Project: Recognize mines and rocks using sonar

Task

The data set was used by Gorman and Sejnowski in a study of the classification of sonar signals using neural networks. The reference to the original work is

Gorman and Sejnowski (1988). "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets" in *Neural Networks*, Vol. 1, pp. 75-89.

The task is to train a network to discriminate a metal cylinder a roughly cylindrical rock, based on the sonar echo from these objects. During measurement, both cylinders and stones were lying on a sand surface, and the sonar chirp projected at them from different angles (aspect-angles) produced the variation in the data. The data was filtered and a spectral envelope of 60 samples was extracted (more information can be found in the reference above).

Data

There are 208 samples in total, the training set contains 104 of these and the test set contains the other 104 samples. You are given, in the file *sonarTrainData.mat*, the inputs and outputs for the training set as well as the inputs for the test set. I keep the outputs of the test set so that the final evaluation of your algorithm is blind. The samples are sorted in increasing order of aspect-angle. This is because Gorman & Sejnowski concluded, not surprisingly, that it was important to have a test set that was similar to the training set to be able to do the task well. (The separation into aspect angles is in a way cheating. The separation was based on a clustering scheme on the inputs so that the training and test sets have very similar inputs, which is not guaranteed in real life.)

The file *sonarTrainData.mat* contains three matrices and an information string: The matrices are *inputSonarTrain*, *inputSonarTest*, and *outputSonarTrain*, with obvious meanings. If you type `char(info)` then you will be able to read the information string. The input matrices are both 104×60. Each row contains the sonar echo from one measurement. The output matrix is 104×1 (i.e. a column vector) with the class labels for the measurements. A "0" means a mine and a "1" means a rock.

The spectra have 60 channels, and the data must (should) be reduced in dimensionality before models are built. There are many options for this, and you should do two of these. A simple "downsampling" version, and a principal components version. The downsampling consists in summing the power in six bands, thus lowering the number of inputs to ten (see below):

```
for i=1:10,
inputSonarTrainDS(:,i)=sum(inputSonarTrain(:,(i-1)*6+1:i*6),2);
end
```

This will produce a new input matrix, *inputSonarTrainDS* with size 104×10.

Steps and subgoals

1. Figure out how to do principal component analysis (PCA).
2. Get acquainted with the data. Plot it and try different transformations. Try to get a feel for the possible relationship between input and output.
3. Construct a linear classifier that uses the downsampled data as input and see how well you can separate mines from rocks. Try logistic regression, linear Gaussian classifier, and a perceptron with no hidden layer and a logistic output. For each of these three models, optimize the set of inputs with respect to the generalization error (estimated using cross validation).
4. Do the same thing as above, but using the principal components as input and see how well you can separate mines from rocks. It is wise to use both training and test input data when computing the principal components, to improve the statistics.
5. Build a k-nearest neighbor (kNN) classifier for the problem, using both the downsampled representation and the principal component representation. Try $k \in \{1, 3, 5\}$.
6. Construct a multilayer perceptron (MLP) classifier, trying both the downsampled representation and the principal component input. Optimize the number of inputs, and the number of hidden units with respect to the generalization error.
7. Hand in the test results for your best linear model, your best kNN model, and your best MLP model, together with your estimate of the generalization error.

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.