STA 35C: Statistical Data Science III

Lecture 11: Generative Models (cont'd) & Review for Midterm 1

Dogyoon Song

Spring 2026, UC Davis

Announcements

Midterm 1 is in class on Fri, Apr 24

- You may bring *one **hand-written** letter-sized (8.5 $\times$ 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:

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

Today:

- Wrap-up on generative approaches for classification: LDA, QDA, Naive Bayes
- Midterm 1 review by theme:
  - probability and random variables
  - regression
  - classification

Recap: Discriminative vs. generative models

Discriminative methods directly model the posterior class probabilities:

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

Generative methods instead model

- the class prior $\pi_k := \Pr(Y = k)$
- the class-conditional PMF/density $f_k(x)$ of X given $Y = k$

Then use Bayes' theorem to compute posterior probabilities:

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

Classification rule:

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

Recap: LDA essentials in 1D

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 classes may have different means, but share a common variance σ^2

Decision rule: predict class 1 if $\pi_1 f_1(x) \geq \pi_0 f_0(x)$ or equivalently if

$$\delta_1(x) \geq \delta_0(x), \qquad \text{where} \qquad \delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k.$$

Cutoff in 1D:

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

- If $\pi_0 = \pi_1$, then $x^* = \dfrac{\mu_0 + \mu_1}{2}$
- Changing the class priors shifts the cutoff toward the less likely class

Recap: LDA visualization (1D)

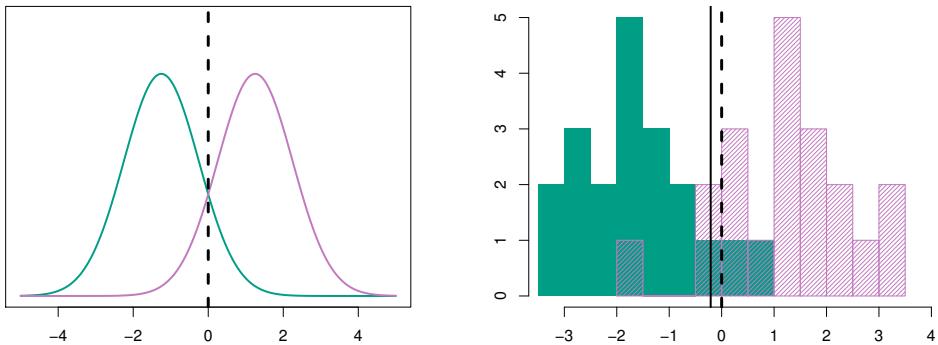

Figure: **(Left)** Two one-dimensional normal density functions. The dashed vertical line is the Bayes decision boundary. **(Right)** Histograms of 20 observations from each class. The dashed vertical line again shows the Bayes decision boundary, while the solid vertical line represents the LDA decision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

LDA visualization (2D, K=3)

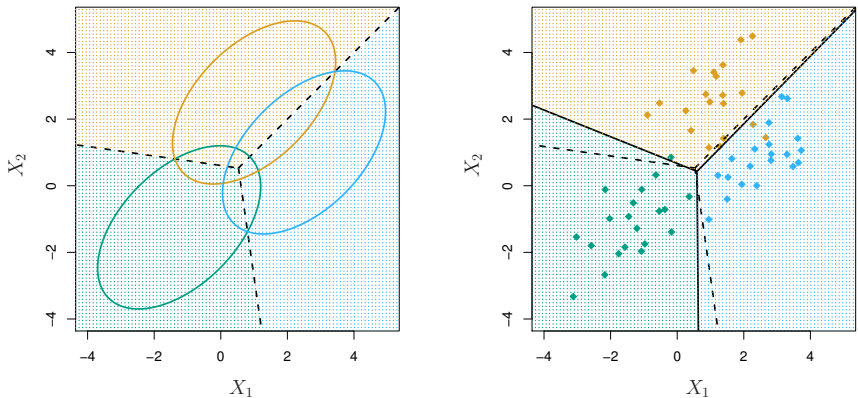

Figure: An illustration of LDA decision boundaries. Observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, a class-specific mean vector, and a common covariance matrix. **(Left)** Ellipses indicating the 95% probability region for each of the three classes, with 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].

Beyond LDA: QDA and Naive Bayes

Different assumptions about $f_k(x) = p(x \mid Y = k)$ lead to different generative classifiers

Quadratic Discriminant Analysis (QDA)

- Assumes $X \mid (Y = k) \sim \mathcal{N}_p(\mu_k, \Sigma_k)$
- Unlike LDA, each class has its own covariance matrix Σ_k
- More flexible than LDA; decision boundaries can be curved
- Needs more data because it estimates more parameters

Naive Bayes

- Useful when X is high-dimensional or discrete, and modeling the full joint density $f_k(x)$ is difficult
- *Naively assumes* that predictors X_j are conditionally independent given the class $Y = k$
  $\implies \quad f_k(x) = \prod_{j=1}^{p} f_{k,j}(x_j)$
- Fast and often effective, even when the independence assumption is only approximately true

Visual comparison between LDA and QDA

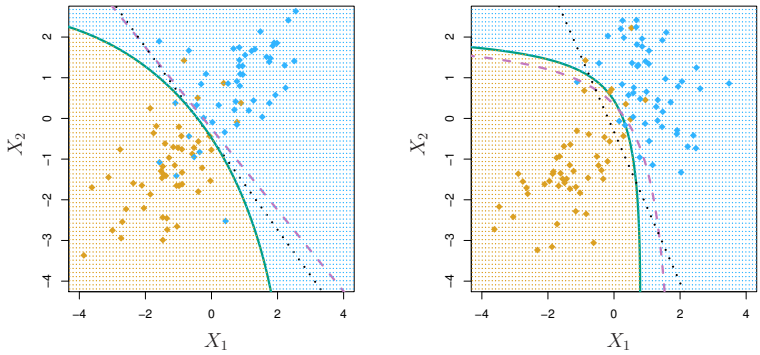

Figure: A comparison of LDA and QDA decision boundaries. The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) boundaries are shown. LDA always produces a linear boundary, whereas QDA can be curved [JWHT21, Figure 4.9].

Takeaway: LDA is less flexible but more stable with limited data; QDA is more flexible but estimates more parameters and can overfit.

Midterm 1 review

What you should be able to do:

- Translate between words, notation, formulas, and interpretations.
- Define the relevant objects clearly: events, random variables, predictors, responses, fitted values, probabilities, and class labels.
- Compute core quantities carefully: probabilities, expectations, variances, least-squares estimates, fitted probabilities, and classification metrics.
- Interpret results in context: coefficients, R^2, intervals, p-values, odds ratios, thresholds, and error rates.
- Distinguish commonly confused ideas: conditional probability vs. independence, prediction vs. inference, CI vs. PI, regression vs. classification, discriminative vs. generative.
- Recognize model assumptions and explain what can go wrong when they fail.

Exam strategy: It is a good practice to clearly define your setup, compute cleanly and carefully, and write one-sentence interpretations for numerical answers.

Review: Probability and random variables

Probability language

- Outcome ω, sample space Ω, event $A \subseteq \Omega$
- Set operations: $A \cup B$, $A \cap B$, A^c, $A \setminus B$
- Probability law: nonnegativity, normalization, and additivity for disjoint events

Random variables

- A random variable is a function $X : \Omega \to \mathbb{R}$
- Discrete: PMF $p_X(x) = P(X = x)$
- Continuous: PDF f_X, with $P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$
- CDF: $F_X(x) = P(X \leq x)$, for both discrete and continuous X

Review: Expectation, variance, and covariance

Expectation and variance:

$$\mathbb{E}[X] = \sum_x x p_X(x) \quad \text{or} \quad \mathbb{E}[X] = \int x f_X(x)\, dx$$

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

For linear combinations,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y],$$

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y).$$

In particular,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).$$

Covariance and correlation

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \qquad \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

- If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Review: Conditional probability, Bayes, and dependence

Conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0.$$

Interpretation: restrict attention to outcomes inside B.

Law of total probability

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c).$$

Bayes' theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}.$$

Independence

$$A \perp B \quad \Longleftrightarrow \quad P(A \cap B) = P(A)P(B).$$

If $P(B) > 0$, then independence implies $P(A \mid B) = P(A)$.

Review: Statistical learning big picture

Supervised learning: data (x_i, y_i), with predictors X and response Y

Problem types

- Regression: Y is quantitative
- Classification: Y is categorical

Goals

- Accurately predict Y for new X (prediction)
- Understand how Y is related to X (inference)

Working model

$$Y = f(X) + \varepsilon.$$

- Prediction error has a reducible part from estimating f and an irreducible part from noise ε that no method can eliminate.

Review: Linear regression basics

Linear regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

Least squares: Choose coefficients that minimize

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

For simple linear regression,

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Interpretation

- In simple regression, β_1 is the average change in the mean response per one-unit increase in X
- In multiple regression, β_j is interpreted holding the other predictors fixed
- This is an association unless causal assumptions are justified

Review: Linear regression inference and model fit

Inference for coefficients

$$t = \frac{\hat{\beta}_j}{\widehat{\mathrm{SE}}(\hat{\beta}_j)} \qquad \text{for testing} \qquad H_0 : \beta_j = 0.$$

$$\text{95\% CI:} \qquad \hat{\beta}_j \pm t_{n-p-1,0.975}\widehat{\mathrm{SE}}(\hat{\beta}_j).$$

Model fit

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-p-1}}, \qquad R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \qquad R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{RSS}/(n-p-1)}{\mathrm{TSS}/(n-1)}.$$

- RSE: typical residual size, in units of Y
- R^2: proportion of variation in Y explained by the model
- Adjusted R^2: penalizes unnecessary predictors

Confidence interval vs. prediction interval

- Confidence interval: for a mean response
- Prediction interval: for a new observation; wider because it includes irreducible noise

Review: Single predictor vs. multiple predictors

Interpretation changes

- In simple regression, β_1 describes the marginal association between X_1 and Y
- In multiple regression, β_j describes the association between X_j and Y, holding the other predictors fixed

Why conclusions can change

- If predictors are correlated, a variable can look significant by itself but not after adjustment
- Thus, simple and multiple regression can lead to different signs, sizes, or significance levels

Two types of tests

- Individual t-test: $H_0 : \beta_j = 0$
- Overall F-test: $H_0 : \beta_1 = \cdots = \beta_p = 0$, asking whether *any* predictor is useful

Model comparison

- Adding predictors never decreases training R^2
- Adjusted R^2 can decrease if the new predictors are not actually helpful

Review: Regression extensions and pitfalls

Model extensions

- Polynomial terms: $X, X^2, X^3, \ldots$
- Interaction terms: $X_1 X_2$
- Dummy variables for categorical predictors
  - With an intercept and K categories, use $K - 1$ dummies

Common pitfalls

- Nonlinearity: residual plots show systematic curvature
- Correlated errors: standard errors and p-values become unreliable
- Non-constant variance: residual spread changes with fitted values
- Outliers / high-leverage points: individual points strongly affect fit
- Collinearity: coefficients become unstable and hard to interpret

Diagnostic habit: Always look at residual plots and check whether the assumptions are plausible.

Review: Classification and logistic regression

Classification

- Y is categorical
- Goal: estimate class probabilities and assign a class label

Binary logistic regression

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

- The logistic function converts the linear predictor into a valid probability in $(0, 1)$.

Interpretation

- A one-unit increase in x_j changes the log-odds by β_j
- Equivalently, the odds are multiplied by e^{β_j}, holding other predictors fixed

Estimation and prediction

- Logistic regression is fit by maximum likelihood, not least squares
- Predict class 1 if $\hat{p}(x) \geq p^*$, where the threshold p^* controls the FP/FN tradeoff

Review: Classification boundaries and LDA

Logistic regression decision boundary

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \log\left(\frac{p^*}{1-p^*}\right).$$

For $p = 2$, this is a line.

Discriminative vs. generative

- Discriminative: model $P(Y \mid X)$ directly, e.g. logistic regression
- Generative: model $P(Y)$ and $P(X \mid Y)$, then use Bayes' rule

LDA

- Assumes $X \mid (Y = k)$ is Gaussian within each class
- Classes have different means but share a common variance/covariance
- Classify by comparing $\pi_k f_k(x)$, equivalently by comparing linear discriminant scores $\delta_k(x)$
- Decision boundaries are linear

Review: Classification assessment

Confusion matrix quantities

$$\text{Error rate} = \frac{FP + FN}{TP + TN + FP + FN}, \qquad \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

$$\text{FPR} = \frac{FP}{TN + FP}, \qquad \text{FNR} = \frac{FN}{TP + FN}, \qquad \text{TPR} = \frac{TP}{TP + FN} = 1 - \text{FNR}.$$

Threshold choice

- Lower threshold p^*: predict positive more often
- Usually increases TPR but also increases FPR
- Choose p^* based on the relative costs of false positives and false negatives

ROC and AUC

- ROC curve plots TPR vs. FPR as the threshold changes
- AUC summarizes overall ranking/discrimination ability across thresholds

Wrap-up: How to prepare

Prioritize fluency over memorization: For Midterm 1, focus on being able to:

- Define the relevant objects clearly: events, random variables, predictors, responses, fitted values, or class labels.
- Connect formulas to interpretations: know not only how to compute a quantity, but also what it means in context.
- Work fluently with probability tools: conditioning, Bayes' theorem, expectation, variance, covariance, and independence.
- Explain the statistical-learning workflow: supervised learning, regression vs. classification, prediction vs. inference, and model assumptions.
- Interpret regression outputs: least squares estimates, coefficients, standard errors, confidence intervals, R^2, dummy variables, and diagnostics.
- Interpret classification outputs: logistic probabilities, thresholds, confusion-matrix metrics, ROC/AUC, FP/FN tradeoffs, and the generative logic behind LDA.

Advice: Work through the practice midterm once without notes, then review the concepts, formulas, and interpretations you could not recall comfortably.

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.