Project: Estimate cetane number for diesel fuel

Task

Your task is to see whether it is possible or not to estimate the cetane number for diesel fuel from the near infrared spectrum for the fuel.

Data

There are a total of

There are a total of 245 observations. You receive 133 of these for training and validation, and 112 are kept for out-of-sample testing.

You are given the file *cnDieselTrain.mat*, which contains three matrices: *cnTrainX* (401×133), *cnTrainY* (1×133), and *cnTestX* (401×112). The first matrix (*cnTrainX*) contains the IR-spectrum for each sample, one column per sample. The second matrix (*cnTrainY*) contains the output value *cetane number* for each diesel fuel. The third matrix (*cnTestX*) contains the input (IR-spectra) for each sample in the test data set.

The spectrum has 401 channels, and the data must first be reduced in dimensionality, e.g. by transforming to a principal component basis (which is how it is done in the commercial application).

Steps and subgoals

1. Work out how to do the 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 model that uses the principal components as input and see how well the cetane number can be estimated. That is, estimate the *generalization* error with a linear model. Optimize the number of principal components, i.e. determine how many and which components you need to get good generalization performance. The function *lsselect* is quite handy for this purpose. It makes sense to report both an MSE and an RMS error. Do a residual analysis (this comment goes for all items). It is wise to use both training and test set input data when computing the principal components, to get better statistics.
4. What you did under item 3 is called a principal component regression model (a linear regression model that uses the principal components as input). Another, and often better linear model, is the partial least squares (PLS) linear model. This model optimizes the projections automatically (i.e. there is no need to do the principal component selection). Construct a PLS model for the cetane number, using the NIR spectra as input (i.e.

no principal components). You are given a toolbox, the PLS toolbox, which you can use for this.

5. Construct a principal component multilayer perceptron (MLP) model. Optimize the number of principal components, and the number of hidden units with respect to the generalization error.

6. Hand in the test results (cetane number for the test data) for your best linear model, and your best MLP model, together with your estimate of the generalization error.

   These results must be handed in no later than 48 hours before your oral presentation.

7. Write a report.

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.