Project: MS diagnosis based on SELDI-TOF spectra

Task

Tell if a person suffers from Multiple Sclerosis or not, from looking at two SELDI-TOF mass spectra.

(You can find an animation about the SELDI technology at Ciphergen's home page http://www.ciphergen.com - check the "Technology & Apps" tab.)

Data

You are given the file *SELDImsTrain.mat*. It contains three matrices: *trainInput* (80 × 10), *trainOutput* (80 × 1), and *testInput* (16 × 10). The two first matrices contain the input and output values for the training data and the latter is the input values for the hold-out test data set. The output value is one (1) for patients with multiple sclerosis (MS) and zero (0) for normal healthy individuals. The inputs correspond to the signal intensity in 10 mass intervals from two different SELDI arrays.

The task is to use the information in these mass intervals to classify a person as being either healthy or suffering from MS.

Steps and subgoals

1. Get acquainted with the data. Plot it and try to get a feel for the possible relationships between input and output. Try some different variable transformations (e.g. Box-Cox transformations) and see if this improves the discrimination power of the variables. The discrimination power can be measured using a Fisher-index measure.

2. Construct a quadratic Gaussian classifier for the problem, using all the variables, transformed if you want, and estimate the generalization error.

3. Prune the Gaussian classifier by successively removing the variable that results in the least degradation of the generalization error, until the degradation is significant. Note the classification error (generalization).

4. Construct a k-nearest neighbor (kNN) classifier, trying $k \in \{1, 3, 5\}$. Optimize the input variable set with respect to the generalization error (i.e. choose the set of input variables that maximize the generalization error).

5. Construct a multilayer perceptron (MLP) model for this. Optimize the number of hidden units (one hidden layer) with respect to the generalization error.

6. See if you can improve the MLP model by adding or removing any variables. Optimize the number of hidden units for the final model. Note the classification error (generalization).

7. Produce a test output file from your best Gaussian classifier. best kNN classifier, and your best MLP classifier. Estimate the expected test error and hand in the test files to me.

Report and presentation of results

You will present the results from your project in two ways: (1) A written report where the main conlusions are presented together with figures and tables supporting your conclusions. (2) An oral presentation, of about 20 minutes, to your course colleagues.

The report should be about 10 pages, including figures and tables, and should contain the elementary report constituents:

- Introduction (brief presentation of problem, 1 page)
- Methodology (brief listing of methods, 1 page)
- Data (presentation of your data set with important observations, 1-2 pages)
- Results (4-5 pages)
- Discussion

The report writing should not take more than one full day.

When you are finished with your report, and it has been accepted, then you should produce a postscript file with it, and pack it together with your dataset and other important parts of your project (like MATLAB M-files). The idea being that someone else could unpack it and repeat the main steps in your analysis without rewriting everything.