Columbia GSB: Machine Learning
Homework 4

Dr. George A. Lentzas

This Homework is due by 11:59pm on Sundary April 30th. Submit your results in .pdf format in Canvas. Good Luck!

1. Go to the ISLR website `http://www-bcf.usc.edu/~gareth/ISL/data.html` and import the 'advertising.csv' data. For a review of the data you can refer to Chapter 2 of ISLR. Use set.seed(1) and sample 150 observations for your training data and 50 observations for your test data.

   (a) Fit a one hidden layer Neural Network to predict sales and calculate your test MSE.

   (10 points)

   (b) Explain how you chose the number of hidden units and any optimization parameters (e.g. learning rate) in your Neural Network and its calibration.

   (10 points)

2. Using the same dataset build an ensemble of 30 Neural Networks; these can have up to a maximum of three hidden layers, different starting values, different number of hidden units per layer and different activation functions. Calculate the ensemble test MSE and explain your findings.

   (20 points)

3. Here you will use the Dow Jones 30 constituents log returns from $1987-03-16$ to $2009-02-03$ from Yahoo Finance found in the R package 'rugarch' to fit an Elman recurrent Neural Network. Once you import the data calculate the average return for all the constituents. This will be your target (Y) variable; the features (Xs) will be the one-lag returns of all the Dow Jones 30 constituents. In other words, you will use last week's returns for all 30 constituents (the Xs) to predict this week's average return (the Y).

   (a) Split your data into a training set from $1987-03-16$ to $2005-12-31$ and a test set from $2006-01-01$ to $2009-02-03$. Fit a one layer, five hidden unit Elman network using the package RSNNS. Plot the Iterative Error of your fit.

   (10 points)

(b) Use your model above to make predictions on the test set and report your test MSE. Hint: look at the help for predict.rsnns function.

(10 points)

(c) Discuss your findings; especially what you think about over-fitting, model selection and why this is or is not a good model for this data.

(10 points)

4. Last you will use the weekly percentage returns for the S&P 500 stock index to fit a Deep Belief Network (this is similar to the R Lab Stacked Autoencoder example we did in class). The features (X variables) will be the 5 lagged returns (numeric values) and the target variable (the Y) will be today's numeric return. The training data will be the first 989 observations and the test data the remaining 200 observations. You can use either of the 'deepnet' or the 'RcppDL' packages.

(a) Fit a RBM-DBN consisting of at least 2 hidden layers.

(10 points)

(b) Use the fitted model to make predictions on the test data and calculate the test MSE.

(10 points)

(c) Discuss your findings, including your choice of number of hidden layers, number of hidden units and the model's performance.

(10 points)