Deep Learning – Project Proposal

Title: "High-Frequency Trade Sign Classification"

Group members:

- Kegan Penovich
- Matt Isaac
- Jared Hansen

1. **The goal. What problem are you trying to solve?**

   Our goal is to create (a) neural network(s) that classifies high-frequency trade (HFT) initiation better than current algorithms in the academic literature. In HFT, it is of interest who is the initiator (later-arriving) in a trade: the buyer or the seller? Thus, this is a binary classification problem. Specifically, we want to surpass accuracies in Rosenthal's 2012 "Modeling Trade Direction".

2. **The data. What data will you use to solve this problem?**

   The main data that we'll be working with is the ITCH data feed from Nasdaq. This data feed provides all messages on the Nasdaq exchange for a specific security on a specific day. For example, Nasdaq keeps track of all bids, asks, cancels, etc. every trading day for the AAOI ticker (Applied Optoelectronics).

   We'll have to parse down and clean these data before training or testing any networks. (Jared already has code that does this and has cleaned some data.) Our plan is to scale up, cleaning many files so we have more training data. We have data for 21 trading days, for 12 tickers. A single trading day for a single ticker has around 5,000 trades. This will give us a total of about 1.2 million observations if we clean all the data (and assuming that all tickers have roughly 5,000 trades executed per day).

3. **The methods. What neural network architectures do you plan to use to solve the problem? Also, what other machine learning algorithms will you compare to for a baseline?**

   The main architectures we plan to employ are variants of recurrent neural networks. RNN's work well for sequence data, and are typically used for speech recognition, text prediction, and time-series-type analysis. Although the response for our data is binary, they are still certainly sequence data, here coming in relative to each other in terms of time of trade execution. They are like other forms of sequence data in that trades that are close to each other in the sequence are likely to have a similar sign (buy or sell). This is due to dynamics of these markets, and the fact that many smaller trades comprise upward or downward movement in an asset price.

   Other ML methods we'll compare to include random forests, SVM, and logistic regression.

4. **Planned contributions of each member.**

   - Kegan:
     - proofread and edit the project proposal
     - go-to math expert (i.e. cost function modification or design, architecture choices relying on deeper mathematical understanding, helping Matt and I understand what's going on w/math)
   - Matt:
     - proofread and edit the project proposal
     - go-to other-ML-methods expert (i.e. for comparing RNN's, Matt will know and advise on forests, SVM, logistic regression, etc., and will take the lead in training these models)
   - Jared:
     - draft the project proposal
     - data: clean data, go-to data-generating process expert (i.e. "what does this feature mean?")
   - Shared duties:
     - build and debug networks
     - hopefully figure out CHPC for scaling up on larger numbers of training data points