Matthew Lynn

Zhaolun Shi

Ujwal Madhavapeddi

Recurrent Neural Network for Stock Market Movement Prediction

1. Techniques and Algorithms we plan to implement
   1. Recurrent Neural Network
      1. Forget Gate
      2. Input Gate
      3. Output Gate
      4. LSTM
      5. Backpropagation through time
2. Dataset details, such as number of features, instances, data distribution
   1. Data comes from Dow Jones Industrial Average (DJI or DJIA) from Yahoo finance
      1. Data is updated every day (even now, yahoo hasn’t died yet)
   2. Dimensions of the data from 1985 to present: 5865 2890
   3. Our features will be 'Date','Open','High','Low','Close',’Volume’ with some added features from preprocessing such as ‘Direction’ and ‘Midpoint’.
   4. Stock data is said to be log-normally distributed
3. Coding language / technique to be used
   1. Python will be the main driver
   2. R may be utilized for toy examples and testing
   3. Tensorflow Backend for complex computations and back testing
   4. Keras for back testing
   5. Scikit Learn for preprocessing
   6. Pandas for data import and data structuring
   7. Numpy for array management and data structuring
4. Preliminary Results (if available)
   1. Scratch.py is attached (skeleton of the code comes from [here](# https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/))
      1. 100 Epochs were ran with batch size = 1
      2. Mean Square Error with optimizer [Adam](https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/)
   2. Attached is scratchOUTPUT.txt which has the output for our first run at the data. This preliminary result utilizes no self made functions (utililzes Keras and Tensorflow for the RNN). The Final report will showcase no RNN functions from these libraries.
      1. Train Score: 118.29 RMSE
      2. Test Score: 1231.41 RMSE
   3. Also attached (Below) is a plot showing the distribution of the data over time
      1. The Blue portion of the graph is the actual data
      2. The Orange portion is the training prediction (practically perfect)
      3. The Green portion is the testing prediction (almost similar but underestimates overall)
   4. An additional plot (second plot down) shows midprice by Date
      1. notice the crash ~2007-2008

