6.4: Resampling Methods

Kamilah Ebrahim

The University of Toronto

Introduction

Resampling methods work by first drawing a sample from the training data set and then fitting the statistical model of interest on this subset of the data. This process is repeated many times which allows us to gain more information about a model without requiring more data.

This section will cover two resampling

- methods: Cross-validation
- The bootstrap

The Validation Set Approach

The **validation set approach** estimates the test error associated with fitting a statistical model to a set of observations.

- Randomly divide the set of observations into a training set and a validation set.
- Fit the model on the training set.
- Predict responses on the validation set.
- The error rate of the validation set gives an estimate of the test error rate.

Finding the error rate of the validation set depends on the statistical model.

- In the regression setting mean squared error (MSE) is usually used.
- In the classification setting, the number of misclassified observations is used.

The Validation Set Approach

There are two potential drawback to this approach:

- The validation set error rate can be highly variable since it depends on which observations are included in the training set versus the validation set.
- The validation set error rate may overestimate the test error rate since the model is being fit on a smaller training set.

Cross-validation is an extension of the validation set approach that accounts for these issues.

Exercises: The Validation Set Approach

Open the Resampling Methods R Markdown or Jupyter Notebook file.

- Go over the “Getting Started” section together as a class.
- Go over the “The Validation Set Approach” section together as a class.
- 5 minutes for students to complete the questions from “The Validation Set
- Approach”. Questions should be completed at home if time does not allow.

Leave-One-Out Cross-Validation

Leave-one-out cross-validation (LOOCV) follows the same steps as the validation set approach except the validation set is just one single observation. Then,

- Fit model on the training data set.
- Make prediction for the response of the one validation observation using the fitted
- model. Compute the validation set error.
- Repeat this process for every observation.

The LOOCV estimate of the test error is the average of all $n$of the validation set errors.

Leave-One-Out Cross-Validation

How is the LOOCV approach better than the validation set approach?

- It has less bias and does not overestimate the test error rate as much since it is using a lot more observations to train the model.
- The LOOCV result will not vary since there is no randomness in the training and validation set splits.

Exercises: Leave-One-Out Cross-Validation

Open the Resampling Methods R Markdown or Jupyter Notebook file.

- Go over the “Leave-One-Out Cross-Validation” section together as a class.
- 10 minutes for students to complete the questions from “Leave-One-Out Cross-Validation”.
- Questions should be completed at home if time does not allow.

k-Fold Cross-Validation

k-fold cross-validation involves randomly dividing the set of observations into k approximately equal sized groups. Then,

- Fit the model using the observations from all but one of the groups.
- Make predictions for the response of the observations in the remaining group. Compute the validation set error.
- Repeat this process for each group.

The k-fold cross-validation estimate of the test error is the average of the k validation set errors.

LOOCV is a special case of the k-fold cross-validation approach using $k = n$(\$n = \$ number of observations).

Comparing LOOC and k-fold Cross Validation

- The computational time/effort for k-fold CV for $k < n$ is less since we are fitting fewer models in the process.
- LOOCV is less biased in its estimation of the test error rate since it trains the model on more observations.

-LOOCV has a test error estimate that has higher variance than k-fold CV ($k < n$) the models

- in the LOOCV process are fit with nearly identical training sets
- thus, each test error result is much more correlated with one another than they would be for k-fold CV
- averaging highly correlated quantities has a higher variance than if they were not correlated

Thus, there is a bias-variance trade-off when it comes to choosing k for k-fold cross validation. Typically $k = 5$ or $k = 10$ is used.

Exercises: k-fold CV

Open the Resampling Methods R Markdown or Jupyter Notebook file.

- Go over the "k-fold CV" section together as a class.
- 5 minutes for students to complete the questions from "k-fold CV".
- Questions should be completed at home if time does not allow.

The Bootstrap

Suppose we wish to find the average of the population of Toronto μ and we have a sample of size n. We can find the mean of the sample μ_s but this does not give any indication for how this compares to the true population mean μ.

The bootstrap can be used to quantify the uncertainty of an estimate in the following way:

- Randomly sample n observations from the original sample to acquire a new sample of the same size (repeat observations are allowed).
- Compute the desired statistic (i.e. average age) of this new sample.
- Repeat steps 1-2 many times.
- Compute the standard error (SE) of the estimates.

This method is able to give us an estimate of the variability associated with our sample mean

Exercises: The Bootstrap

Open the Resampling Methods R Markdown and Jupyter Notebook file.

- Go over the “The Bootstrap” section together as a class.

References

Chapter 5 of the ISLR2 and ISLP books:

James, Gareth, et al. “Resampling Methods.” An Introduction to Statistical Learning: with Applications in R, 2nd ed., Springer, 2021.

James, Gareth, et al. “Resampling Methods.” An Introduction to Statistical Learning: with Applications in Python, Springer, 2023.