DS 303 Final Exam Practice Problems

Concept Review

1. True or False? Justify your answer. When carrying out a hypothesis test, as n increases, the type 1 error will also decrease.
2. True or False? Justify your answer. When carrying out a hypothesis test, as n increases, the ability to reject the H_0 (power) will also increase.
3. Suppose your colleague fits a logistic regression model with all predictors and finds that roughly 30% of the predictors have non-significant p-values. He wants to drop those predictors and only keep all the remaining significant predictors. He asks for your advice. What do you tell him?
4. True or False? Justify your answer. Since ridge regression (with $\lambda > 0$) introduces a penalty term to protect us from overfitting, its training MSE will always be smaller than that of least square regression.
5. True or False? Justify your answer. As K increases, the KNN classifier because more flexible.
6. True or False? Justify your answer. For a given data set, suppose we want to prune a decision tree. We can directly calculate the bias and variance of a tree for different sizes. Based on this, we can choose an optimal tree size.
7. $\hat{Y}$ is an estimator for _____. Is it an unbiased estimator? Explain.
8. Explain what (if anything) happens to a multiple linear regression model ($\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots \hat{\beta}_p X_p$) under the following scenarios:
   (a) We have a dataset where we have redundant information among the predictors.
   (b) The response Y does not follow a normal distribution.
   (c) The assumption $E(\epsilon_i) = 0$ does not hold.
9. Suppose you have a dataset with $n = 1000$ observations and $p = 10,000$ predictors. This is call the high-dimensional setting. For each technique, explain whether it is affected by the curse of dimensionality.
   (a) Boosting
   (b) KNN classification
   (c) LDA
   (d) K-means clustering
   (e) Ridge regression
10. Your colleague (who is inexperienced at machine learning) states that all your models are data-dependent. That means if you had used a different training set, your trained model would look different: your predictions for Y would change. This worries your colleague - how do they know which model is the 'right' one? What tools do you have at your disposal to address their concerns? Explain.

11. In a well known competition for the Netflix, many teams decided to merge together before the competition finished for the one million dollar prize. Explain briefly what strategy the teams likely used to combine their algorithms, and why one can expect that this strategy will improve the final prediction results.

Coding

1. Design a simulation study to calculate the variance of random forest as a function of m. Create a plot that shows that as m decreases, the variance of the bagged model decreases as well. Use simple statistical reasoning to explain why we see a reduction in variance as m decreases. You can use the same setup as HW 11, Problem 3.

2. We will use the `NCI60` cancer cell line microarray data from `library(ISLR2)`. This is also part of the `ISLR2` library. Carry out the following pre-processing:

```
nci.labs <- NCI60$labs
nci.data <- NCI60$data
```

   (a) Should we scale our features, which are gene expressions, in this setting? Justify your answer. If you decide to scale the features, do so.

   (b) Implement K-means clustering on the (possibly scaled) data. Experiment with $K = 2$ and $K = 4$. Report the total within-cluster sum of squares for both $K = 2$ and $K = 4$.

   (c) Implement hierarchical clustering with complete linkage and Euclidean distance on the (possibly scaled) data. Cut the dendrogram to obtain 2 clusters. How does this compare to the K-means results we obtained in part (b) for $K = 2$? Report a confusion matrix to compare the results.

   (d) Discuss how you might validate your clustering results.

3. Implement PCA on the `USArrests` dataset. The proportion of variance explained by the first two principal components is 87%. Bootstrap the standard error for this quantity.