36-350: Data Mining

Homework 12
Date: November 30, 2001 **Due: start of class December 7, 2001**

---

1. (Projection) Sonar works by broadcasting a well-chosen sound wave and analyzing the frequency content of the echo. A useful task is to discriminate the echo of metal objects from those of rocks. The files `Sonar-tr.dat` and `Sonar-te.dat` contain 208 sonar echoes, some of which came from metal cylinders and others from rocks (the `Class` variable indicates which). Each echo is represented by the energy in 40 different frequency bands.

   (a) Use projection to visualize the boundary between the classes. Argue whether you expect a logistic regression or nearest neighbor classifier to perform better. Remember that some projections hide information that other projections reveal.

   (b) Train a logistic regression classifier and examine the fitted boundary using `cplot.project.glm`. How many regression coefficients are there, relative to the number of training points?

   (c) Make a cross-validated nearest neighbor classifier and compare its test set performance to the logistic regression classifier. Do the results agree with what you found in part (a)?

   (d) Use the results of the last three parts to argue that the logistic regression classifier is overfitting.

2. (PCA) The file `Boston-small.dat` contains a reduced version of the Boston housing data. To reduce confusion, the variable `black`, which indicates a non-black neighborhood, has been renamed to `white`.

   (a) Use Principal Components Analysis, as shown in class, to display the main variations among houses. Treat `Price` not as a response but just another variable. The plot will contain a lot of information and needs to be disgested slowly.

   (b) Describe the different housing groups; there are at least four of them, starting with the high-crime group and working counterclockwise. Speculate on what they correspond to in real life.

   (c) What variable correlations and anti-correlations are revealed by the plot? Do they make sense? It may help to make scatterplots.

3. (Clustering) Use Ward's method to cluster the Boston housing data. Pick an interesting number of clusters and show the cluster means with a star plot. Describe the clusters and what they might represent in real life. Your answer need not be the same as the last problem.