Statistical Data Mining I
Final Homework

Due: Sunday December 15th (11:59 pm)
30 points

Directions: Complete two exercises. A third may be done for extra credit.

1) (15 points) The Cleveland heart-disease study was conducted by the Cleveland Clinic Foundation. The response variable is "diag" (diagnosis of heart disease: bugg = healthy, sick = heart disease). There were 303 patients in the study, and 13 predictive variables, including age, gender, and a range of biological measurements.

   Fit a neural network, CART model and a random forest to the Cleveland heart-disease data. Compare the results, and comment on the performance.

2) (15 points) A pen-based handwritten digit recognition (pendigits) was obtained from 44 writers, each of whom handwrote 250 examples of the digits 0,10,2,….,9 in a random order. The raw data consists of N=10,992 handwritten digits extracted from tablet coordinates of the pen at fixed time intervals.

   a) Compute the variance of each of the 16 variables and show that they are very similar. How many PCs explain 80% and 90% of the total variation of the data? Display biplots for the first few PCs, color the plots by class (digit). Create a three-dimensional score plot for PC1, PC2 and PC3, color the samples by class.

   b) Divide the data into test and training. Fit a kNN model over a range of "k" to the (a) raw data, and (b) PCs from part (A) that capture at least 80% of the variation. Comment on your results.

   Extra Credit (15 points):

   a) (2 pts) Explain how cross validation is implemented.
   b) (4 pts) What are the advantages and disadvantages of k-fold cross validation relative to:
      i) The holdout method (division of the data into test and training).
      ii) LOOCV?
   c) (9 points) Write an *.R **function** to implement k-fold cross validation. Apply it to a data set of your choice. Use it to compare the results of 10-fold, 5-fold and LOOCV. Submit your fully commented function, as well as the application of the dataset that you selected.