Homework 3

Math 189Z – Covid-19: Data Analytics and Machine Learning

Nico Espinosa Dice
April 23, 2020

Paper 1: Hidden Markov Model for Stock Trading

Author: Nguyet Nguyen

In this paper, Nguyen presents a novel method of using a hidden Markov model for stock trading. The goal of all stock traders is to buy stocks at low prices and sell them at high prices, thus making profits. For decades, traders have attempted to create models to allow them to better predict when to buy and sell stocks, in addition to knowing what stocks to select (see Nguyen and Nguyen*,* 2015).

Stock prices are a form of time-series data, and a difficulty in creating models is incorporating time-series data into them. For example, when attempting to predict a stock's price at the end of the day, it is important to not only know the current stock price but the past price during the last few days. (This, of course, leads to the question of how to weigh the recent information in the model).

Nguyen developed and tested the performance of HMMs with between two and six states. It was found that the HMM with four states performs best and yields higher returns than the "Buy and Hold" strategy – a strategy where the stock is bought immediately and held until the end of the time frame. Consequently, Nguyen shows that HMMs have significant potential in applications of stock trading.

Paper 2: Gene Finding and Hidden Markov Models

Within the field of computational biology, there have been significant advancements in gene finding – a process that attempts to locate regions of DNA that encode genes. Genes are sequences of nucleotides that encode – meaning they provide the instructions for – the process of synthesizing proteins and RNA (National Institutes of Health: National Library of Medicine).

The paper focuses on two types of organisms: prokaryotes and eukaryotes. The cells of prokaryotes do not have nuclei, whereas the cells of eukaryotes do. Gene finding in eukaryotes is more difficult than in prokaryotes because the structure of eukaryotic genes have both coding and non-coding regions, called exons and introns respectively. In order to assist in gene finding of eukaryotes, the researchers attempted to use hidden Markov models, which were used to differentiate between exons and introns. The HMM employed the Viterbi algorithm, which allowed them to use maximal likelihood estimates to determine the particular hidden states that generated the event, in this case the DNA sequence. However, after further review, the paper states that because exons, coding sequences, were "interrupted" by introns, non-coding regions, HMMs and the Viterbi algorithm were not successful in the application of gene finding in eukaryotes.

Project Source: Creating The Twitter Sentiment Analysis Program in Python with Naive Bayes Classification

Author: Anas Al-Masri

In this post, Al-Masri describes how to create a sentiment analysis program to analyze tweets using Naive Bayes classification methods. The program is written in Python, and Al-Masri includes links to the program's code.

A Naive Bayes classification algorithm is a type of probabilistic classifier that uses Bayes' theorem for predictions. By applying Bayes' theorem, the algorithm is able to make a prediction based on previous "odds" or probability estimates of past events.

The goal of sentiment analysis is to determine the attitude and emotion present in a given form of communication. Sentiment analysis applies natural language processing techniques in order to do so. In this post, Al-Masri applies sentiment analysis to Tweets.

I will use this paper because I hope to conduct sentiment analysis on Tweets regarding the Covid-19 pandemic. In doing so, I hope to gain an understanding, and possibly a predictive method, of determining how the severity of the virus spreads geographically, using Tweets from those areas.