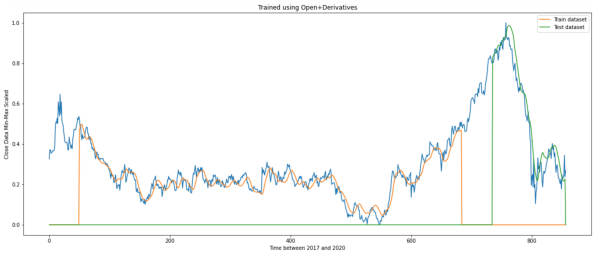
Predicting Stock Behavior on K-Means Clustered Data using an LSTM

By David Chaparro and Spencer Shortt

University of Colorado at Boulder Machine Learning CSCI 5622

Problem Space

Stock market data holds a large amount of data over time. Many parties, from hedge funds to individuals are interested in which stocks to buy. Predicting the behavior of stocks holds great monetary importance for use in investing and individual use. Our project uses an LSTM to predict future stock behavior for the next day given data of the 10 days prior. From these 10 days, our goal is to determine which features are best to predict the stock over its training period, including our own custom features which we derived from stock data.



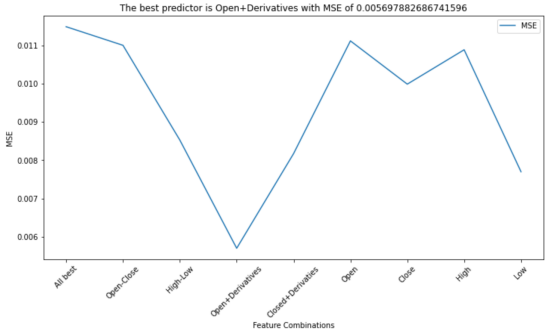
Data

Each stock has 250 Entries of 7 features each. Since we are only examining the NYSE, we will be examining 1000 stocks over the time period from 2017 to 2020. This means each stock will have 13 thousand data points and the whole 15 million data points In addition, we created 12 additional features this amounts to help characterize the stock trends. We will be performing feature optimization on only one stock initially but will see if the same set of features holds for multiple stocks. Data is either Min-Max Scaled or Absolute Value Max Scaled.

Original Features	Created Features
Date	Day of Year
Open	Weekday
Low	Year
Volume	1st Derivative
High	2nd Derivative
Close	Nightly Difference
Adjusted Close	~~Adjusted Close~~

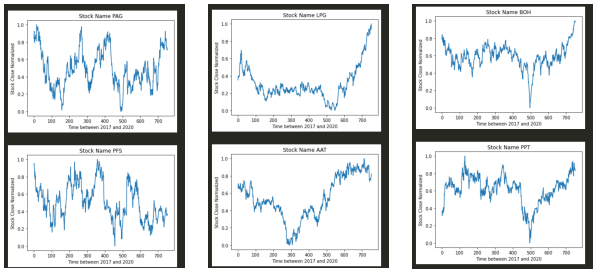
LSTM – Feature Selection

To judge the effectiveness of our LSTM, we decided to use a MSE of the predicted next day value of a stock, and the actual next day value of the stock. The minimum MSE of the predicted stock price and would give us the best combination of features for predicting a stocks value. The LSTM is trained on the closing price given an input of features from the stocks data. We especially want to test if other features, or combinations of features have better predicting power than only using the closing price as your training set.



K-Means Stock Selection

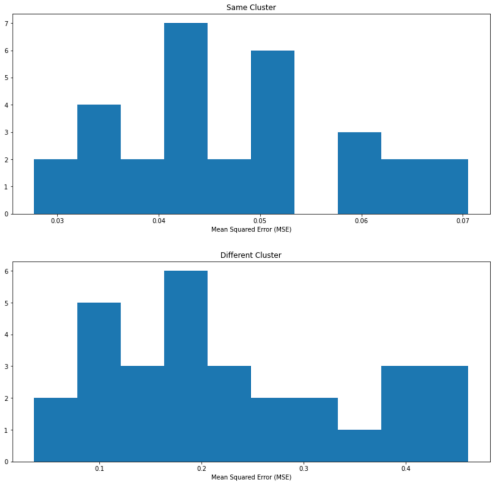
K-Means was used for grouping similar behaving stocks. We split our dataset into two clusters initially, then examined the clusters in one of the two initial clusters to group similarly behaving data. We used Dynamic Time Warping "dtw" distance metric to characterize the behavior of a stock closing price. This was primarily used to choose stocks to perform feature analysis and see if the methods are applicable to other similar stock.



Group 0 Group 4 Group 5

LSTM Effectiveness

LSTM Performance on the same cluster shows good predictive ability, meanwhile outside the cluster, the model performs an order of magnitude poorer.



Conclusion

- Created a method for training LSTMs on specific stock data and behavior
  - Works well, works best with raw features
  - Derivatives help but only somewhat
- K-Means has incredible depth, more time could be spent categorizing stocks
  - Classify stock types
  - Classify larger time periods
- Unsupervised Learning
  - Train for best buy and sell times
  - Build your own loss function

References

Dataset
https://www.kaggle.com/datasets/paultimothymooney/stock-market-data
Time Series K-Means
https://tslearn.readthedocs.io/en/stable/gen_modules/clustering/tslearn.clustering.TimeSeriesKMeans.html#tslearn.clustering.TimeSeriesKMeans.fit
LSTM
https://medium.com/@kumarpal.nagar/stock-price-prediction-using-artificial-recurrent-neural-network-part-1-5955f93b6734
https://medium.com/@kumarpal.nagar/stock-price-prediction-using-artificial-recurrent-neural-network-part-2-589ad903957e
DTW and K-Means:
https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3