Datamining & Neural Networks: Excercise Session 3

In this session we investigate methods of dimensionality reduction and input selection using the MATLAB Neural Network Toolbox and Netlab 3.2 (downloadable from http://cl.ly/Rw0P). The code in this document was written for MATLAB version R2010b or later.

Consider the following problems

(1) **Dimensionality reduction by PCA analysis**

Consider the example of a biomedical application discussed in doc chodataset. The dataset consists of 264 data points with 21 inputs and 3 outputs.

(a) Investigate PCA analysis in order to achieve a dimensionality reduction

```
%% Load the choles data
% This will create a 21x264 choInputs matrix of 264 input patterns
% and a 3x264 matrix choTargets of output patterns
doc cho_dataset
load cho_dataset
%% Standardize the variables
 doc mapstd;
[pn, std_p] = mapstd(choInputs);
[tn, std_t] = mapstd(choTargets);
%% PCA
doc processpca;
[pp, pca_p] = processpca(pn, 'maxfrac', 0.001);
[m, n] = size(pp)
```

In this case the 21 inputs are reduced to 4 inputs.

(b) For the case of 21 inputs, define a training, validation and test set and apply the Levenberg-Marquardt algorithm:

```
%% Set indices for test, validation and training sets
 Test_ix = 2:4:n;
 Val_ix = 4:4:n;
 Train_ix = [1:4:n 3:4:n];
 %% Configure a network
 net = fitnet(5);
 net.divideFcn = 'divideind';
net.divideParam = struct('trainInd', Train_ix, ...
'valInd', Val_ix, ...
'testInd', Test_ix);
[net, tr] = train(net, pn, tn);
%% Get predictions on training and test
Yhat_train = net(pn(:, Train_ix));
Yhat_test = net(pn(:, Test_ix));
```

Investigate whether the performance can be improved by means of Bayesian regularization (*trainbr*). Compare the results on test data.

(c) Compare the training results of the case of 21 inputs (original inputs) and 4 inputs (after dimensionality reduction by PCA) by applying *trainbr*. Which choice would you make between the two options? Motivate your choice.

(2) **Input selection by Automatic Relevance Determination (ARD)**

In order to apply ARD, download the Netlab software from http://www.ncrg.aston.ac.uk/netlab/ or use the following link to get to the download's page directly: http://cl.ly/Rw0P

You can run the Netlab programs within Matlab. Consider the following experiments:

- Run the demo *demard*

- Run the demo *demev1*

- Consider the UCI dataset ionosphere data (file ionstart.mat) with 351 data points and 33 inputs. Solve this binary classification problem as a nonlinear regression problem with targets ±1 by writing a matlab program which is based on *demard.m*. Which inputs are most relevant? Try to reduce the network by taking the most relevant inputs and retrain. Discuss the obtained results.