

Chemistry 696D – Analytical Informatics

Exam 1 – Due Monday February 17, 2003

Training data sets A1,A2, B and C are on the CD. (TrainA1.dat, TrainA2, TrainB.dat, TrainC.dat)  
There are also three Test sets (A, B and C) on the CD (TestA.dat, TestB.dat, TestC.dat)

Training sets have class information, Test sets do not.

Turn in all your work: plots, calculations, source code and explanations.

1. Describe when a linear classifier is Bayes optimal.
2. Describe what is meant by a parametric vs. a non-parametric approach to building a classifier.
3. Using data set A, train a linear classifier using a response matrix approach. Obtain the following:
  - a. Train on TrainA1.dat, and compare the errors obtained on TrainA2.dat, plot the data, the decision line; provide the parameters to the decision model; all source code or calculations.
  - b. Update the training set with TrainA2.dat and plot the new classifier. Describe the new classifier in terms of the new parameters – provide data and decision plots.
  - c. Classify the Test set (TestA.dat) with the best linear classifier you can develop from the data.
4. Repeat part 3 using a knn classifier
  - a. Report the classification error rate on the training set (TrainA1.dat) from  $k=1$  to 30
  - b. Report the error on the training set A2 with the final classifier and  $k$  value chosen from the training set A1.
  - c. Update the classifier with TrainA2.dat and give the new parameters, showing the new  $k=1$  to 30 classification error plot.
  - d. Classify the Test set (TestA.dat) with the best knn classifier you can develop from the data.
5. Classify Test Set B (TestB.dat) using Training Set B (TrainB.dat). Fully describe your classifier and defend your design choices.
6. Classify Test Set C (TestC.dat) using Training Set C (TrainC.dat). Fully describe your classifier and defend your design choices.