CS5831 Final Project: Status Update

The Mining Avengers
Blake Krouth, Ganesh Vannam, Nandhika Rajmanikandan

March 2025

We hope to explore market regime and trend detection since financial data is abundant, easily accessible, and both statistical and machine learning approaches have proven successful in financial applications. The biggest challenge will be dealing with the noisy nature of financial markets, which requires careful feature selection and hyperparameter tuning. We aim to automate the detection of market conditions (such as bull, bear, or sideways trends). Using stock price data, technical indicators (e.g., moving averages, RSI, MACD), and macroeconomic data, we plan to experiment with applying different clustering methods (e.g. K-Means, DBSCAN, HMM, GMM) to time-series financial data. Time permitting, it would be interesting to explore transformers and other recent trends in machine learning. Market regime detection is important for making informed financial decisions, whether that be for trading or investments.

We are retrieving data from Yahoo Finance via the Python library yfinance. Using yfinance, we have acquired closing prices for the NASDAQ-composite for the last 10 years. Of course, we are not limited to the NASDAQ-composite, or data in the last 10 years, and this can be adjusted as we see fit. We have so far calculated a 7-day moving average for the data, and then the return and log-return for this 7-day moving average. The motivation for the moving average is to reduce the noise that is evident in time-series financial data [1], [2]. The motivation for using log-returns is that they are often more symmetric and convenient to work with [1]. [2] used PCA to select 8 macroeconomic indicators as features for detecting market regime. While this required significant pre-processing, it seems plausible that it will result in better model performance than just log-return. Depending on our initial model performance with log-return, we may shift to using macroeconomic indicators, which can be obtained from the FRED API. It should be noted that other literature has reported success with only log-return, such as [3].

For a starting point, we are considering replicating and evaluating the best of the three models (agglomeration clustering, Gaussian mixture model, hidden Markov model) evaluated in [1], the hidden Markov model (HMM), and its variations as discussed in [2], along with models we will cover in class (e.g., K-Means). Since cross-validation is difficult with unsupervised learning, we hope to follow the procedure outlined in [2] in which model families are compared using marginal likelihood and hyperparameters are selected using Silhouette score. To evaluate our model performance, the most logical option appears to be to plot our model predictions and visually analyze the model output (the detected regimes) as done in both [1] and [2] by comparing the classifications to (known) historical market trends. [3] performs both a qualitative visual validation with historical data and a quantitative validation using synthetic data that has been labeled beforehand. They then compare the results of their clusters with the synthetic labels. While having quantitative validation results would be ideal, it seems unlikely that we will have time to generate quality synthetic data. We have not yet begun model evaluation.

References

[1] H. Aramyan, J. Ramchandani, and M. Skevofylakas. "Market regime detection using statistical and ml based approaches," LSEG Developer Community. (Feb. 2023), [Online]. Available: `https://developers.lseg.com/en/article-catalog/article/market-regime-detection`. (accessed Mar. 1, 2025).

[2] A. Hirsa, S. Xu, and S. Malhotra, "Robust rolling regime detection (r2-rd): A data-driven perspective of financial markets," Available at SSRN, Feb. 2024.

[3] B. Horvath, Z. Issa, and A. Muguruza, "Clustering market regimes using the wasserstein distance," Available at SSRN, Oct. 2021.