TEAM - Wall Street Wizards

AUQUAN Project Documentation

DATA Preprocessing and Prediction Algorithm

We were provided with 111 stocks containing prices of stocks after each minute. The problem is to predict the price accurately so as to increase the profits by longing and shorting the stocks appropriately.

We first downsample the data for each stock by a window of 4 leaving only the data at every 4th minute. Then we use a convolutional neural network, a state of the art Machine Learning model, used for audio, speech and even in stock prediction, all of which are sequential data forms.

The model is being fed with 5 kinds of data in a window of 1000 samples:

1. Raw data
2. Rolling averaged data in a window of 100
3. Rolling averaged data in a window of 50
4. Rolling averaged data in a window of 25
5. Momentum across 25 samples.

Now we predict the price change after 100 samples. All the data is normalized to zero mean and variance 1 before being input into the algorithm.

Optional preprocessing:

We normalize the price change for each stock also for creating the labels. Note: It would change your prediction to normalized prediction, therefore we have to denormalize it before actually using for trading.

We train a single model on 111 stocks and backtest on the 20 % future data to get an accuracy of 70 % within 60 epochs. We are using a quadro P5000 GPU for training with 16GB GPU RAM and 32 CPU RAM.

The code will save the model and weights is the directory model. User can use the saved model for prediction by skipping the training phase and loading the model with same preprocessing.