Assignment 3

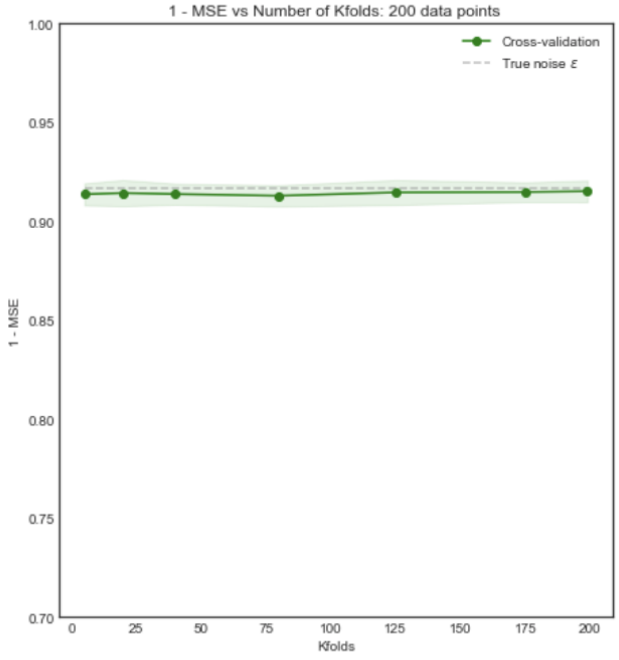
ISLR2 Ch5. Ex 3.
ESL2 Ex 7.3, 7.4, 7.5, 7.6

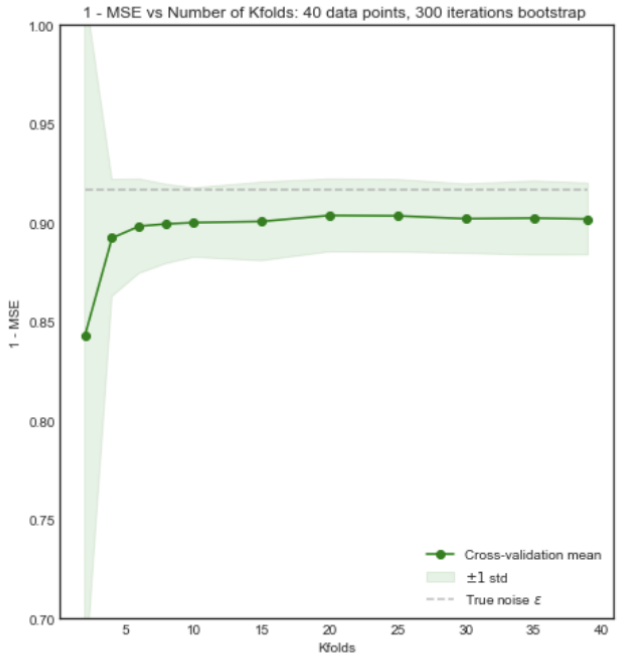
Experiment about Variance/Bias in K-fold Cross-Validation

1. Generate 10,000 points from the distribution $\sin(x) + \epsilon$ where the true variance of ϵ is known
2. Iterate i times (e.g. 100 or 200 times). At each iteration, change the dataset by resampling N points (try different Ns, e.g. 40, 200 to see the difference) from the original distribution
3. For each data set ii:
    - Perform K-fold cross validation for one value of K
    - Store the average Mean Square Error (MSE) across the K-folds
4. Once the loop over i is complete, calculate the mean and standard deviation of the MSE across the i datasets for the same value of K
5. Repeat the above steps for all K in range $\{5, \dots, N\}$ all the way to Leave One Out CV (LOOCV)

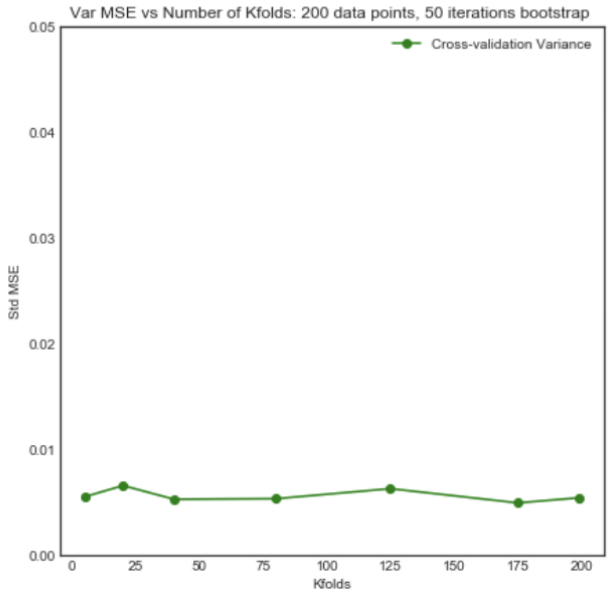
You should get plots like these

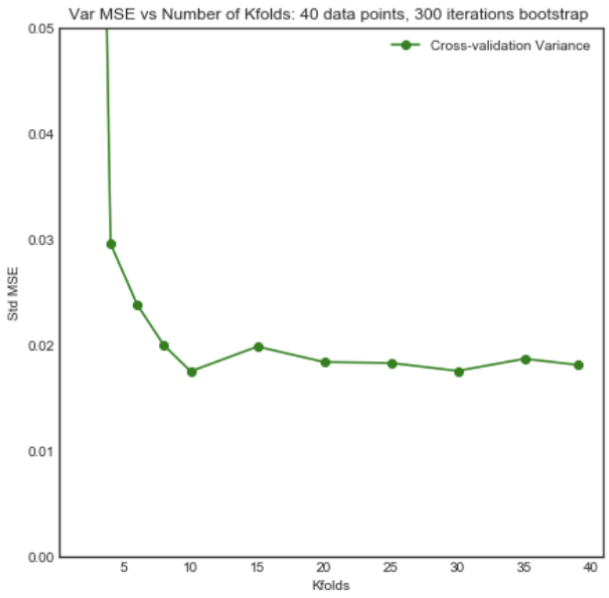
Left Hand Side: Kfolds for 200 data points, **Right Hand Side**: Kfolds for 40 data points





Standard Deviation of MSE (across data sets i) vs Kfolds





What does this experiment tell you?

Experiment about Wrong and Right Ways to do Cross-Validation

1. Generate 50 samples for 5000 standard Gaussian random variables as predictors and binary label (0 or 1 with 50/50 odds). All predictors and label are independent.
2. Find the 100 predictors having the largest correlation with the class labels.

3. Find 5-fold CV error (% of incorrectly classified cases) for 1 nearest neighbor classifier based on these 100 selected predictors.
4. Repeat step 1-3 50 times to get a distribution for CV error.
   This corresponds to the wrong way in class.
   Repeat this with the right way of doing CV.
   Since the label is independent of the predictors, any classifier should have roughly 50% error rate. What do you find in the wrong way and right way of doing CV?