Lecture 6:
k-Fold Cross-Validation
STAT 452, Spring 2021

Review

- Last time we discussed the **validation set approach** or **holdout method** for doing cross-validation.
- For this approach, the data is randomly split into two parts: a **training set** and **test set**.[1]
- The regression model is fit on the training set, and then the fitted model is used to make predictions for the response variable on the test set.
- The performance of the model on the test set is evaluated using the MSE, or some other metric.

[1] Note that the test set is also sometimes called the **validation set**. We'll use these terms interchangeably in this class.

Drawbacks

The holdout method is conceptually simple, but it has two potential drawbacks:

1. The test set MSE can be highly variable, depending on which observations are included in the training set and which observations are included in the test set.
   - For example, each random 70% training / 30% test set split of the data set can result in substantially different values for the MSE.

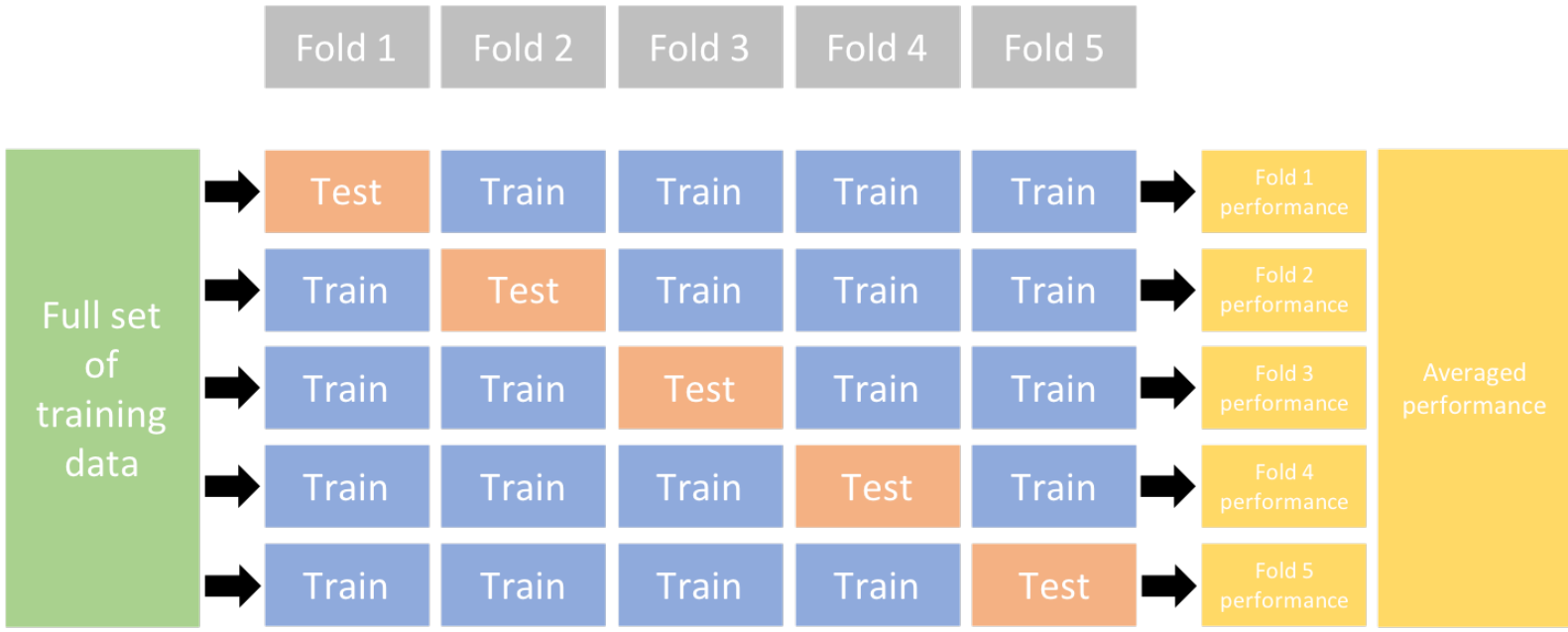
2. A relatively large portion of the data (e.g., 30%) is usually used for the test set. For smaller data sets, this might not be feasible. Statistical models also tend to perform better when using more training data.

k-fold cross-validation is a more robust approach, that overcomes these drawbacks.

k-Fold Cross-Validation (CV)

1. Randomly divide the data into k groups, or folds, of approximately equal size.
2. For $i = 1, \cdots, k$:
   (a) Hold out fold i as a validation set, and fit the model to the other $k - 1$ folds.
   (b) Calculate the mean squared error, MSE_i, on the observations in the validation set.
3. Compute the average MSE over the k folds:

$$\frac{1}{k} \sum_{i=1}^{k} \text{MSE}_k$$


Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Full set of training data
Test
Train
Train
Train
Train
Fold 1 performance
Train
Test
Train
Train
Train
Fold 2 performance
Train
Train
Test
Train
Train
Fold 3 performance
Train
Train
Train
Test
Train
Fold 4 performance
Train
Train
Train
Train
Test
Fold 5 performance
Averaged performance

Remarks

- In practice, one typically uses $k = 5$ or $k = 10$ folds.
- One advantage of k-fold CV over the holdout method is that k-fold CV uses the entire data set to compute the test set MSE.
- k-fold CV requires re-fitting the statistical model k times, and can therefore be computationally intensive. However, for linear regression, and many other statistical learning methods, this is usually not an issue (since the models run fast).
- The `train()` function from the `caret` package can be used to implement k-fold CV in R (much easier than manually coding!)