STA 35C: Statistical Data Science III

Lecture 12: Mid-course Review / Resampling Methods Overview

Dogyoon Song

Spring 2025, UC Davis

Announcement

Midterm 1 solution and scores are available online

- **Discussion tomorrow** will review the midterm questions
- You may look over your graded exam there (pick it up at the start, return it at the end)

Grade disputes/adjustments

- If you believe your score should be changed for any question, please email the TA by noon on Wednesday (April 30) with:
  - The specific problem(s) you want regraded
  - A clear explanation of why you believe you deserve a different score (e.g., pointing out the key elements in your answer that match the official solution)

Mid-course survey

- Please take 10 minutes to complete the survey on Canvas
- All feedback and any constructive suggestions/requests are welcome

Agenda

- **Brief review of what we've covered:**
  - Supervised learning
  - Regression
  - Classification
  - Model assessment & the bias-variance tradeoff
- **Overview of what's next (next three weeks):**
  - Resampling methods
    Q: How can we estimate test MSE using training data?
    Q: How can we enable inference beyond linear models?
  - Model selection
    Q: How can we systematically select relevant predictors?
  - Multiple hypothesis testing
    Q: What is the correct inferential framework after using data to select models?

Recap: Supervised learning

$$\underbrace{X}_{\text{predictors}} \rightarrow \underbrace{Y}_{\text{response}}$$

Goal: "Explain" or model Y using X

- Estimate $f : X \rightarrow Y$ so that $y \approx f(x)$

Why?

- **Prediction:** e.g., forecasting sales, predicting house prices
- **Inference:** identifying significant predictors, relationships among variables

Depending on the type of Y,

- **Regression**: Y is numeric
- **Classification**: Y is categorical

Recap: Regression

Problem setup

$$\underbrace{X}_{\text{predictors}} \longrightarrow \underbrace{Y}_{\text{numeric}} \in \mathbb{R}$$

Goal: Estimate $f : X \to Y$ to fit a regression line (or curve)

For what?

- **Prediction:** Given x_{new}, predict $y_{\text{new}} = \hat{f}(x_{\text{new}})$
- **Inference:** Estimate how X influences Y and assess significance

If we knew the distribution of (X, Y)...

- We might use $\hat{Y} = \mathbb{E}[Y \mid X]$
- In reality, we only have finite data, so we estimate from samples

Linear regression: 1) Estimation & Prediction

Linear regression model: $Y = \beta_0 + \beta_1 X$

- Simple and interpretable

Parameter estimation: Find β_0, β_1 that minimize

$$\mathrm{RSS} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{where} \quad \hat{y}_i = \beta_0 + \beta_1 x_i$$

Prediction: $\hat{y}_{\mathrm{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\mathrm{new}}$

Model fit:

- $R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \in [0, 1]$: proportion of variance in Y explained by the model
- Higher R^2 indicates better explanatory power
- Adding more predictors always increases R^2; R^2_{adj} penalizes for extra variables

Linear regression: 2) Inference

Significance test: Is $\beta_1 \neq 0$? (i.e., is X truly related to Y?)

- Null hypothesis $H_0 : \beta_1 = 0$ (no linear relationship)
- If $t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$ in magnitude, we reject H_0 and conclude significance

Why this test? You may have got a nonzero slope purely by luck, and want to verify it

- Under H_0, $\frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$ follows a t-distribution
- Observing a value far out in the tail suggests H_0 is unlikely, so reject it
- If you see a moderate value, you may not be able to reject H_0 (not enough evidence)

z	0.5	1.0	1.5	2.0	2.5	3.0	3.5
Approx. p-value	0.6171	0.3173	0.1336	0.0455	0.0124	0.0027	0.000465

Linear regression: 3) Interpretation

Interpretation of β_1:

- On average, Y changes by β_1 per unit increase in X
  - Individual outcomes may vary (noise)
  - The true slope could differ across X if the relationship is not perfectly linear
- It does not imply causation; only correlation

Interpretation in multiple linear regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

- β_1 is the effect of X_1 holding X_2 fixed (conditional effect)
- **Confounding:**
  - $\beta_{1,\text{simple}}$ vs. $\beta_{1,\text{multiple}}$ may differ if X_1 and X_2 are correlated
    - *Why?* $\beta_{1,\text{simple}}$ may include indirect effects through X_2
  - Including X_2 in regression model can change the estimated effect of X_1

Recap: Classification

Problem setup

$$\underbrace{X}_{\text{predictors}} \longrightarrow \underbrace{Y}_{\text{classes}} \in \{0, 1\}$$

Goal: Estimate f to define a decision boundary between classes

For what?

- **Prediction:** Given x_{new}, predict its class label
- **Inference:** Understand which predictors significantly affect the probability $\Pr(Y = 1)$

Key ideas:

- If we knew $\Pr[Y = 1 \mid X]$, we could classify $Y = 1$ if $\Pr[Y = 1 \mid X] \geq p^*$
- In reality, we need to estimate $\Pr[Y = 1 \mid X]$ from data, and use it
- Two approaches:
  - *Discriminative* approach: directly model $\Pr[Y = 1 \mid X]$
  - *Generative* approach: model $\Pr[X \mid Y]$, then use Bayes' theorem

Logistic regression: A discriminative approach

Model:

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

- Similar to linear regression, but the response is the log-odds of $Y = 1$

Parameter estimation: Find β_0, β_1 that maximizes the likelihood

$$\text{Likelihood}(\beta_0, \beta_1) = \Pr(\text{data} \mid \beta_0, \beta_1) = \prod_{i=1}^{n} \Pr(y_i \mid x_i;\, \beta_0, \beta_1)$$

- A higher likelihood means the observed data are more probable under the model

Prediction in two-steps:

- Calculate $\hat{p}_{\text{new}} = \sigma(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}})$, where $\sigma(z) = \frac{1}{1+e^{-z}}$
- Predict $Y = 1$ if $\hat{p}_{\text{new}} \geq p^*$

Example: Classifying 5 crabs via logistic regression

Data: 5 crabs, 2 species, single predictor (weight):

$$\text{Species A (label 0): } \{1.5, 2.5\} \quad \text{vs.} \quad \text{Species B (label 1): } \{2.0, 3.0, 4.0\}$$

Goal: Classify based on weight X

Fitted Model:

$$\log\left(\frac{p_B(X)}{p_A(X)}\right) = \beta_0 + \beta_1 X \quad \Longrightarrow \quad \hat{\beta}_0 \approx -5.30, \;\; \hat{\beta}_1 \approx 2.10$$

- Decision boundary near $x \approx 2.52$
- One misclassification is unavoidable (points at 2.0 vs. 2.5)
- Best overall likelihood is achieved by this compromise

Generative models for classification

Bayes' theorem:

$$\Pr[Y = 1 \mid X] = \frac{\Pr[Y = 1 \;\&\; X]}{\Pr[X]} = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

- $\pi_k = \Pr[Y = k]$: proportion of class k
- $f_k(x) = \Pr[X = x \mid Y = k]$: probability of $X = x$ conditioned on class k

Classification rule:

- Choose class k that maximizes $\pi_k \, f_k(x)$
- Requires modeling assumptions for $f_k(x)$
- Note that the marginal or prior probability for class k, π_k, also matters

Generative models for classification: Illustration

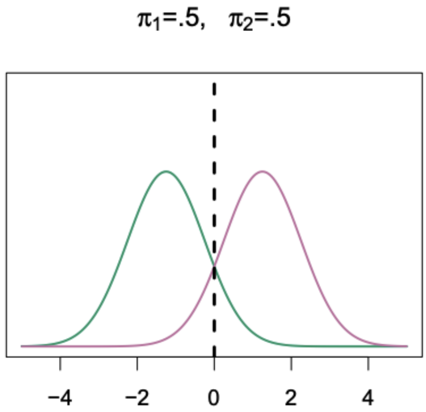

Figure: Generative classification compares likelihoods $f_k(x)$ weighted by π_k (Source: ISLR2 Ch. 4 Slides https://hastie.su.domains/ISLR2/Slides/Ch4_Classification.pdf).

Linear discriminant analysis: A generative approach

To move forward, modeling assumption required for $f_k(x) := \Pr[X = x \mid Y = k]$

Gaussian density assumption $\rightarrow$ LDA

- Assume $f_k(x)$ is Gaussian with mean μ_k and common variance σ^2
- Then $\Pr[Y = k \mid X = x]$ can be expressed using linear *discriminant functions*

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

  - Why this form?

$$k \text{ maximizes } \Pr[Y = 1 \,\&\, X = x] = \pi_k f_k(x) \quad \Longleftrightarrow \quad k \text{ maximizes } \log\left(\pi_k f_k(x)\right)$$

  - At any given $X = x$,

$$\log\left(\frac{\Pr[Y = 1 \mid X = x]}{\Pr[Y = 0 \mid X = x]}\right) = \delta_1(x) - \delta_0(x)$$

- Predict class k for which $\delta_k(x)$ is largest

Example: Classifying 5 crabs via LDA

Data: 5 crabs, 2 species, single predictor (weight):

$$\text{Species A (label 0): } \{1.5, 2.5\} \quad \text{vs.} \quad \text{Species B (label 1): } \{2.0, 3.0, 4.0\}$$

Goal: Classify based on weight X

Steps:

- Estimate class priors: $\hat{\pi}_A = \frac{2}{5}, \quad \hat{\pi}_B = \frac{3}{5}$
- Estimate means: $\hat{\mu}_A = \frac{1.5+2.5}{2} = 2, \quad \hat{\mu}_B = \frac{2+3+4}{3} = 3$
- Estimate common variance: $\hat{\sigma}^2 = \frac{1}{5-2}\left[(0.5^2 + 0.5^2) + (1.0^2 + 0^2 + 1.0^2)\right] = \frac{5}{6}$
- Form discriminants:

$$\delta_A(x) = \frac{\mu_A}{\sigma^2}x - \frac{\mu_A^2}{2\sigma^2} + \log \hat{\pi}_A = \frac{12}{5}x - \frac{12}{5} + \log\left(\frac{2}{5}\right),$$

$$\delta_B(x) = \frac{\mu_B}{\sigma^2}x - \frac{\mu_B^2}{2\sigma^2} + \log \hat{\pi}_B = \frac{18}{5}x - \frac{27}{5} + \log\left(\frac{3}{5}\right)$$

- Compare $\delta_A(x)$ vs. $\delta_B(x)$ to classify: $\delta_A(x) > \delta_B(x) \quad \Longleftrightarrow \quad x < \frac{5}{2} - \frac{5}{6}\log\left(\frac{3}{2}\right)$

Assessing models: 1) Error metrics

Regression models: Commonly use **MSE** (Mean Squared Error):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

- Lower MSE indicates better fit

Classification models: Often use **error rate**:

$$\text{Error Rate} = \frac{\#\ \text{Misclassified}}{\#\text{Sample size}}$$

- **False Positives (FP)** vs. **False Negatives (FN)** are also important
- A confusion matrix helps visualize these counts
  - e.g., in the crab example: (1) Which crabs were misclassified? (2) How do FP/FN rates shift if p^* changes?

Assessing Models: 2) Bias-variance tradeoff

Training vs. test performance:

- More flexible models fit the training data better (lower training MSE or error rate)
- However, they may generalize poorly to new (test) data
- This illustrates the **bias-variance tradeoff**:
  - High flexibility $\implies$ low bias but potentially high variance
  - Low flexibility $\implies$ higher bias but lower variance

Questions next:

- How do we estimate test error using only training data?
- How do we perform valid inference (e.g., confidence intervals, significance tests) for flexible or complex models?

Overview of what's coming next

1. **Resampling methods**
   - **Cross-validation**: Approximate test error from training data
   - **Bootstrap**: Enables inference (e.g. confidence intervals) when analytical formulas are unavailable
2. **Model selection**
   - Techniques for systematically choosing a subset of predictors (e.g. forward/backward selection, regularization)
   - Balances model complexity against predictive accuracy
3. **Multiple hypothesis testing**
   - After model selection using the data, standard inference can be misleading
   - We will learn how to adjust p-values and confidence intervals to maintain valid statistical inference