36-350: Data Mining

Homework 8
Date: October 26, 2001 **Due: start of class November 2, 2001**

1. Hans and Frans each train classifiers on a training set of size 500. Hans has a classifier which makes 25 errors on the training set. He tests it on 100 new points and it makes 5 errors. Frans has a classifier which makes zero errors on the training set. He tests it on 200 new points and it makes 7 errors. Frans says his classifier is better. Hans says the test sets aren't big enough to tell. Who is right?

2. In the next three problems, you will use classification trees to predict the onset of diabetes. The files `Pima-tr.dat` and `Pima-te.dat` contain a training set and test set, respectively, collected by the US National Institute of Diabetes and Digestive and Kidney Diseases.

   (a) Train a classification tree on the training set to predict the `type` variable (diabetic or not) from the two predictors `glu` and `ped`.

   (b) Make a `cplot`. Some of the splits separate regions which make the same decision ("Yes" or "No"). Do these splits have any effect on the misclassification rate of the tree? Why are these splits made?

   (c) Use cross-validation to determine how large the tree should be. Use a division into 50 blocks and plot the expected misclassification rate for each tree size.

   (d) Make a `cplot` of the best-sized tree that results from pruning the original tree. How does the probability of diabetes vary with `glu` and `ped`?

3. Suppose it is three times more costly to make a false negative classification ("No" when truth is "Yes") than a false positive classification. A patient has `glu = 100, ped = 0.5`. Using the tree in problem 2(a), which classification of the patient has minimum expected cost?

4. (a) Train a classification tree on the training set to predict the `type` variable from all other variables. The tree should be fairly complex (20 leaves).

   (b) Use 50-block cross-validation to determine how large the tree should be. Plot the expected misclassification rate for each tree size.

   (c) What is the difference between using a division into 2 blocks versus 50 blocks? Hint: try running it multiple times.

   (d) Compute the misclassification rate of the best-sized tree and the original tree on the test set. Which is better?

   (e) Using the method of problem 1, is this difference statistically significant?

5. A data miner trains a classification tree on a dataset with 64 predictors. The tree ends up using 3 different predictor variables and has high accuracy on a holdout set. Does this mean that the classes do not differ significantly with respect to the other predictors? Explain.