ECEN 649 Pattern Recognition – Spring 2016

Computer Project 2

Due on: April 25

This computer project will consist of the application of classifier design, error estimation, and feature selection techniques to a gene expression data set. The data come from the following cancer classification study:

van de Vijver,M.J., He,Y.D., van't Veer,L.J., et al. (2002), "A gene-expression signature as a predictor of survival in breast cancer." *New Eng. J. Med.*, 347, 1999-2009.

This paper analyzes gene expression in breast tumor biopsies from 295 patients. Of the 295 patients, 115 belong to the "good-prognosis" class, whereas the remaining 180 belong to the "poor-prognosis" class. The expression data corresponds to a previously-published 70-gene signature.

For the purposes of this project, the gene expression data was randomly divided into a training data set (containing 60 patients) and testing data set (containing 235 patients). The latter will be used for test-set error estimation of the true classification error. The proportion of good and poor prognosis patients was kept approximately the same in the training and testing data. The data data can be retrieved in tab-delimited text format from the following links (opening these files in a spreadsheet program such as Microsoft Excel allows for good visualization of the data).

- Training data set: `http://www.ece.tamu.edu/~ulisses/ECEN649/Training_Data.txt`
- Testing data set: `http://www.ece.tamu.edu/~ulisses/ECEN649/Testing_Data.txt`

Note that the first row contains the gene symbol names, whereas the first column contains the patient ID. The last column contains the label (1 = good prognosis, 0 = poor prognosis).

We are going to search for gene feature sets that best discriminate the two prognosis classes on the training data, for a given classification rule, different error estimation criteria and feature selection methods.

We will consider the following classification rules:

- DLDA
- 3NN

The criterion for the search will be simply the resubstitution error estimate of the designed classifier for the current feature set (wrapper feature selection).

Finally, we will employ two simple feature selection methods:

- exhaustive search (for 1 to 3 genes)
- sequential forward search (for 1 to 8 genes)

Therefore, each person will determine 11 gene sets for each of the 2 classification rules, for a total of 22 gene sets.

Each person will submit a report containing the code and a table with the gene sets found. For each row of the table (gene set found), the corresponding error estimate and the test-set estimate of the true classification error should be indicated. There should be a section at the end with your conclusions. In particular, here are examples of questions you should address:

- How do you compare the error estimators and feature selection methods used based on the gene sets found and the estimates of the true error?
- How do you think the results might change if there were more training samples available or different classification rules?