Learning from Data, Assignment Sheet 2

School of Informatics, University of Edinburgh

Instructor: Amos Storkey

Submission Deadline: 23:59 Tues 5 December 2006.

Please remember that plagiarism is a university offence. Do not show your work to anyone else. The number of marks assigned to each task is given in square brackets. In total, this assignment will contribute 12% to your overall mark for LFD.

The questions ask you to do some MATLAB programming. You are asked to provide some written answers. You should try to make your code clear and comment it. The code will be looked at, and will be referred to if there are any questions regarding the originality of the answers given, or where an unforseen ambiguity arises. The code will not contribute to the marks itself. The marks will be dependent on the quality of the written answers.

This assignment is about Neural Network modelling.

The following marking scheme will be used:

A results/answer correct plus extra achievement at understanding or analysis of results. Clear explanations, evidence of creative or deeper thought will contribute to a higher grade.

B results/answer correct or nearly correct and well explained.

C results/answer in right direction but significant errors.

D some evidence that the student has gained some understanding, but not answered the questions properly.

E/F/G serious error or slack work.

Marks are given for each question.

1 Assignment

Please "hand in" your submission electronically, using the `submit` program. Put your answers in plain ascii text (no html etc) in a file named `answers.txt` in a directory

named `answers`, along with any code you wrote for the project. Then from a DICE directory that contains answers as a subdirectory type

```
submit msc lfd-5 2 answers
```

if you are an MSc student or

```
submit ai4 lfd-4 2 answers
```

if you are a third or fourth year ai/cs student. Failure to follow these instructions could result in no marks being given. Note the 2 to distinguish this submission from that for the first assignment. Typing a 1 instead will result in your submission not being received and zero marks being given. Be warned. Be careful.

I repeat: Typing a 1 instead will result in your submission not being received and zero marks being given. Be warned. Be careful.

The data for this assignment is in `lfdassignmenttwo2006.mat`. This data should be loaded in to MATLAB using the `load` command, and is to be downloaded by going to the course assignments page: `http://anc.ed.ac.uk~amos/lfd/lfdassignments.html`. Do not attempt to run this file. Download it to your working directory. The `.mat` file contains training dataset `xmu`, training targets `ymu`, and validation/test datasets `xnu`, `xnu2`, and corresponding targets `ynu`, `ynu2`. In this assignment we are interested in supervised learning. This dataset is derived from the `mpg` UCI dataset relating to the miles per gallon of different cars. The targets are the miles per gallon a car does, and the remaining data are various features of the car. The three datasets correspond to cars from different regions.

1 [25 Marks] Perform various sensible visualisations of the data, commenting on what you see. You should comment on each of the different variables, any differences between the different datasets, and any other salient observations you wish to back. If possible back up your statements numerically.

2. [20 Marks] **Rescaling** You will have noticed that the data above has not been scaled. How would you rescale the training data for use in a neural network? Write at most four lines of code to transform some training data into the rescaled version. Give the code as part of your written answer. Write a description of the reasons for your choice of rescaling.

Remember that the validation and test data must also be transformed in the same way. Why can this not be done with another run of your pseudocode using the validation or test sets instead? Provide a description indicating how you rescaled the validation and test data, and the reasons for your choice.

3. [35 Marks] Train a suitable neural network with one hidden layer on the rescaled training data (`xmu,ymu`). Using (`xnu,ynu`) as a validation set assess a good choice of number of hidden layer neurons, random seeds etc. Choose about 8-10 different runs. Carefully describe the form of your choice of network form, and the exact process you went through in obtaining a good final network (give a table of the settings for each run

and the training and validation errors). Justify the decisions you made in this process. What is the performance of the final network of the test data (`xnu2,xnu2`)? Comment on the relationship between the performance of the final network on the training, validation and test data. You should use the `NETLAB` toolbox for neural networks. Type `help netlab` in matlab. See `http://anc.ed.ac.uk~amos/lfd/lfdassignments.html` for information on how to set the matlab path to include the netlab toolbox.

4. [20 Marks] In light of your visualisations, comment on your use of (`xnu,ynu`) as a validation set given the training data you have, and given that you will be using the network for prediction scenarios similar to that in the test data.

Notes on Matlab

- Remember, there are only 100 licences. If you have finished using MATLAB, quit from your session so that others can work.
- On-line documentation about NETLAB can be found at `http://www.dai.ed.ac.uk/dai/computing/software_manuals/netlabhelp/index.htm`, at the netlab website: `http://www.ncrg.aston.ac.uk/netlab/index.php` and in the text book Netlab: Algorithms for Pattern Recognition which is available in the library. You should run the netlab demos and refer to the netlab demo code (e.g. `type demmlp1`, `type demmlp2`) to help you. Allow yourself some time to familiarise yourself with netlab. `help netlab` gives a list of all the netlab functions.
- You can find out more about NETLAB functions (and many MATLAB functions) by typing `help` followed by the function name. Also, you can find the `.m` file corresponding to a given function using the `which` command, for example

  ```
  >> which mlp
  /usr/local/lib/matlab/toolbox/local/netlab/mlp.m
  ```

  You can then examine the code (but this should not be necessary for the assignment). Note you should ensure the netlab toolbox is on the path using `addpath`.
- The command `clf` is very useful; it clears the current figure. This is useful as the functions that you have have `hold on` set, which means that successive plots are plotted on top of each other. If this has become confusing, `clf` clears the figure so you can start again. `clear` clears your workspace of variables (check that the command `who` returns nothing).
- If you wish to find out how much cputime a function is taking, use the online help `help cputime`.
- The functions `mlp`, `netopt` and `mlpfwd`. These are the main NETLAB functions used in this assignment. `mlp` sets up a MLP network. In the call

  ```
  net = mlp(nin, nhidden, nout, transfunc);
  ```

where transfunc determines the type of transfer function the output unit has.

`netopt` is a wrapper function that calls the various optimization routines, `scg.m` in this assignment.

`mlpfwd` forward propagates inputs through the network, to produce the corresponding outputs.