**ISQA 8080 Assignment 4 Due: By Tuesday, Nov. 25 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, A4, and the course # (Example: LastName-A4-ISQA 8080).

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

1. **Support Vector Machines**

As the last main type of classification approach, you will build different types of support vector machines for diagnosing malignant cancer.

Notes:

* This data set is different from the ones’ before as you will see very high values of accuracy, sensitivity, and specificity. As a typical case of medical diagnostics, we want to have both a high sensitivity and a high specificity. You might not see large differences between the kernels, but each additional correctly classified patient makes a difference in this case.
* The variable to predict is “class”, and it is either Malignant or Benign. The remaining variables are medical measurements for the specific patient and tumor.
  1. Maximum Margin Classifiers, Support Vector Classifiers, and Support Vector Machines are all based on the concept of hyperplanes. **In your own words**, describe what a hyperplane is and how it can help us to classify observations into 2 categories.

-In this context, a hyperplane consists of a subspace with its number of dimensions determined by 1 - # of predictors. Its purpose is to optimally separate positive and negative training observations in a given data-set, thereby allowing for classifications of test observations into one of the two groups based on which side of the hyperplane it resides.

* 1. In your own words, what is the difference between Maximum Margin Classifiers, Support Vector Classifiers, and Support Vector Machines?

Note: illustrations can be helpful to explain the differences (if you decide to use visualizations, use your own visualizations and not just copied images).

Maximum Margin Classifiers: Can only be applied if there is perfect separation of positive and negative observations within the training set. In this case, the hyperplane splits the separation of the positive and negative classes so that the margin, or minimal distance to the observations, is maximized (maximum margin).

Support Vector Classifiers: When perfect separation between classes is not possible, a support vector classifier can be used instead. Unlike maximum margin classifiers, support vector classifiers permit certain observations to violate, or have a shorter distance, than the margin, or even fall on the wrong side of the hyperplane. The number of observations, along with their degree of violation, are controlled by the tuning parameter ***C***, which can be adjusted to balance the bias-variance trade-off.

In cases where the hyperplane dividing the two classes is not linear, support vector machines can be used to generate non-linear decision boundaries for classes by implementing various polynomial, radial, or otherwise non-linear kernels for quantifying the similarity between training observations. According to the textbook, a given kernel will be applied to the inner products for all pairs of training observations, providing for a computationally efficient way to enlarge the feature space and drastically improve the accuracy of the classifications in ways linear hyperplanes cannot.

* 1. Now, let us build SVM models for the cancer data set. Start by splitting the data into 70% / 30% training and test set. Use a unique random seed to create the split (you can use set.seed(sample(10000,1)) for this).

For the first model, build a Support Vector Machine with linear kernel on the training set (using cross-validation). Use the default parameters (e.g., default Cost parameter) for this. Show the performance of the model on the test data using the confusionMatrix() function. Also calculate the AUC (Area under the Curve) on the test data.

* 1. Building on part c, let’s find a good value for the Cost parameter C. Use the tune() function (on the training dataset) with k=10 folds to try different values for the Cost parameter. For example, use cost = c(0.001,0.01,0.1,0.5,1,5,10,100). Select the best model and calculate the confusionMatrix for the test set. Also calculate the AUC (Area under the Curve). How does it compare to the model in part c?
  2. Now, let us try different decision boundaries. Select a radial kernel and try different parameters for your svm model (e.g., different cost and gamma parameters). Select the best model and calculate the confusionMatrix for the test set. Do you see improvements over the model in part d? Look at the metrics in the confusion matrix (accuracy, sensitivity, specificity), as well as AUC.

1. **Advanced Regression Models – Shrinkage Methods**

Instead of classification models, let us use some advanced regression models that help us to build a prediction model with good performance on the test set.

Specifically, we will use the Ames Housing dataset to predict the sales price of houses in Ames, Iowa. The data and description is available as part of this assignment.

1. Both Ridge Regression and LASSO are examples of Shrinkage or Regularization methods. In your own words, compare the two approaches against standard regression that uses Least Squares to find the beta parameters. What do you have to do to convert Ridge Regression or LASSO to Least Squares Regression?
2. How do Ridge Regression and LASSO differ from each other?
3. What are some of the advantages of LASSO compared to standard regression using OLS (Ordinary Least Squares)?
4. Now, let’s start by building a Ridge Regression (RR) for predicting the Sales Price of a house. Split the data into a training and test set (70% / 30%). Use k-fold cross validation to select the best lambda parameter for the given training set. Note: consider using larger values of lambda, e.g., in the range 1 – 100,000. You should eventually see a U-shaped RMSE curve when you plot the RMSEs for different lambdas, so experiment with the ranges that works best for you.

Use this parameter to predict the house prices (SalePrice) for the test dataset. What is the corresponding Test MSE?

1. Use the same training/test split, only now you build a LASSO regression. Again use k-fold cross validation to select the best lambda parameter for the given training set (same reasoning for lambda range as before). What is the corresponding Test MSE for the house prices? Also, which variables are selected to be included in the LASSO model (i.e., which variables have non-zero beta coefficients)?
2. Finally, let’s compare the previous two models (RR, LASSO) against a full linear regression using all variables and OLS (Ordinary Least Squares). What is the corresponding test MSE (note: ignore potential warning messages)?
3. Last but not least: Compare the RR, LASSO, and OLS results. Which one would you prefer overall? Use both the test MSE and the number of variables used in the final model for the comparison and decision.