**ISQA 8080 Assignment 2 Due: By Tuesday, Oct. 8 2019, 5:30 PM**

**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**NOTES:**

1. Use R for the calculations and implementation.
2. Submit all documents in a zip file and upload it to Canvas. Name your Zip Folder with your name, A2, and the course # (Example: LastName-A2-ISQA 8080).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Classification – Conceptual (15 points)**

When we talked about the k-nearest-neighbor (KNN) algorithm, we discussed that when the number of features p is large, there tends to be a deterioration in the performance of KNN (and, in fact, other local approaches) that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that parametric approaches often perform poorly when p is large. We will now investigate this curse.

1. Suppose that we have a set of observations, 1 to n, each with measurements on p = 1 feature, X (i.e., we have one predictor and one target variable). We assume that X is uniformly (evenly) distributed on [0, 1]. Associated with each observation is a response value Y. Suppose that we wish to predict a test observation’s response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with X = 0.6, we will use observations in the range [0.55, 0.65]. On average, what fraction of the available observations will we use to make the prediction?
2. Now suppose that we have a set of observations, each with measurements on p = 2 features, X1 and X2. We assume that (X1,X2) are uniformly distributed on [0, 1] × [0, 1]. We wish to predict a test observation’s response using only observations that are within 10% of the range of X1 and within 10% of the range of X2 closest to that test observation. For instance, in order to predict the response for a test observation with X1 = 0.6 and X2 = 0.35, we will use observations in the range [0.55, 0.65] for X1 and in the range [0.3, 0.4] for X2. On average, what fraction of the available observations will we use to make the prediction.
3. Generalize the previous argument by studying p predictors, where p is large (e.g., p = 100). Which fraction of overall observations will be used for prediction?
4. Given the previous results, discuss the issue(s) that you see when p gets large. For example, look at the fraction of observations that is used in the predictions.
5. **Resampling – Conceptual (20 points)**

In chapter 5 of the book, we introduced different resampling methods that allow us to get an estimate of the prediction performance on new data. Answer following questions in your own words (note: visualizations could be helpful for explanations).

1. Explain how k-fold cross-validation is implemented.
2. What are the advantages and disadvantages of k-fold cross-validation relative to:

i. The validation set approach?

ii. LOOCV?

1. Assume that you have a Machine Learning algorithm that requires you to set a parameter, e.g., the number of predictors to include. Setting this parameter will affect your model performance, e.g., measured by RMSE or Accuracy.

You want to find the optimal value of the parameter by using resampling techniques. Compare following two approaches:

1) you use cross-validation on the entire dataset and select the best model, and 2) you split your dataset into 80% training and 20% test, further run cross-validation on the training set, select the best model based on cross validation, and then use the 20% test set to determine the performance of the selected model.

What advantage(s) do you see in the first approach, what in the second?

1. **Classification – Comparing different methods (65 points)**

In this problem, you will develop a model to predict whether a person in the US Census earns more than $50K or not. Consider Income as the target variable and include Age, MaritalStatus, Race, Sex, and WeeklyHours as predictors.

1. Explore the data graphically in order to investigate the association between the Income variable (target) and the predictor variables. Which predictor variable(s) seem to be most relevant for predicting the target variable? Describe your findings.
2. Split the data into a training set and a validation/test set using a 70%/30% split. Run the summary() function on each set and show the results. Do the training and test sets seem sufficiently similar?
3. Create a Logistic Regression model on the training data to predict the income (>$50K or <$50K). Use cross-validation on the training set. Calculate and show the confusion matrix for both the training and the test set. What is the performance with respect to accuracy, sensitivity, and specificity, and ROC? Create and print the ROC curve.
4. Now, perform LDA on the training. Use cross-validation on the training set. Calculate and show the confusion matrix for both the training and the test set. What is the performance with respect to accuracy, sensitivity, and specificity, and ROC? Create and print the ROC curve.
5. Next, perform QDA on the training data. Use cross-validation on the training set. As before, calculate and show the confusion matrix for both the training and the test set. What is the performance with respect to accuracy, sensitivity, and specificity, and ROC? Create and print the ROC curve.
6. Finally, perform KNN on the training data, with several values of K (e.g., 3, 5, 7). Use cross-validation on the training set. Calculate and show the confusion matrix for both the training and the test set. What is the performance with respect to accuracy, sensitivity, and specificity, and ROC? Which value of K seems to perform the best on this data set?
7. What would be your recommendation for the best model, based on the previous results **for the training set**? Which factors / metrics did you take into account for your recommendation (you are not restricted to accuracy)?
8. If you compare the previous recommendation with the test set performance of the different approaches, would you have chosen the same algorithm?