Jacob Sauther

COMP\_SCI-5565

Midterm

October 10th, 2022

**3.a: Explain how k-fold cross-validation is implemented**

K-fold cross validation is a resampling method that divides the observations into k equal groups (aka folds). One of these folds is used as a validation set and then the other k-1 folds are used to fit the data. This process is repeated over each fold so that each fold is used as a validation set at some point. Each time this process is repeated we can calculate the MSE for that k fold. The average of each MSE for each fold then will give us the test error.

**3.b: What are the advantages and disadvantages of k-fold cross-validation relative to:**

**i. The validation set approach**

The advantage to the validation set approach is that the is fewer computations required to complete. Even with the smallest k-fold cross-validation, the validation set approach is easier to implement. However, k-fold cross-validation is better with a less variable test error rate. Also, k-fold cross-validation uses all the data as both training and fitting instead of just a subset of the data. This means that k-fold cross-validation is less likely to overestimate the test error rate.

**ii. LOOCV**

LOOCV is a special case of k-fold cross-validation where k = n. Assuming k < n for k-fold cross-validation, this method is less computationally heavy than LOOCV. Though k-fold cross-validation does have slightly higher levels of bias on the data. LOOCV has much less biased estimates of the test error. It should be noted too that LOCCV has higher variance on the test error when compared to k-fold cross-validation.

**4: Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.**

To estimate the standard deviation of our prediction we need to use the bootstrap method as described in chapter 5.2. We first will use the following equation on randomly sampled n paired observations with replacements of X and Y to use as a training set. We then will use that set and a statistical learning method to get a bootstrap model. That model will then allow us to use the predictor X to estimate a response Y. We repeat this a considerable number of times to produce several bootstrap models and corresponding estimates of Y. Once this is completed the formula for standard deviation is all we would need to use to find the standard deviation of our prediction.