Eve Kazarian
Math 189Z: Covid-19 Data Analytics and Machine Learning
April 2020

Homework 3

Summaries of provided readings

Article 1: "Hidden Markov Model for Stock Trading," Nguyet Nguyen
(https://www.mdpi.com/2227-7072/6/2/36/pdf)

Researchers developed a hidden Markov model (HMM) to forecast closing prices of stocks. Four criteria (AIC, BIC, HQC, and CAIC) were used to select the optimal number of hidden states in the HMM. The close, open, high, and low prices of stocks were used as observations in the HMM. The researchers then tested their model against the historical average return (HAR) model.

In their findings, the researchers noted that four hidden states were optimal for their HMM. Their model predicted S&P 500 monthly prices accurately. It also consistently generated a greater profit than the HAR model when simulating S&P 500 trading.

Article 2: "Gene finding and the Hidden Markov models"
(https://www.cs.us.es/~fran/students/julian/gene_finding/gene_finding.html)

The aim of this project was to identify genes and segmentation in genomes. Preliminarily, researchers looked at prokaryotes, which are predominantly protein-coding, to see where open reading frames occurred. They classified longer open reading frames as more likely to contain genes.

Then they moved to eukaryotes, which contain introns and exons that are not protein coding and complicate genome sequencing. They applied hidden Markov models to determine the segmentation of the genome to then narrow in on segments that could contain genes.

Project Proposal

Team: Alex Bishka, Eve Kazarian
Research question: What is known about a vaccine for COVID-19?

Methods: We plan on looking through the literature on COVID-19, filtering for documents pertaining to vaccination. Once the documents that are irrelevant to our research are filtered out, we plan on performing PCA and LDA on the data.

After using PCA to separate the dataset of research projects into topics, we will choose the number of dimensions of our PCA to be three and get a visualization of the topic distribution in three dimensions. We can then analyze the articles in the topics to vaguely discern each topic's

category. If we find a cluster about vaccines, we will look specifically into that cluster to get the latest research in that area.

We are building off an already-existing literature clustering project on Kaggle but modifying it to plot in 3D, changing our choice of stop words and number of clusters, and where exactly in the clusters we do our analysis.

Sources:
Dataset

Article clustering kernel on Kaggle: https://www.kaggle.com/maksimeren/covid-19-literature-clustering

> *Summary:* The author of this project used PCA and LDA to visualize the topic clustering of research articles about COVID-19. After visualizing the distinct clusters, they looked into each of them to assign topics. Topics ranged from surveillance to therapeutics.

3D PCA on Kaggle:
https://www.kaggle.com/luisblanche/cord-19-use-doc2vec-to-match-articles-to-tasks