STAT 621 HOMEWORK 7

Due: Monday November 11

Problem 1: The data set `HeartDisease.txt` contains data on 462 patients. Patients were classified as having experienced coronary heart disease (`chd=1`) or not (`chd=0`). Nine additional variables were measured on each patient including systolic blood pressure (`sbp`), LDL cholestero (`ldl`), a body mass medasurement `adiposity` and others. In this question you will try to predict coronary heart disease from the other variables.

1. First create a training data set with approximately 75% of the data, and a test set with the remaining 25%. Use the training set to construct the following classifiers. Use each classifier to predict heart disease on the training set. Report the confusion matrix for each method. Is there a clear winner?
   (a) Logistic Regression
   (b) Naive Bayes
   (c) Linear Discriminant Analysis
   (d) Support Vector Machine
2. Plot ROC curves for each of the classifiers above as evaluted on the test data. Compute the AUC (Area Under the Curve) for each classifier. Which is best in terms of AUC?
3. Do 5-fold crossvalidation to select the best classifier. Report the average accuracies for each method.

Problem 2: The data set `skulls.txt` contains measurements on 30 Egyptian skulls from each of 5 eras. The eras are Early Predynastic (`EP`), Late Predynastic (`LP`), 12th and 13th Dynasties (`TTD`), Ptolemaic Period (`PP`) and Roman Period (`RP`). Four measurements were taken on each skull: X_1 = maximum bredth, X_2 = basibregmatic height, X_3 = basialveolar length and X_4 = nasal height.

1. Fit the Naive Bayes model to a training set to classify the `Era` response on a test set. Summarize preditions with plots and a table.
2. Do the same using the LDA model.
3. Use 5-fold crossvalidation to compare prediction accuracy between the two methods.

Problem 3: Now consider using SVM to classify skulls in terms of `Era`.

1. Find the best SVM model using a linear kernel (so the support vector classifier). Use the `tune` funciton to find the optimal value of the Cost parameter.

2. Find the best SVM model with a polynomial kernel. Find the optimal values of degree and Cost using the `tune` function.

3. Repeat for a SVM modle using radial kernel, optimizing over gamma and Cost.

4. Use 5-fold crossvalidation to compare prediction accuracy among the three kernels.