Homework 3

Directions: Complete all exercises.

1. We have seen that as the number of features used in a model increase, the training error will necessarily decrease, but the test error may not. Let's examine this in simulation.

   (a) Generate a data set with $p = 25$ features, $n = 1,000$ observations, and an associated quantitative response vector generated according to the model: $Y = X\beta + \epsilon$, where β has some elements that are exactly equal to zero. (be sure to use "set.seed")

   (b) Split your data set into a training set containing 500 observations and a test set containing 500 observations.

   (c) Perform subset selection (best, forward or backwards) on the training set, and plot the training and test MSE associated with the best model of each size.

   (d) For which model size does the test set MSE take on its minimum value? Comment on your results. If it takes on its minimum value for a model containing only an intercept a model containing all the features, then play around with the way that you are generating the data in (a) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size

   (e) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient values.

   (f) Create a plot containing $\sqrt{\sum_{i=1}^{p}(\beta_j - \hat{\beta}_j^r)^2}$ for a range of values r, where $\hat{\beta}_j^r$ is the jth coefficient estimate for the best model containing r coefficient estimate for the best model containing r coefficients. Comment on what you observe. How do these results compare to part D.

2. Consider the Diabetes dataset (posted with assignment). Assume the population prior probabilities are estimated using the relative frequencies of the classes in the data.

   (a) Produce pairwise scatterplots for all five variables, with different symbols or colors representing the three different classes. Do you see any evidence that the classes may have difference covariance matrices? That they may not be multivariate normal?

   (b) Apply linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). How does the performance of QDA compare to that of LDA in this case?

   (c) Suppose an individual has (glucose test/intolerence= 68, insulin test=122, SSPG = 544. Relative weight = 1.86, fasting plasma glucose = 184). To which class does LDA assign this individual? To which class does QDA?

   (d) Apply RDA (regularized discriminant analysis). What is the optimal value of α in this case? Does this support your observations about the covariance matrices in (a).

3. This problem concerns the Boston data set (ISLR2 package).

   (a) Fit classification models in order to predict whether a given census tract has a high or low crime rates. Explore logistic regression, LDA, QDA and KNN models using various subsets of the predictors. Describe your findings.

(b) Fit classification models in order to predict whether a given census tract has a high, medium or low crime rates. Explore logistic regression, LDA, QDA, and KNN models using various subsets of the predictors. Describe your findings.

(c) Reflect on the results from (a) and (b). Is this within your expectation, why or why not?