Assignment Details

This assignment consists of two parts: Part 1, which is a regression problem, and Part 2, which is a classification problem. In order to achieve an overall pass for the assignment, you must complete both parts. The task for each part is described individually below.

Part 1: Predicting the median value of house prices

The Dataset

In 1978, Harrison & Rubinfeld published data on various factors which might affect housing values in the Boston metropolitan area in 1970. The data originate from several sources, including the 1970 US Census and the Boston Metropolitan Area Planning Committee.

The dataset consists of 506 samples of data containing 14 features of housing areas and the housing value in that area. The objective is to predict the median value of homes (the 14th feature) using the other 13 features. The 14 features are:

CRIM	1.	per capita crime rate by town,
ZN	2.	proportion of residential land zoned for lots over 25,000 sq.ft.,
INDUS	3.	proportion of non-retail business acres per town,
CHAS	4.	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	5.	nitric oxides concentration (parts per 10 million)
RM	6.	average number of rooms per dwelling
AGE	7.	proportion of owner-occupied units built prior to 1940
DIS	8.	weighted distances to five Boston employment centres
RAD	9.	index of accessibility to radial highways
TAX	10.	full-value property-tax rate per $10,000
PTRATIO	11.	pupil-teacher ratio by town
B	12.	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	13.	% lower status of the population
MEDV	14.	the median value of owner-occupied homes in $1000's.

(learn how load an influence on house price http://www.quantlet.com/mdstat/scripts/mva/htmlbook/mvahtmlframe

For the purposes of this assignment, the original data set has been divided into three sets: a training set, a validation set, and a test set. These are available as the files `Boston_Training.dat`, `Boston_Validation.dat` and `Boston_Test.dat`. Note that features 1 to 13 have been normalized by subtracting the mean, and dividing by the standard deviation. Feature 14 has been scaled by a factor of 0.02, so that the maximum value is 1. The data files are available from the subject web page.

The Task

Your task is to investigate how the many factors involved in MLP training affect the performance of the network on the Boston housing dataset. The factors you should consider include:

- network size (restrict networks to one hidden layer, but vary the number of units in the hidden layer);
- learning algorithm (back-propagation, scaled conjugate gradients)
- training parameters (learning rate and momentum);
- early stopping;
- weight regularization;

The performance indicators you consider should include:

- generalization performance (*i.e.*, ability of the network to predict values of novel examples not used for training);
- learning speed.

The investigation is open-ended, and you are invited to also perform any other experiments that you think may provide further insight into neural networks training on regression tasks.

Notes and advice:

- For each combination of parameters you use, plot the error on training examples, validation examples, and test examples. Make sure that you record what the actual parameter values were, so that you can later analyse and interpret the results.
- If you have plotted the training, validation, and test error on the same axes, then it is not actually necessary to implement early stopping, as you can obtain the test set error from the plot. $e = t - y$
- It is suggested that you start with the back-propagation training algorithm. Once you are familiar with how the various factors affect its performance, you should have some idea about what is the best performance that you can achieve with this algorithm on this data set. You can then use this as a benchmark with which to compare the performance of the scaled conjugate gradients algorithm.
- You will obviously have to run experiments using various combinations of parameters (network size, learning rates, weight regularization coefficients value, etc.). DO NOT TRY TO USE EVERY POSSIBLE COMBINATION – just do enough so that you can see the effect of the various parameters on network performance, and draw meaningful conclusions from your results. You will need to be selective in what you include in your report, anyway.

Part 2: Predicting the presence of heart disease

The Dataset

The Cleveland Heart Disease dataset consists of 303 examples containing 14 features of medical data thought top be related to the presence of absence of heart disease. The objective is to predict the presence or absence of heart disease (Feature 14) using the other 13 features. The 14 features are:

1. age: age in years
2. sex: (1 = male; 0 = female)
3. cp: chest pain type
   - Value 1: typical angina
   - Value 2: atypical angina
   - Value 3: non-anginal pain
   - Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
   - Value 0: normal
   - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)