Tutorial 2: Cross-Validation and Subset Selection

Teaching Team of STA314

Department of Statistical Sciences
University of Toronto

Monday September 16, 2024

Recall: Approaches for Model Selection

In an ideal scenario, we train a set of models on a training set, estimate their expected MSE by evaluating them on a separate test set, and then choose the model with the lowest test MSE.

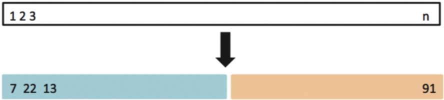
But what happens when we don't have a test set?

As we saw in Lecture 3, there are two approaches we can consider:

- 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

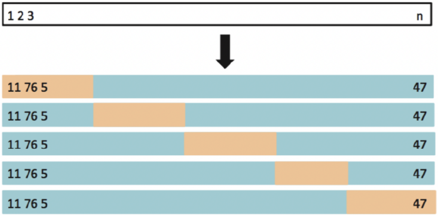
Validation Set Approach

- 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
    - And so on...
- 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

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

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 so that $L(\hat{f})$ is the maximum of the likelihood function. The *Akaike information criterion* (AIC) is

$$\mathrm{AIC}(\hat{f}) = -2\log L(\hat{f}) + 2p. \tag{1}$$

In lecture, we said that if $\hat{f}$ is a linear model with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ 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

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

Therefore, given a dataset $\mathcal{D}^{\text{train}} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, 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 RSS(β).

AIC and C_p Equivalence

So, up to constant terms, we have

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

But multiplying this by $\frac{\sigma^2}{n}$ gives us exactly $C_p(\beta)$. Since this constant factor does not depend on β, minimizing AIC and $C_p(\beta)$ give 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

For the remainder of the tutorial, let's take a look at how we implement cross-validation and forward selection.