STA 35C: Statistical Data Science III

Lecture 12: Mid-course Review

Dogyoon Song

Spring 2025, UC Davis

Announcements

Midterm 1 scores are posted on Canvas

- Exam and solutions are posted on the course webpage
- Please review the solutions carefully and identify what to revisit
- You can review graded exams in discussion section tomorrow

Post-midterm review

- Homework 3 will include selected Midterm 1-style problems with modified numbers
- The TA will go over Midterm 1 problems in discussion section on Thursday

Mid-course survey

- Please take 10 minutes to complete the survey on Canvas (until Friday, May 1)
- All feedback and constructive suggestions/requests are welcome

Agenda

Quick post-midterm recap

Today: Bias-variance tradeoff

- Model accuracy: MSE and classification error
- Training error vs. test error
- Bias, variance, and irreducible error
- Why overly simple and overly flexible models might both fail

Looking ahead: estimating test error, controlling flexibility, and multiple testing

Recap: Supervised learning

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

Goal: learn a rule or function $\hat{f}$ that uses X to explain or predict Y

- **Prediction:** use $\hat{f}(x_{\text{new}})$ to predict a future response
- **Inference:** understand how Y is related to the predictors

Two major problem types

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

Note: We fit models using training data, but we care about performance on future data.

Recap: Probability and random variables

Probability language

- Outcome ω, sample space Ω, event $A \subseteq \Omega$
- Conditional probability:

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

- Bayes' theorem updates beliefs after observing evidence

Random variables

- A random variable is a function $X : \Omega \to \mathbb{R}$
- PMF/PDF/CDF describe the distribution of X
- Expectation and variance summarize center and spread

Useful formulas

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

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

Recap: Regression

Regression setup

$$\underbrace{X}_{\text{predictors}} \longrightarrow \underbrace{Y}_{\text{quantitative response}}$$

Linear regression model

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

- Coefficients are estimated by least squares

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

- β_j is interpreted as the average change in the mean response per unit increase in X_j, holding other predictors fixed
- Standard errors, confidence intervals, and t-tests quantify uncertainty
- R^2, adjusted R^2, and RSE summarize model fit

Recap: Classification

Classification setup

$$\underbrace{X}_{\text{predictors}} \longrightarrow \underbrace{Y}_{\text{class label}}$$

Logistic regression

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

- Predict class 1 if $\hat{p}(x) \geq p^*$
- Changing p^* changes the false positive / false negative tradeoff
- Confusion matrices, ROC curves, and AUC assess classification performance

Generative approach

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

- LDA assumes Gaussian class-conditional distributions with common covariance

Model accuracy: regression vs. classification

Regression: mean squared error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

- Smaller MSE means predictions are closer to observed responses
- Least squares minimizes training MSE, equivalently RSS

Classification: error rate

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

- Smaller error rate means fewer misclassified observations
- Other metrics may matter when false positives and false negatives have different costs

Training error vs. test error

Training error is computed on the same data used to fit the model:

$$\mathrm{MSE}_{\mathrm{train}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

Test error is computed on new, unseen data:

$$\mathrm{MSE}_{\mathrm{test}} = \frac{1}{n_{\mathrm{test}}}\sum_{j=1}^{n_{\mathrm{test}}}\left(y_j^{\mathrm{test}} - \hat{f}(x_j^{\mathrm{test}})\right)^2$$

Key distinction

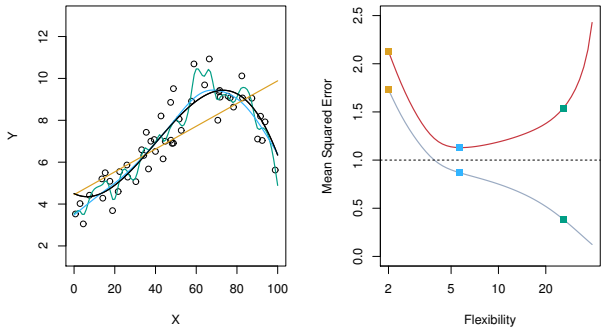
- Training error measures how well the model fits the data it saw
- Test error measures how well the model generalizes to new data
- We care most about test error, but it is usually not directly available

The challenge in practice

Reality: We usually do not have a large independent test set

- We fit models using training data
- But the model with the smallest training error may not have the smallest test error

Low training error $\not\Rightarrow$ low test error



Bias-variance tradeoff

At a fixed test point x_0, suppose

$$Y = f(x_0) + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \mathrm{Var}(\varepsilon) = \sigma^2.$$

The expected test error decomposes as

$$\mathbb{E}\left[\left(Y - \hat{f}(x_0)\right)^2\right] = \underbrace{\left(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2\right]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible error}}$$

where the expectation is over both the training sample used to fit $\hat{f}$ and the new noise ε.

Interpretation

- **Bias:** systematic error from using a model class that is too simple
- **Variance:** sensitivity of $\hat{f}$ to the particular training sample
- **Irreducible error:** noise that no method can eliminate

Interpreting bias and variance

High bias

- Model is too simple to capture the true relationship
- Example: fitting a line when the true relationship is strongly curved
- Leads to underfitting

High variance

- Model is too sensitive to small changes in the training data
- Example: fitting an overly flexible curve through noisy observations
- Leads to overfitting

Goal

- Choose enough flexibility to reduce bias
- But not so much flexibility that variance becomes too large

Model flexibility and test error

As model flexibility increases:

- Training error usually decreases
- Bias tends to decrease
- Variance tends to increase
- Test error may decrease at first, then increase

$$\text{Test error} \approx \text{Bias}^2 + \text{Variance} + \text{Irreducible error}$$

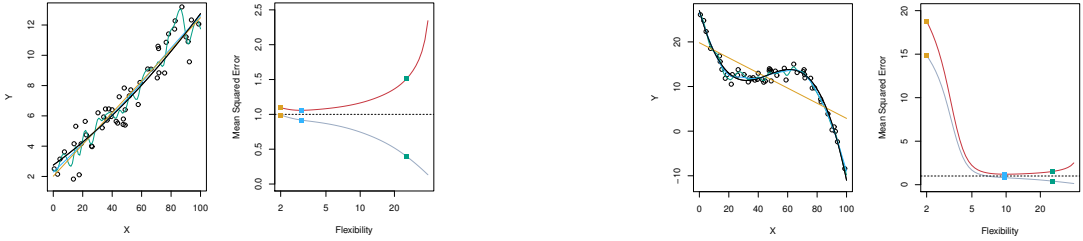


Figure: The same bias-variance idea appears across many settings [JWHT21, Figures 2.10 & 2.11]

Summary of the bias-variance tradeoff

Overly simple models

- High bias
- Low variance
- Underfit the data

Overly flexible models

- Low bias
- High variance
- Overfit the training data

Best predictive model

- Usually lies between these extremes
- Balances bias and variance to minimize test error
- Must be chosen using evidence about future performance, not training error alone

Looking ahead: natural questions

The bias-variance tradeoff raises several practical questions.

1. **How can we estimate test performance using training data only?**
   - Cross-validation
2. **How can we quantify uncertainty when formulas are difficult?**
   - Bootstrap
3. **How can we allow flexibility but avoid overfitting?**
   - Regularization and model selection
4. **When many predictors or tests are considered, which findings are reliable?**
   - Multiple testing and false discovery control

Roadmap until Midterm 2

Main theme: Building models that generalize well

- **Resampling methods**
  - Cross-validation for estimating test error
  - Bootstrap for estimating uncertainty
- **Model selection and regularization**
  - Choosing a relevant subset of predictors
  - Controlling flexibility via regularization
- **Multiple testing**
  - Adjusting inference when many hypotheses are considered

Unifying question: How do we learn useful patterns without mistaking noise for signal?

Wrap-up

- Before Midterm 1, we built the foundations: probability, regression, classification, and model interpretation.
- After Midterm 1, the main focus shifts from fitting models toward model performance and generalization.
- Training error measures fit to observed data; test error measures performance on new data.
- The bias-variance tradeoff explains why both underfitting and overfitting can lead to poor test performance.
- Upcoming lectures will cover tools such as cross-validation, bootstrap, regularization, and multiple-testing adjustments that help us choose models and quantify uncertainty more responsibly.

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.