A STOCHASTIC TIME SERIES MODEL FOR PREDICTING FINANCIAL TRENDS WITH NLP

Pratyush Muthukumar
Department of Mathematics
Department of Engineering, Computer Science, and Technology
California State University Los Angeles
Los Angeles, CA 90032
pmuthuk2@calstatela.edu

Jie Zhong
Faculty Mentor
Department of Mathematics
California State University Los Angeles
Los Angeles, CA 90032
jie.zhong@calstatela.edu

February 13, 2020

ABSTRACT

We consider a stochastic time series model for predicting the risk of a corporation's assets through linguistic analysis of earnings conference calls (ECC). ECCs are vital performance indicators of a company's assets over the fiscal year. By translating specific phrases into word vectors and discerning similarities between related concepts and phrases, we can improve the current strategies of time series forecasting. We also introduce a novel method for time series forecasting by computing the stochastic volatilities of stocks based on the subtleties of ECC sessions. Our neural network model has similarities to the structure of a convolutional character decoder, calculates along a time series of data, and can accurately predict the stochastic risk-reward payoff of a company by factoring in human sentiments gathered through conference call analysis. We hope the model can open up new discussion on predicting the human sentiment portion of stochastic noise that is so widely unclear within the financial prediction field. We gratefully acknowledge support from CSU-LSAMP, supported by the National Science Foundation under Grant HRD-1302873 and the CSU Office of the Chancellor.

1 Introduction

This program will predict aerospace stock trends based on years of financial data and financial news texts. We used machine learning on this combination of both words and numbers because ML is more efficient and less biased than humans. By using both numbers and text to predict stocks, we are getting the full picture of the stock's performance and its public impressions.

2 Methodology

- Build a Generative Adversarial Network (**GAN**) for **time-series** data (set of datapoints indexed by time).
- Add **stochasticity** by using the results of the Natural Language Processing (**NLP**) analysis of financial text as an input into the GAN.

A GAN is a neural network (NN) that has a special way of improving its learning. Neural networks are computer models that contain a series of nodes connected to a series of synapses, similar to a human brain. GAN puts two NNs against each other, a generator and a discriminator. The generator tries to make a stock prediction as similar as possible to the true time-series input stocks. The discriminator tries to find the difference between the real and generated stocks. We used NLP Naive Bayes Sentiment Analysis to categorize the overall message of every sentence in the financial texts. Every sentence was classified with Naive Bayes into a positive, neutral, or negative (1,0,-1) effect. These values were fed into the latent space input of the GAN.

3 Mathematical Intuition

Naive Bayes Sentiment Analysis

Let C be the set of all classes (positive, neutral, negative). Let D = $\{x_1, x_2, \ldots, x_n\}$ be the dataset. By Bayes' Theorem, the most likely sentiment is

$$\arg\max_{c \in C} \frac{P(D|C)P(C)}{P(D)} = \arg\max_{c \in C} P(x_1, x_2, ..., x_n|C)P(C).$$

By assuming independence by the nature of language, we get

$$\arg\max_{c \in C} P(x_1, x_2, ..., x_n|C)P(C) = \arg\max_{c \in C} P(C) \prod_{x \in X} P(x|C).$$

Generative Adversarial Network Let generative network G have latent space input z (density p_z) and output $x_g = G(z)$. Let discriminative network D have input either x_t (density p_t) or x_g (density p_g) for true/generated data and output $D(x) = P(x = x_t)$. The error function is

$$E(G, D) = \frac{1}{2}(\mathbb{E}_{x \sim p_t}[1 - D(x)] + \mathbb{E}_{x \sim p_g}[D(x)]),$$

with the goal of the algorithm being $\max_G(\min_D E(G, D))$.

4 Conclusion

Model (1-month Forecast Horizon)	RMSE
Time-Series GAN + NLP Input	4.32
LSTM Only	13.24
ARIMA(5,1,0)	32.43
NLP Only	174.87

There were **3,105,536** data points corresponding to 3466 days worth of 112 daily features for 8 aerospace portfolios. The GAN ran a 500-epoch set with 6211 datapoints each in **1.2** minutes total after training, analyzing **43,132.44** stock features per second. We would like to understand theoretically how much of an impact the sentiment analysis stochasticity had on the final accuracy.

5 Future Insights

By creating an accurate and efficient prediction model, our program could be a viable option for stock brokers or financial experts to gain insight into the economy. Through our novel use of public sentiment analysis, we hope to usher in new methods of time-series prediction.

6 References

[1] Dormehl, Luke. *What Is an Artificial Neural Network? Here's Everything You Need to Know.* Digital Trends, Digital Trends, 6 Jan. 2019.
[2] Nicholson, Chris. *A Beginner's Guide to Generative Adversarial Networks (GANs).* A.I. Wiki, Pathmind, 18 June 2019.