Cross-Validation and Subset Selection

Teaching Team of STA314

Department of Statistical Sciences
University of Toronto

Monday January 19, 2026

Road Map for today

1. Review of Model Selection Methods
2. Coding Examples + Exercises
3. Quiz 2 in the last 5 minutes

Recall: Approaches for Model Selection

Today, we will consider two approaches to model selection that we've seen in lecture:

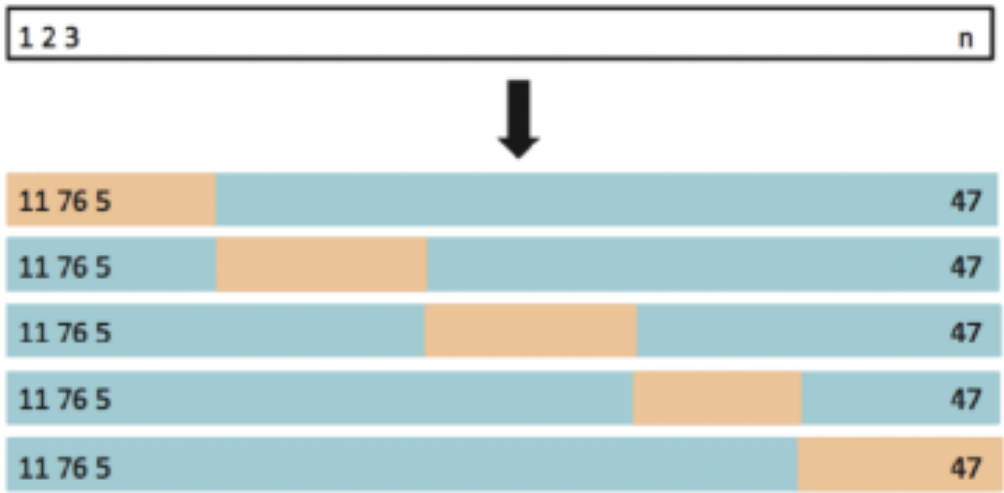
- Estimate the expected MSE by "holding out" a portion of your training data for validation:
  - Validation set approach
  - Cross-validation approach
- Make an adjustment to the training error to penalize more complex models:
  - Mallow's C_p
  - Adjusted R^2
  - AIC and BIC

Validation Set Approach

- Randomly divide the dataset into a *training set* and a *validation set*
- Train on the training set and then compute MSE on the validation set

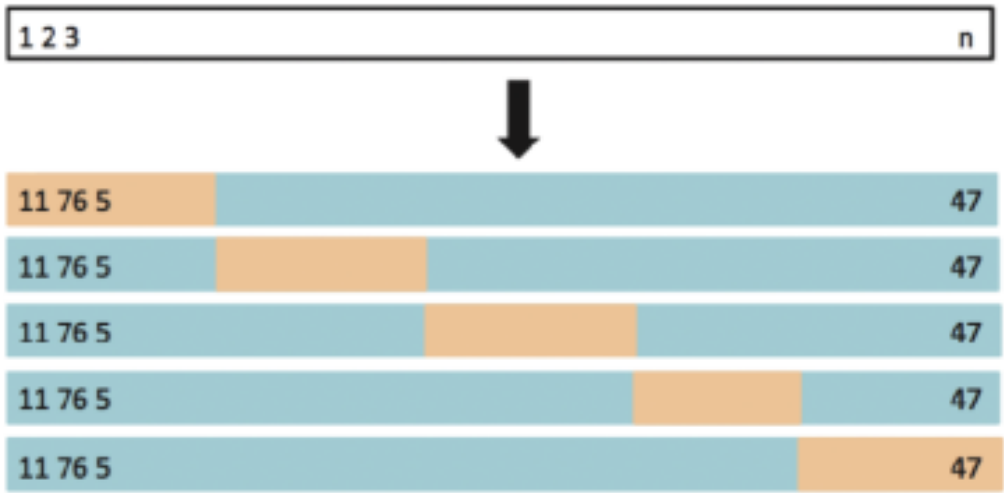
k-Fold Cross-Validation

- Randomly divide the data into k (roughly) equal-sized *folds*
- Treat the first fold as the validation set, and take the remaining folds to be the training set
- Repeat with the second fold as the validation set, and the other folds as the training set.
- Continue in this way until each fold has taken a turn as the validation set.



k-Fold Cross-Validation

- Evaluate the model by taking the average of the k validation MSEs obtained
- If $k = n$, we have *leave-one-out* cross-validation
- For larger datasets, $k = 5$ or $k = 10$ is more common



Mallow's C_p

Let p be the total number of parameters in the model. Then Mallow's C_p is defined as

$$C_p(\hat{f}) := \frac{1}{n}\mathrm{RSS}(\hat{f}) + \frac{2p\sigma^2}{n}.$$

Usually σ^2 is unknown and we replace it with a consistent estimator $\hat{\sigma}^2$.

AIC

Let $\hat{f}$ be the fitted model obtained from the MLE approach. In other words, the likelihood function is maximized at $\hat{f}$. The *Akaike information criterion* (AIC) is defined as

$$\mathrm{AIC}(\hat{f}) = -2 \log L(\hat{f}) + 2p.$$

In lecture, we said that if $\hat{f}$ is a linear model with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d., then $\mathrm{AIC}(\hat{f})$ and $C_p(\hat{f})$ select the same model.

Let's prove it!

AIC and C_p Equivalence

Recall that in a linear model, with Gaussian noise, we write

$$y = x^T \beta + \epsilon.$$

Therefore, given a dataset $D^{\text{train}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the likelihood of β is

$$L(\beta) = \frac{1}{(2\pi\sigma)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right).$$

Thus, the negative log-likelihood is (up to terms not depending on β)

$$\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2,$$

which is proportional to $\text{RSS}(\beta)$.

AIC and C_p Equivalence

This tells us that, up to constant terms, we have

$$\mathrm{AIC}(\beta) = \frac{1}{\sigma^2}\mathrm{RSS}(\beta) + 2p.$$

But if we multiply this by $\frac{\sigma^2}{n}$ we get exactly $C_p(\beta)$. Because this constant factor does not depend on β, minimizing $\mathrm{AIC}(\beta)$ and $C_p(\beta)$ will give us the same solution.

BIC

The *Bayesian information criterion* (BIC) is very similar to AIC, but applies a stronger penalty (that depends on sample size) for more complex models:

$$\mathrm{BIC}(\hat{f}) = -2 \log L(\hat{f}) + (\log n)p.$$

Note that AIC and BIC will *not* necessarily give the same solution for linear models with Gaussian noise.

Coding Example

Next, we will take a look at how to implement cross-validation and forward selection in Python. After this we will have a brief quiz.