CDA Spring 2026

Sneha Das

February 10, 2026

Week 2: MODEL SELECTION & ASSESSMENT

Last week, we witnessed the 'Wobble' of OLS and learned how Ridge Regression has a complexity knob to operationalize bias-variance tradeoff. This week, we transition from building to 'auditing'. In these exercises, you will play the role of a Scientific Auditor. You will investigate why high accuracy can be a lie (Leakage), how to choose models that a clinician could/would trust (One-SE Rule), and how to quantify the uncertainty of your findings (Bootstrap).

Exercise 2.1: Investigating information leakage

Goal: Empirically show that preprocessing outside the data-partition loop (eg: CV) leads to information leakage.

In file `ex_2_1_Q.py`, you are given a dataset of 50 patients and 1000 features. **Crucially, both X and y are pure random noise.** There is no relationship.

1. **Workflow A: Leakage -** Standardize the whole dataset, select the top 10 features most correlated with y, split the data, and fit an OLS model. Report the R^2 on the test set. Why is it so high despite the data being noise?
2. **Workflow B: non-leakage -** Split the data *first*. Perform the exact same standardization and feature selection *only* using the training set. Report the R^2 on the test set.
3. **Verdict:** Compare the results. Reflect on the differences between Workflow A & workflow B and why.

Exercise 2.2: Scientific Parsimony (The One-SE Rule)

Goal: Implement K-Fold CV on the Wine Quality dataset and justify a simpler model.

1. Open `ex_2_2_Q.py`. Implement a 10-fold Cross-Validation loop for Ridge Regression.
2. For each λ, calculate the mean MSE and the **Standard Error (SE)** of the folds.
3. **The One-SE Rule:** Identify λ_{min}. Now, find λ_{1se} (the largest λ whose error is within one SE of the minimum).
4. **Justification:** Plot the CV curve with error bars. If you were deploying this in a winery, why might you prefer the λ_{1se} model over the λ_{min} model?

Exercise 2.3: Scientific parsimony for the non-Parametric Knob (KNN)

Goal: Understand the complexity of KNN and apply the One-SE rule to a discrete parameter k.

1. Open `'ex_2_3_Q.py'`. We will use KNN Regression on the Wine dataset.
2. **(Selection):** Perform 10-fold CV for $k \in \{1, 2, \ldots, 100\}$. Plot the CV error.

3. **(Complexity):** Unlike Ridge where $\lambda \to \infty$ is the simplest model, in KNN, which direction of k represents the simplest (highest bias) model?
4. **(Audit):** Apply the One-SE rule. Identify the k_{min} and the k_{1se} (the simplest k within one SE). How much smoother is the prediction of k_{1se} compared to k_{min}?

Exercise 2.4: Analytical Guards (AIC vs. BIC)

Goal: Compare theoretical penalties when data is scarce.

1. Using a synthetic model path, calculate AIC and BIC for models of increasing size ($d = 1$ to 50).
2. Plot both criteria. Which one identifies the 'true' simple model earlier?
3. **Thought Experiment:** As the sample size N increases, which penalty ($\ln N$ vs 2) becomes more aggressive? Why is BIC considered a more conservative?

Exercise 2.5: Feature Uncertainty (The Bootstrap)

Goal: Use the Bootstrap to decide which features are reliable.

1. Revisit your best Ridge model from Ex 2.2. Implement a Bootstrap ($B = 1000$) to estimate the distribution of the coefficients $\hat{\beta}$.
2. Calculate the 95% Percentile Confidence Interval for the features.
3. **The Conclusiosn:** Does the interval for any feature cross zero? If so, what is your advice to a clinician who wants to use that feature for patient diagnosis?

Files Provided

- `ex_2_1_Q.py` / `sol.py`: The Leakage Simulation.
- `ex_2_2_Q.py` / `sol.py`: Ridge CV and Wine Audit with 1 SE rule.
- `ex_2_3_Q.py` / `sol.py`: KNN CV and Wine Audit with 1 SE rule.
- `ex_2_4_Q.py` / `sol.py`: AIC & BIC.
- `ex_2_5_Q.py` / `sol.py`: Bootstrap Reliability.