Electrochardiograms

Task:

The task is a classification task: To tell if a patient suffers from Transmural Ischemia (TI) or not, based on the signal from a 12 channel electrochardiogram (ECG). The 12 ECG channels are called V1, V2, V3, V4, V5, V6, aVL, I, -aVR, II, aVF, and III.

You can read more about ECG interpretation and Transmural Ischemia at e.g. http://endeavor.med.nyu.edu/student-org/erclub/skills.html (click on "EKG interpretation practice"). A PDF document with the information on this site is also downloadable at
http://endeavor.med.nyu.edu/student-org/erclub/ekgguide.pdf

Data

There are 300 observations: 150 control subjects (both healthy subjects and subjects suffering from heart infarction, but not TI), and 150 subjects that suffer from TI. For each subject you are given 26 features for each ECG channel, i.e. $26 \times 12 = 312$ features per subject. The features are 26 features for each one of the 12 channels in the ECG ($26 \times 12 = 312$).

The 26 features are denoted *I_dur, Q_dur, R_dur, S_dur, Rp_dur, Sp_dur, K_dur, I_ampl, Q_ampl, R_ampl, S_ampl, Rp_ampl, Sp_ampl, K_ampl, QRS_area, QRS_dur, Tmaxampl, Tminampl, timeTmax, timeTmin, ST_ampl0, ST_ampl20, ST_ampl40, ST_ampl60, ST_ampl80, ST_ampl100*. The "dur" features are the durations (in time) between different parts of the ECG (different points on the ECG are denoted P, Q, R, S, T...etc. as you will see in the www-page referenced above). The "ampl" are the amplitudes at these points. The "QRS_area" variable denotes the are of the QRS peak complex. The "timeTmax" and "timeTmin" are the times for the maximum and the minimum signal in the ECG. The "ST_amplXX" are the amplitudes at different points in the "ST" interval (which are generally thought to be important for doing the classification).

The most important features, believed by the physicians, are the times, i.e. features 19–26 for each channel. These correspond to inputs 19–26, 45–52, 71–78, and so on.

You are given a file, *ECGITtrain.mat*, which contains the matrices *inputECGITtrain* (200×312), *outputECGITtrain* (200×1, i.e. a column vector), and *inputECGITtest* (100×312). The first matrix is the inputs for 200 training patterns. The *outputECGITtrain* vector is the target values for the 200 patterns, where 1 correspond to a TI pattern and 0 to a non-TI pattern. The last matrix contains the inputs for the 100 test patterns. I keep the true outputs for the test data.

You can also have a file containing the raw ECG curves for each channel, but they are large (28 Mb for the training and half that for the test).

Steps and subgoals

1. Get acquainted with the data (also read some information about ECG interpretation). Plot the data and try to get a feel for the possible relationships between input and output.

2. Compute the Fisher index for each variable and rank your variables according to the Fisher index. Estimate how many of your variables that contain significant information about the problem (when considered one by one).

3. Construct a linear classifier (try both Gaussian linear classifier and a logistic regression classifier), trying different feature subsets in forward selection (follow the rankings by the Fisher index).

4. Construct a k-nearest neighbor (kNN) classifier for the problem. Select the variables using a forward selection method, following the variable rankings from the Fisher index tests above.

5. Construct a multilayer perceptron (MLP) model using the best inputs for the *k-NN* classifier. 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. Estimate the classification error (generalization).

7. Produce the classifications for the test inputs and mail to me (for your best linear, kNN, and MLP models), together with your estimates for how well you will do on the test samples.

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 (your results and comparison to other researchers' results, 1 page)

The report writing should not take much more than one full day, since you are two persons sharing the work.

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.