Big Data in Finance 2: Group presentations on neural networks

Instructions for submission and presentation

Please submit both your code, output such as graphs and tables, and your slide deck for the group presentations.

We strongly recommend that you submit all the code as one Jupyter notebook. Use markdown cells to include commentary. At the begining of the answer to each separate question, place a markdown cell that states the question.

Your slide deck should be concise so that you can present it in 15 minutes. The number of slides is up to you, but you should practice to make sure that you can finish on time. We cannot give you credit for any content on slides that you do not manage to talk about in the 15 minutes.

You may use packages such as sklearn to estimate models.

You are welcome to use R for coding, although Python is recommended. Both work with jupyter.

Setup and introduction

Download the stock data saved on the hub as */Data/stocks_presentations/Stocks_data.csv*.

For interpretations of the variables included, download the data dictionary in */Data/stocks_presentations/dictionary.xlsx*.

The data are a stock-time level panel running from 2010 to 2017. As well as stock prices, and returns, the data contains corporate finance variables that describe the characteristics of each firm, and macroeconomic variables that vary over time, and are defined as in Goyal and Welch (2008). For definitions about the macroeconomic variables see the paper by Goyal and Welch (page 3 to 5). For the rest, see the dictionary included.

Allocate the data into an 40% training set, a 40% validation set and a 20% test set. Respect the panel structure of the dataset: The models should be trained data older than the validation data, which should in turn be older than the test data.

Your task it to use state of the art supervised learning methods to predict stock-level returns. We will not give you detailed directions, because the process of tuning a deep learning model is an iterative one: You need to learn from trial and error.

If your end results are not outstanding, don't worry, as long as you can give solid statistical and economic intuition for why this has happened, and how it could be improved with further work.

Question 1: Understand the data

Start with a simple LASSO regression of one-month ahead stock returns on the firm-level and Macroeconomic variables. Tune the hyperparameter using the validation set.

Which variables are important? Are those the variables you expected?

Are there other asset classes and sample periods for which you would/would not expect those relationships to hold?

Question 2: Train a Neural Network

Train a neural network to predict one-month ahead returns. Using the validation set, tune the model by modifiying the architecture, and regularisation procedures. Regularisation will be important here because you do not have a huge dataset.

Gu, Kelly, and Xiu (2018) follow the geometric pyramid rule from Masters (1993) for the hidden layers and number of neurons at each layer, but you can try alternatives if you wish. This is an exploratory process, with local minima: You will want to run many specifications.

Which model performs best on the out of sample test set?

How does it compare to the second best model?

How does it compare with the LASSO?

Which variables are important, and how do they differ from the lasso model? Give an economic interpretation of your findings.

Question 3: Include past returns as predictors

Now add up to three months of past returns as explanatory variables. Again, using the validation set, tune the model and the hyperparameters.

Are the predicitons better than without past returns? What is the economic meaning of this result?