Shanni Lam
Math 189Z
23 April 2020

Homework 3 Summaries

Hidden Markov Model For Stock Trading Paper Summary

Most models require stationary time series, which rarely occurs, so a HMM comes in handy. Nguyen used the AIC, BIC, HQC, and CAIC to evaluate the HMM's performance in predicting stock prices when used on between two to six states (open, close, high, closing price). As a result, his HMM predicted the S&P 500 price better than the historical average return model (HAR). It also is better at stock trading by yielding higher percentage returns. A HMM is a stochastic signal model where each observation at t is made at a finite hidden state. The HMM uses 4 algorithms, the forward, backward, Viterbi, and Baum-Welch, the last of which is used most heavily for multiple observations that may not be independent, while the others were used for independent observations. The Baum-Welch algorithm, also known as the EM method, maximized the log-likelihood of the model. It is important to choose the number of states to use for a stock before applying the HMM to predict the stock's price. Generally, the lower the criterion value, the better the model fits, and according to the graphs, having 4 states worked best.

Nguyen performed many statistical tests to check if the HMM was actually a better model than the HAR model. The out of sample R^2 value tells if the current model outperforms another model; during the out of sample forecasting period. If they got a positive value, meaning the HMM outperformed the HAR. Their p-value was smaller than the significance level of .001, so they accepted the alternative hypothesis that their R^2 value was positive. The cumulative squared predictive errors (CSPE's) shows the efficiency of two competing models for each point estimation. If the function increases, the HMM outperforms the HAR. Their model has some ups and downs but generally had an upward trend CSPE. Nguyen also calculated the absolute percentage error (APE), average absolute error (AAE), average relative percentage error (ARPE), and root mean square error (RMSE) for the HMM model and HAR to calculate the efficiency measure to compare the HMM with the HAR model. Since the efficiency statistic was positive, he concluded that the HMM was a better model.

Gene Finding and the Hidden Markov Models Paper Summary

Searching for genes in prokaryotes is easier than in eukaryotes due to the high density of genes in prokaryotic genomes; more than 90% of the genome is coding. Prokaryotic genes consist of an open reading frame, lacking the structure of eukaryotic genes that have introns and exons. So, to find genes in prokaryotes, just identify all open reading frames and according to a level of significance decide about which of them are genes. The null hypothesis would be that the tested ORF is random, while the alternative hypothesis would be that the ORF is not random, meaning there is a gene. On the other

hand, since gene finding for eukaryotes is harder a HMM can help do segmentation, where the HMM finds the hidden structure of regions with different biochemical makeup.

The Viterbi Algorithm determines the maximum likelihood hidden sequence by recursion and tabular computation. It takes observable sequences, initial probabilities for hidden states p, transition matrix of hidden states T, and emission matrix observable symbols E to get an output of a sequence of hidden states maximum likelihood h*.

There are two ways to train HMM. Through supervised learning, it uses a set of known hidden sequences and then estimates unknown values. In unsupervised learning, it uses observable sequences, an initial estimate of the initial probabilities, transition matrix, and emission matrix, and then it gets the maximum likelihood hidden sequence. After that, repeat to get better estimates. These two techniques are used jointly. First, use a training set to get initial estimates for supervised learning. Next, do unsupervised learning to refine these initial parameters of the observed sequence.

Summary of Source for Final Project (Undercount of COVID-19 Cases)
Source Link:
https://ourworldindata.org/coronavirus?fbclid=IwAR0N-V37byOhETU6xOCYGVkHJPasWZQ_OGeLaOORls1AZPRhTaktWDo6flU

Here are reasons why the number of confirmed deaths is lower than the actual number:
After a death certificate has been completed, inspection by post-mortem or laboratory testing may be required to verify the cause of death; COVID-19 deaths are usually manually coded, and there can be delays in this process, especially when there is a large increase in the number of deaths.
COVID-19 deaths may be recorded differently in different countries (e.g. some countries may only count hospital deaths, whilst others have started to include deaths in homes)
The reported death figures on a given date does not necessarily show the number of new deaths on that day: this is due to delays in reporting. There is also limited testing.

Here is information to consider when analyzing data:
Is there no data – or it is just hard to find? Are negative results included? Are pending results included?
Do the figures include all tests conducted in the country, or only some?
Are all regions and laboratories within a country submitting data on the same basis?
What period do the published figures refer to?
Are there any issues that affect the comparability of the data over time?
What are the typical testing practices in the country? For instance, how many tests does a case investigation require? What are the eligibility criteria to be tested? Are health workers, or other specific groups, being routinely retested?
Might any of the information above be lost in translation?