
  - Why generative modeling?
- Linear discriminant analysis (LDA)
  - Basics: $p = 1$ case & exntension to general $p \geq 1$
  - Example ($p = 2$)
  - Parameter estimation

Recap: Simple logistic regression ($p = 1, K = 2$)

Model:

$$\log\left(\frac{\Pr[Y = 1 \mid X]}{\Pr[Y = 0 \mid X]}\right) = \beta_0 + \beta_1 X$$

or equivalently, $\Pr(Y = 1 \mid X = x) = \sigma(\beta_0 + \beta_1 x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$

What do we do with this? If the model is correct:

- For each $X = x$, "$Y = 1$" is $e^{\beta_0+\beta_1 x}$ times more likely than "$Y = 0$"
- That is,

$$\Pr(Y = 1 \mid X = x) : \Pr(Y = 0 \mid X = x) = e^{\beta_0+\beta_1 x} : 1$$

- To convert this ratio into conditional probabilities, we normalize:

$$\implies \Pr(Y = 1 \mid X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \quad \text{and} \quad \Pr(Y = 0 \mid X = x) = \frac{1}{1 + e^{\beta_0+\beta_1 x}}$$

Recap: Extending logistic regression to $p > 1$

Extension to $p > 1$ is straightforward: Now we have

$$\log\left(\frac{\Pr[Y=1 \mid X]}{\Pr[Y=0 \mid X]}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

What do we do with this?

- For each $X = x$, "$Y = 1$" is $e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$ times more likely than "$Y = 0$"
- That is,

$$\Pr(Y=1 \mid X=x) : \Pr(Y=0 \mid X=x) = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} : 1$$

- Again, normalize to get conditional probabilities:

$$\implies \Pr(Y=1 \mid X=x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

Recap: Extending logistic regression to $K > 2$

If $K = 3$, we model two log-odds *separately* (with class 3 as reference):

$$\log\left(\frac{\Pr[Y=1|X]}{\Pr[Y=3|X]}\right) = \beta_{1,0} + \beta_{1,1}X_1 + \cdots + \beta_{1,p}X_p$$
$$\log\left(\frac{\Pr[Y=2|X]}{\Pr[Y=3|X]}\right) = \beta_{2,0} + \beta_{2,1}X_1 + \cdots + \beta_{2,p}X_p$$

- Note the *double indices* on coefficients: one for the response label ($Y = 1, 2$) and another for the predictors $(X_1, \ldots, X_p)$

What do we do with these? (Assume $p = 1$ for simplicity)

- Letting $p_k(x) := \Pr[Y = k \mid X = x]$, we have

$$p_1(x) : p_2(x) : p_3(x) = e^{\beta_{1,0}+\beta_{1,1}x} : e^{\beta_{2,0}+\beta_{2,1}x} : 1$$

- Again, normalize to obtain conditional probabilities (see Lecture 9, Slide 12):

$$\implies p_k(x) = \Pr(Y = k \mid X = x) = \frac{e^{\beta_{k,0}+\beta k,1x}}{1 + e^{\beta_{1,0}+\beta_{1,1}x} + e^{\beta_{2,0}+\beta_{2,1}x}}$$

Recap: Decision boundary ($K = 2$)

Prediction rule: Once we have $p(X) = \Pr(Y = 1 \mid X)$, we predict

$$\hat{Y} = \begin{cases} 1 & \text{if } p(X) \geq p^*, \\ 0 & \text{otherwise.} \end{cases}$$

where p^* (e.g., 0.5) is a tunable parameter

Under a logistic model:

$$\begin{aligned} p(x) \geq p^* \quad & \iff \quad \log\left(\tfrac{p(x)}{1-p(x)}\right) \geq \log\left(\tfrac{p^*}{1-p^*}\right) \\ & \iff \quad \beta_0 + \sum_{i=1}^{p} \beta_i x_i \geq \log\left(\tfrac{p^*}{1-p^*}\right) \end{aligned}$$

For $p = 2$: if $\beta_2 > 0$ (***Question:*** *What if $\beta_2 < 0$ or $\beta_2 = 0$?*),

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq \log\left(\frac{p^*}{1-p^*}\right) \quad \Longrightarrow \quad x_2 \geq -\frac{\beta_1}{\beta_2} x_1 + \frac{1}{\beta_2}\left[-\beta_0 + \log\left(\frac{p^*}{1-p^*}\right)\right]$$

Error rate

Error rate: Fraction of observations that are misclassified

$$\text{Error rate} = \frac{1}{n} \sum_{i=1}^{n} I(\hat{y}_i \neq y_i)$$

Bayes classifier:

$$X \mapsto \arg\max_k \Pr(Y = k \mid X)$$

- Optimal classifier that minimizes error rate *in theory*
- Usually impossible to compute *in practice*, since $\Pr(Y \mid X)$ is unknown
- **Question:** Even if we could compute Bayes classifier, is the error rate always the best measure?
  - Some classification errors could be costlier than others
  - e.g., missing a cancer is worse than a false alarm

More on error metrics

		True class		
		− or Null	+ or Non-null	Total
Predicted class	− or Null	True Neg. (TN)	False Neg. (FN)	N^*
	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P^*
	Total	N	P	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P^*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N^*	

Figure: **Top:** Possible classification outcomes in a population. **Bottom:** Important measures for classification, derived from the confusion matrix [JWHT21, Tables 4.6 & 4.7].

Minimizing total error rate can be suboptimal if FP and FN have different costs

Threshold selection

Many classifiers (e.g. logistic regression) produce $\hat{p}(x) = \Pr(Y = 1 \mid x)$

- If $\hat{p}(x) \geq p^*$, predict $Y = 1$, else 0
- Changing p^* alters false positives and false negatives

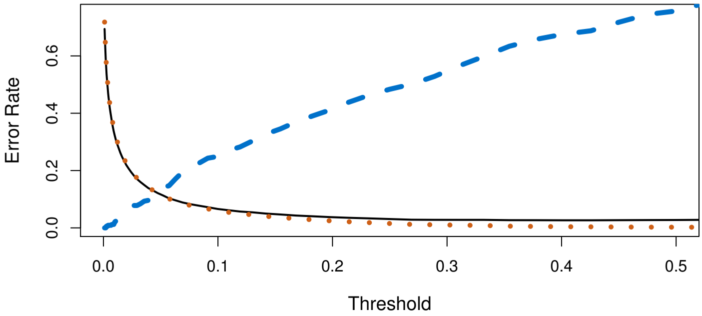

Figure: False positive (orange dotted) and false negative (blue dashed) error rates as a function of the threshold value p^* for the `Default` dataset [JWHT21, Figure 4.7].

Receiver operating characteristic (ROC) curve

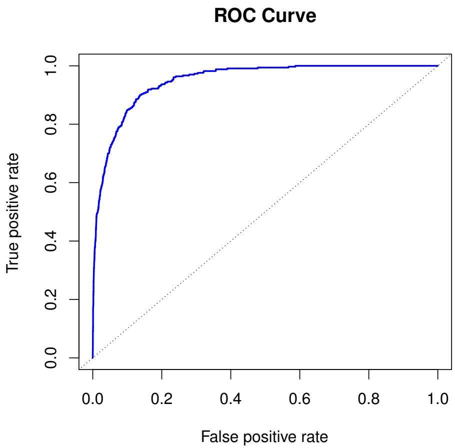

Figure: An example ROC curve, with AUC [JWHT21, Figure 4.8].

ROC curve

- Plot TPR vs. FPR as p^* moves $0 \to 1$
  - $\mathrm{TPR} = \frac{TP}{P} = \frac{TP}{TP+FN}$
  - $\mathrm{FPR} = \frac{FP}{N} = \frac{FP}{TN+FP}$
- Summarize the performance via area under curve (AUC)

Area under curve (AUC)

- Reflects overall discriminative power across thresholds
  - Perfect classifier: $\mathrm{AUC} = 1$
  - Random guess: $\mathrm{AUC} = 0.5$

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

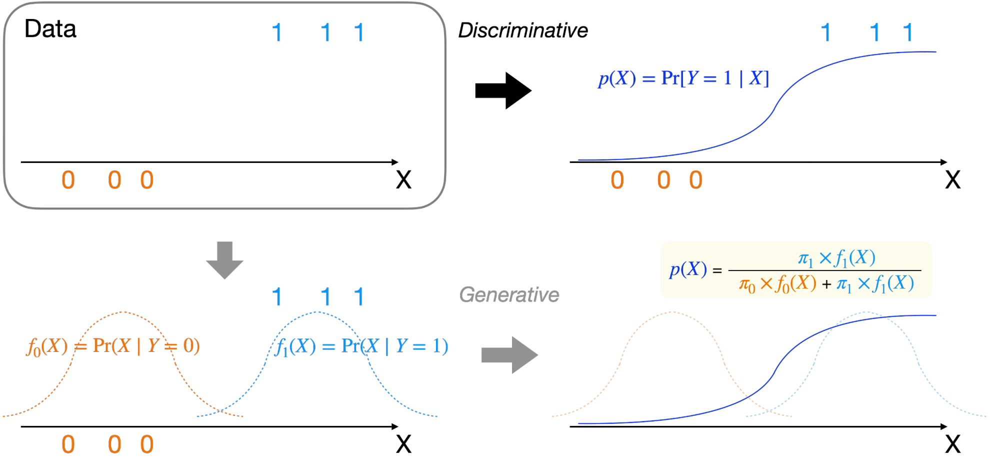

Figure: A schematic contrast: discriminative approaches (**black**) directly learns $\Pr(Y|X)$, while generative (**gray**) models $\Pr(X|Y)$ and $\Pr(Y)$ first, then obtains $\Pr(Y|X)$ via Bayes.

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.

Answer: (C). Proper distribution assumptions can yield a data-efficiency advantage.

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

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

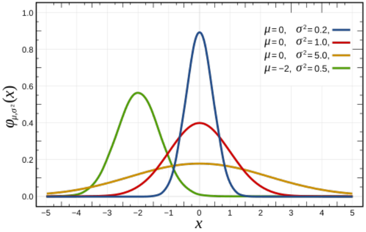

Figure: PDF of 1D Gaussian distribution (Image from Wikipedia[a]).

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

Decision boundary for $p = 1$

By Bayes' theorem:

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

where $\pi_k := \Pr(Y = k)$ and $f_k(X) := \Pr(X \mid Y = k)$

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

- We find k that maximizes $\log(\pi_k\, f_k(x))$; when σ^2 is common across classes,

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

Linear discriminant function: We choose k with largest $\delta_k(x) := x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$; the boundary between class k and class $j \neq k$ is *linear* in x

Extending LDA from $p = 1$ to $p \geq 1$

General assumption:

- $\pi_k = P(Y = k)$
- $X \in \mathbb{R}^p$ and $X \mid (Y = k) \sim \mathcal{N}(\mu_k, \Sigma)$; common covariance Σ, distinct μ_k
- The class-conditional density (multivariate Gaussian):

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

Discriminant function[2]:

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

Again, the boundary between class k and class $j \neq k$ is *linear* in x

[2] Multi-dimensional extension of 1-dimensional version $\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$

Extension from $p = 1$ to $p \geq 1$: Visualization of density

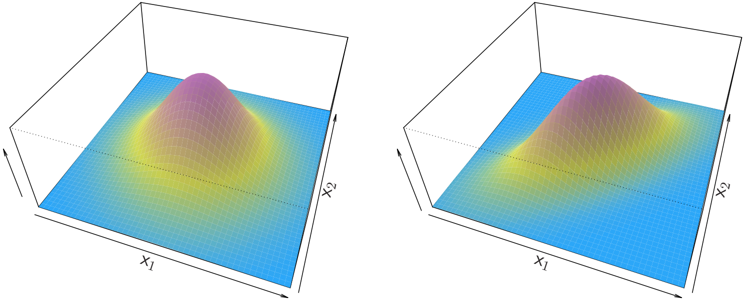

Figure: Illustration of multivariate Gaussian density functions for $p = 2$ Left: The two predictors are uncorrelated. Right: The two variables have a correlation of 0.7 [JWHT21, Figure 4.5].

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.

Answer: (B). With one shared Σ, the $(x - \mu_k)$ quadratic terms cancel, leaving a linear boundary.

Wrap-up

Recapping logistic regression & classification assessment

- From log-odds model to conditional probabilities
- Decision boundary
- Confusion matrix: False positives/false negatives & ROC curve

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

Linear discriminant analysis (LDA):

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

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.