STAT 8330: **Homework 6**

Due on November 10, 2020

1. OJ analysis: Consider the `OJ` data set that is part of the `ISLR` package. This data set has information concerning purchases of two brands of orange juice (Minute Maid (MM) and Citrus Hill (CH)). We are interested in trying to classify an individual into one of these categories (variable `Purchase`) given the other variables in this data set. Using `set.seed(1)`, split the data into a training set (of size 800), with the remaining observations in the test set.

   (a) Fit a support vector classifier to the training set using the `tune` function to select an optimal cost for values in the range $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$. Report the training and test error rates for the model identified as best from the tuning.

   (b) Fit a polynomial SVM of degree two, using the `tune` function to select an optimal cost for values in the range $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$. Report the training and test error rates for the model identified as best from the tuning.

   (c) Same as (2) except fit a radial basis kernel SVM, tuning with the same values of the cost as above, and potential values for the γ parameter, $\{0.001, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100\}$. Report the training and test error rates for the model identified as best from the tuning.

   (d) Same as (3) except fit a neural net ("sigmoid") kernel SVM, tuning with the same values of the cost and gamma as in (3). Report the training and test error rates for the model identified as best from the tuning.

   (e) Fit a logistic regression classifier (using the `glm` command as in the lab from Chapter 4) to these data and report the training and test error rates.

   (f) Plot an ROC curve for the training and test data (on two separate plots) for the best model identified in each of problem (1)- (5). Discuss which procedure works the best in terms of AUC.

   (g) Using the R package `glmnet` (see the package help on the CRAN website) or find some other package to fit an L1 (lasso) and L2 (ridge) penalized logistic regression classifier to these data. Report the training and test error rates for each. How do these compare to the other methods considered in this HW?

2. Extra Credit: Consider the training and test R data sets linked on the Canvas site, `dig_train` and `dig_test`. The training set here is a portion of the famous MNIST handwritten digit dataset that has been used as a test of many learning algorithms. The training set has 5000 cases and 784 features (the file is 5000×785, with the last column the labels (0-9)). The test set has 10,000 cases and the same column format. The size of the data set can be a problem!!

   (a) Build a random forest classifier using the `RandomForest` package (or some other package). Report your R-code that implements your model and your final error rate on the test data.

   (b) Build a deep neural net classifier for these data. Report your R-code that implements your model and your final error rate on the test data.