Chapter 5: Cross-Validation

Name	Content
TYPE	cheat sheet
BOOK	An Introduction to Statistical Learning
AUTHORS	Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
PUBLISHER	Springer

Resampling is repeatedly draw samples from a training set and refit the model of interest on each sample to get more information.

Two Common Methods of Resampling

1. *Cross-Validation:* how to check your model fit
   a. Can be used for model assessment & model selection

Approach	How it works	Drawbacks
Validation Set	Divides observations into training and validation set fit the model on the training set	Validation test error rate can be highly variable, may over estimate the test error rate
Leave One Out Cross Validation (LOOVC)	Splits the data set into the training set and a single validation observation.	Computationally expensive, takes a lot of time if the number of observations is large.
K-fold Cross Validation (K-fold CV)	Randomly divides observations into k folds of roughly equal sizes, the first fold is the validation set, the rest are the training set, process is repeated k times.	High variance when k is less than the number of observations

For classification problems (Y is qualitative), the validation is done on how many observations are misclassified.

2. *Bootstrap:* how to check how accurate your model (or a parameter fit) is.

- Uses the computer to emulate the process of getting new datasets.
- Repeatedly samples observations from the original dataset