ECEN 649 Pattern Recognition – Spring 2018

Computer Project 2

Due on: May 3

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. The authors performed feature selection to obtain 70 genes; hence, the full data matrix is 70×295. This is a **retrospective study**, meaning that the patients were tracked over the years and their outcomes recorded. Using this clinical information, the authors labeled the tumor samples into two classes: the "good' prognosis" group were disease-free for at least five years after first treatment, whereas the "bad prognosis" group developed distant methatasis within the first five years. Of the 295 patients, 216 belong to the "good-prognosis" class, whereas the remaining 79 belong to the "poor-prognosis" class.

For the purposes of this project, the gene expression data was randomly divided into a training data set (containing 120 patients) and testing data set (containing 175 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 from e-campus. 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 and design classifiers that best discriminate the two prognosis classes on the training data, and use the testing data to determine their accuracy.

We will consider the following classification rules:

- LDA, $p = 0.75$.
- Linear SVM, $C = 1$.
- Nonlinear SVM with Gaussian RBF kernel, $C = 1$.
- NN with 5 neurons in one hidden layer.

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

We will consider the following feature selection methods.

- Top 2 genes (exhaustive search).
- Top 3–5 genes (sequential forward search).
- All genes (no feature selection).

Therefore, each person will determine 5 gene sets for each of the 4 classification rules, for a total of 20 classifiers.

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 different classifiers based on the dimensionality 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?