Advanced Statistical Methods
Homework 5

Brandon Hosley

University of Illinois - Springfield

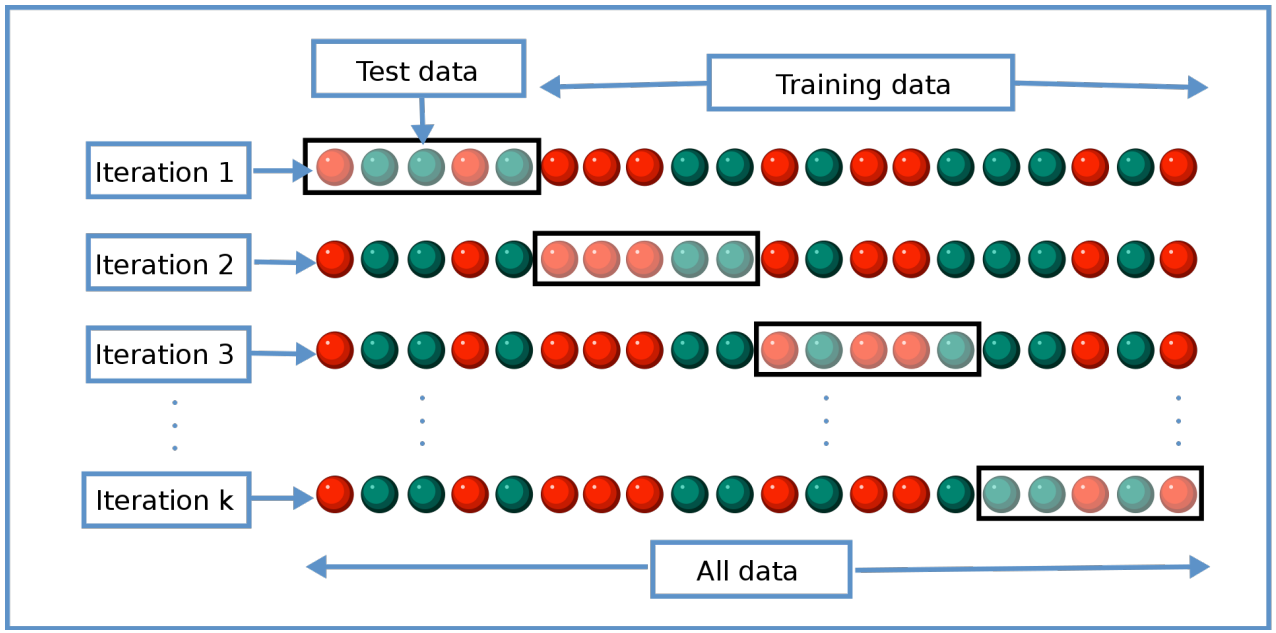
October 12, 2020

Overview

1. Q1A: Issues/mistakes with cross-validation
2. Q1B: Issues/mistakes with bootstrap?
3. Q2: Hastie and Tibshirani Summary

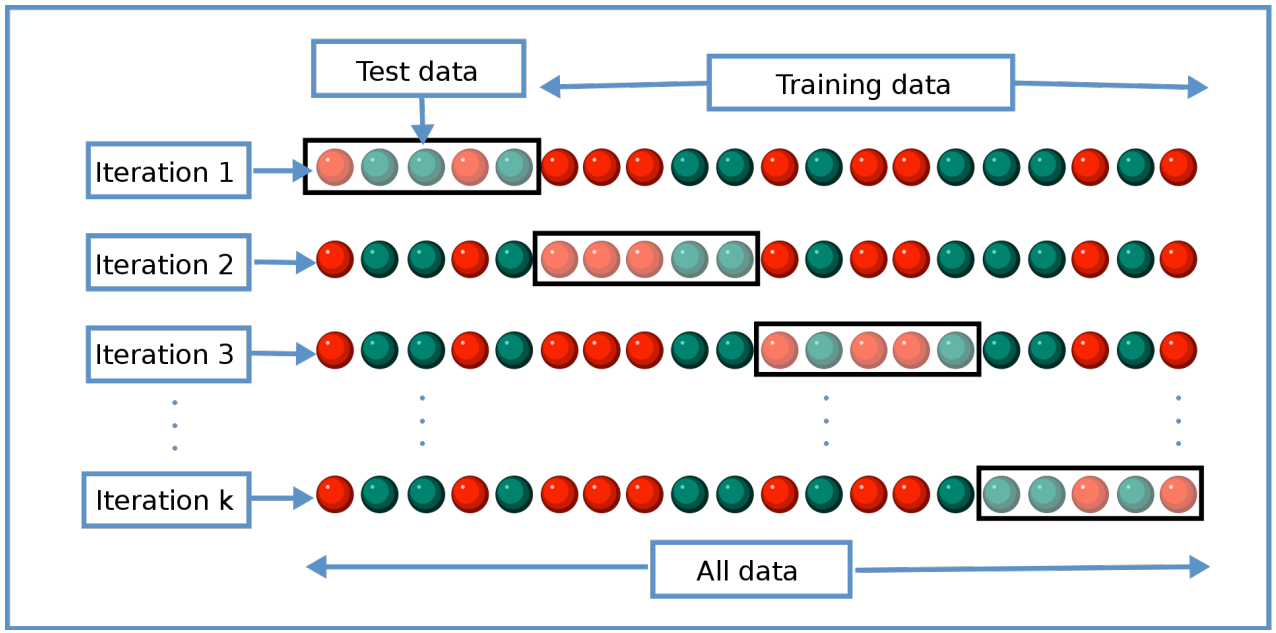
Issues/mistakes with cross-validation

- K-fold cross validation biases toward increased prediction error.



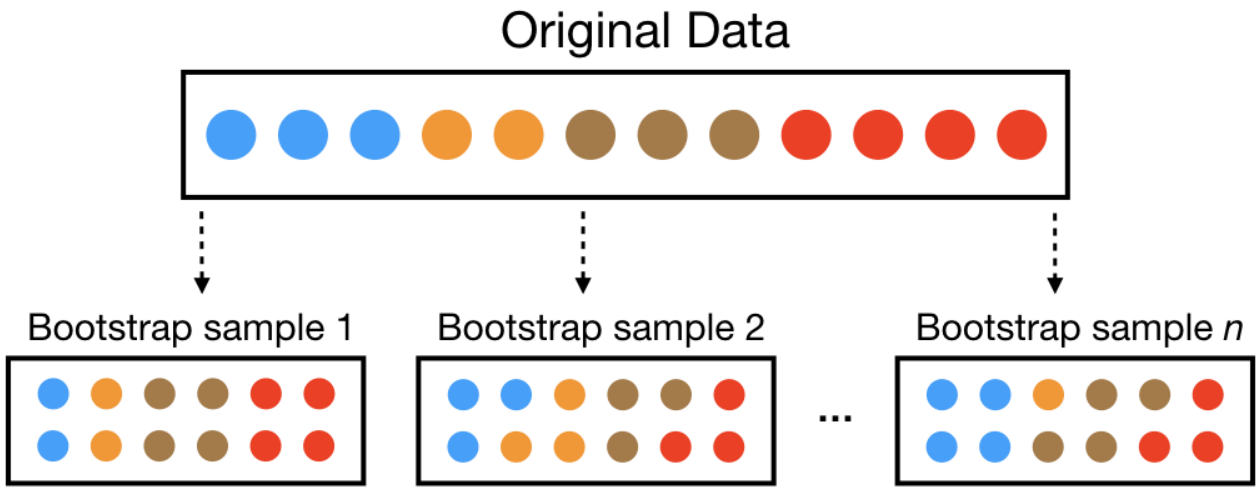
Issues/mistakes with cross-validation

- K-fold cross validation biases toward increased prediction error.
- Filtering data before placing into validation groups can cause problems with fitting;
  over-fitting to 0% training error.



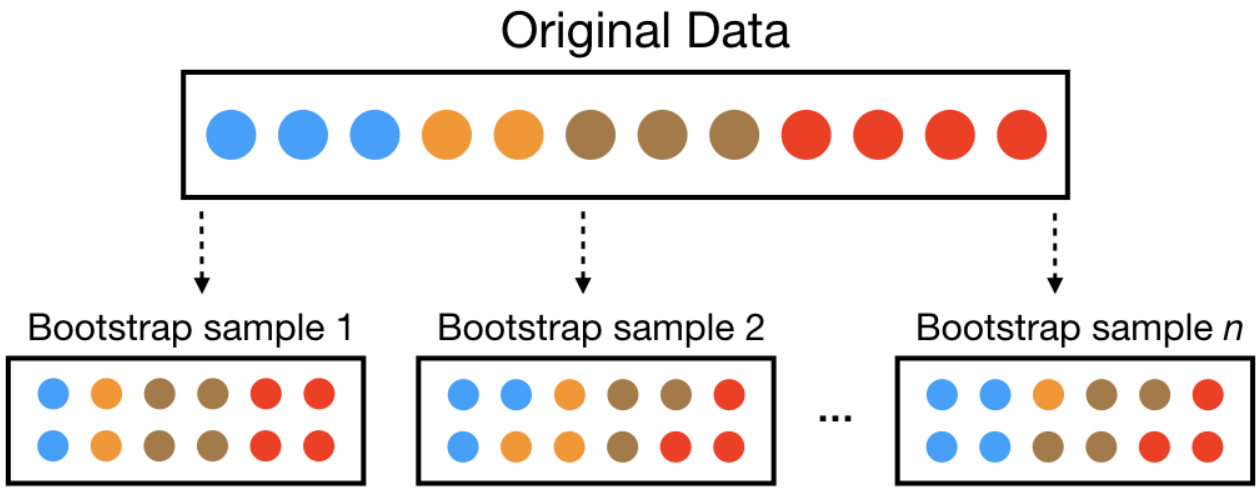
Issues/mistakes with bootstrap

- Datasets possess significant overlap.



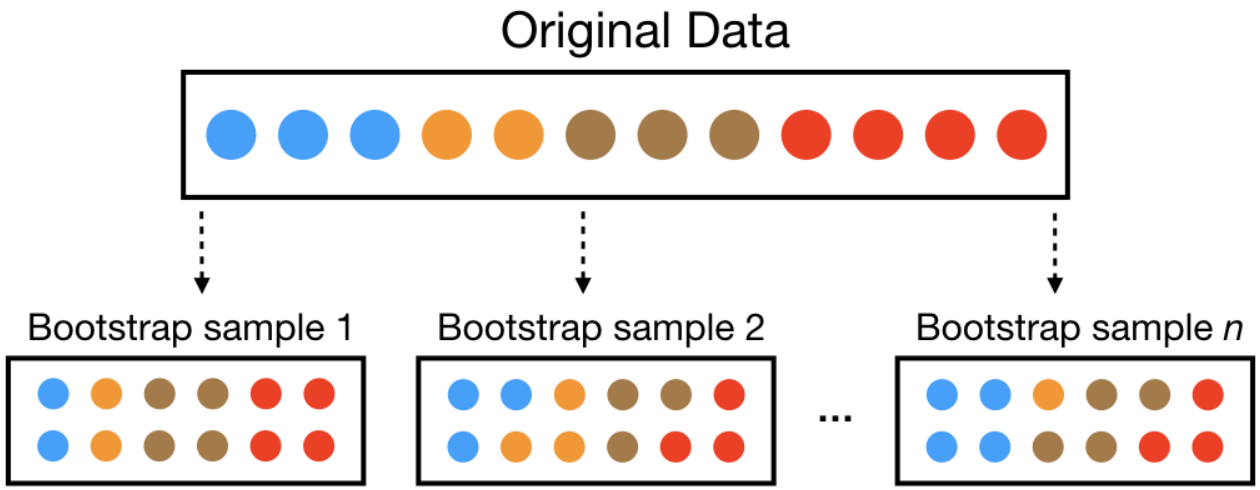
Issues/mistakes with bootstrap

- Datasets possess significant overlap.
- Severely underestimates the prediction error.



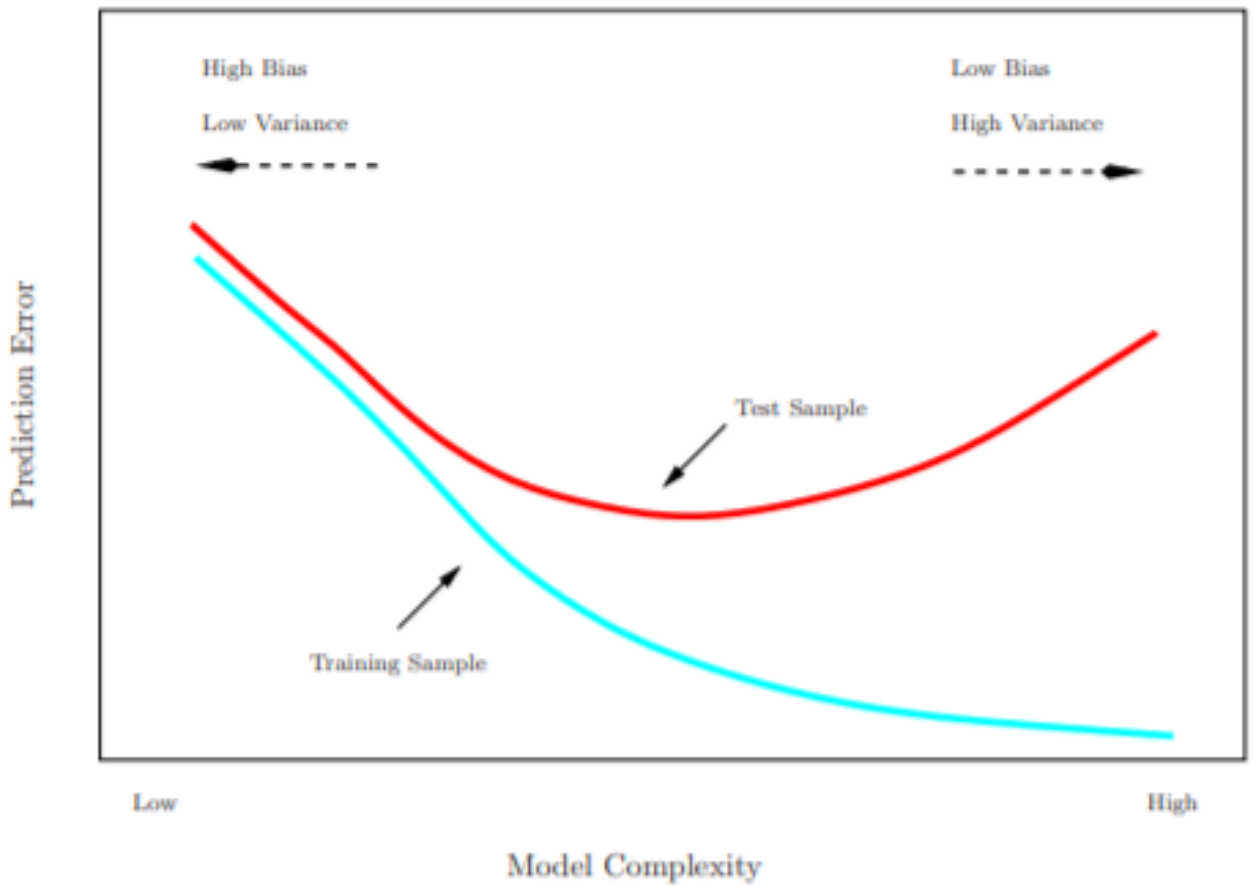
Issues/mistakes with bootstrap

- Datasets possess significant overlap.
- Severely underestimates the prediction error.
- May be improved if validated with samples that did not end up in any of the bootstrap samples.



Tibshirani Lecture: Resampling Methods

Testing the accuracy of our model.
The goal is to minimize testing error:



Tibshirani Lecture: Resampling Methods

Possible approaches:

Validation Set Random splitting of data to provide a set for testing error.

K-fold Cross Validation Splitting data into K parts, using one as the validation set, and the other $K - 1$ as training sets. Afterward select the model that provided the lowest test error.

Bootstrap Multiple data sets produced from the original by sampling from the original with some data replaced with random selections from the original data-set.

Tibshirani Lecture: Permutation v. Bootstrapping

Permutation

-

Bootstrap

-