MATH 189Z Homework 3

Jeremy Tsai

HMM for Stock Trading Article Summary:

A Hidden Markov Model was used to predict the fluctuations of the stock market and compared with the historical average model. Many other models have also been made in order to try and determine the best time to buy/sell stock, such as the exponential moving average or head and shoulders models. However, the stock time series data is nonstationary, so the models need to also be nonstationary in time. In the current study, they used an HMM with six states and kept track of the open, low, high, and closing prices of the S&P500. The HMM was then pitted against the HAR model to trade stock according to predictions and then the results were compared.

The HMM was introduced in 1966 and has several properties. There are observations at time *t*, finite hidden states that fit the first-order Markov property, transition probability matrix is constant, and observation at a certain time has a specific probability distribution linked to a hidden state. There are two types of HMM: discrete and continuous. If probabilities are continuous, then the HMM is continuous. In the study, they utilized the three HMM algorithms of forward, backward, and Baum-Welch, to tune parameters of the HMM based on observations.

The study trained the HMM with monthly data on the S&P500 gathered from Yahoo finance and calculated 4 criteria (AIC, BIC, HQC, CAIC). They trained in 10 year blocks and shifted the starting month up by 1 for a total of 120 times, calculating the criteria every time. They did this for HMM with varying states and found that one with 4 states worked best for them. They then used the HMM to predict stock prices and compared these predictions with the actual stock prices. To do the prediction there were 3 separate steps involved. The parameters were tuned with training data and probabilities of an observation were calculated. Another data set with similar probabilities is found and the difference of the stock prices between months is used to predict stock prices at a later time. To measure how accurate the predictions were when compared to the other model, they used an out-of-sample R^2 statistical method. A positive R^2 value would mean the HMM performed better than the HAR model. In the end, the R^2 values were positive so the HMM was able to perform better than the HAR model in predicting stock prices. Comparisons between the two models also included other error estimators like absolute percentage error, average absolute error, average relative percent error, and root mean square error. The HAR model beat the HMM model only in the APE indicator, showing that overall the HMM is better than the HAR model.

They then let these models trade stock on the S&P500. Purchasing/selling of stocks was done by the following rules: if the stock had a predicted positive return, then it would be bought, if the stock had a predicted negative return, then it would be sold. The HMM generated more profit than the HAR model and also beat the buy and hold technique where 100 shares are bought at the beginning and held until the end.

Gene Finding Article Summary:

Proteins are made from blocks of amino acids and the structure of the composition of amino acids determines protein functions. Proteins can be classified by an sequence of letters, where each letter represents an amino acid. This is commonly used in order to employ computational/statistical methods on proteins.

To assemble amino acids together in the right order to create a protein, information is taken from genes. The genetic code is made up of groups of codons, or three nucleotides, which code for specific amino acids. Open reading frames mark the groups of codons that begin with an ATG start codon and end with a stop codon. The process of finding genes in prokaryotes is then first finding all possible open reading frames, then using significance to classify them as genes or not. For example, longer ORFs have a higher probability of being a gene since it is less likely it happened by chance. Hypothesis testing is done where the null hypothesis is that the ORF is random while the alternate is that the ORF is a gene.

Gene finding in eukaryotes is harder than gene finding in prokaryotes since introns and exons exist. A common technique to find genes within eukaryotes is Hidden Markov Models. The concept behind the HMM when applied here is that the observed DNA sequence is based off of a hidden sequence. The hidden sequence itself follows a Markov model where a symbol depends on the previous symbol. The observed sequence is then generated by the corresponding hidden symbol.

The HMM has observable states, hidden states, an emission matrix, a transition matrix, and an initial probability distribution of hidden states. Using the HMM, the hidden sequence can be derived from an observed sequence, called the maximum likelihood hidden sequence. The Viterbi algorithm is used to determine this sequence. It utilizes dynamic programming, where a large problem is broken up into smaller problems which are solved first, then aggregated to get the final answer. Since exons and introns exist within the sequences of eukaryotes, standard gene finding methods cannot be used and the HMM is employed to segment the exons and introns.

Hidden states are used so that segments in a sequence are tied to a hidden state. The transition between hidden states models when one segment transitions to another segment. If we know the initial probabilities, the transition matrix, the emission matrix, and have an observable sequence, the Viterbi algorithm can generate the maximum likelihood hidden sequence.

HMM with two states was used to model transmembrane proteins, where one state corresponds to hydrophobic segments while the other state corresponds to hydrophilic segments. The HMM can be trained with supervised learning or unsupervised learning. For supervised learning, it uses sequences where the hidden sequence is known. For unsupervised learning, it has an initial estimate for probabilities, transition matrix, and emission matrix, and then the observed sequence/maximum likelihood hidden sequence are used to make better predictions. This is done iteratively for a set number of times. The HMM was used on various transmembrane proteins in cyanobacteria and it was able to segment out various sections correctly (although there were definite errors present too).

Source Summary: https://monkeylearn.com/blog/sentiment-analysis-of-twitter/#What-is-Sentiment-Analysis

This source goes over the basics of sentiment analysis and also how it can be used to analyze tweets. Sentiment analysis is a process that is used to extract data from texts. The most common type is polarity detection which classifies text into three groups: positive, negative, or neutral. Lots of companies use

sentiment analysis to get an idea of their customers' perception of their goods/services. Since social media is commonly used to express opinions, it is helpful to companies to do sentiment analysis on these platforms, such as Twitter.

To actually conduct sentiment analysis on tweets, you must first gather your data, prepare your data, create an analysis model, and lastly, visualize the results. To actually gather the data, there are various services that make it easy to search for tweets within a certain timeframe or about a certain topic. Twitter also has their own API that you can use to get a stream of tweets. To prepare the data, you should remove information that is not relevant, like emojis, duplicates, spam, and blank spaces. You then need to pick what type of text analysis you want to do and the appropriate model to do so. For example, there is topic classification, sentiment analysis, and intent classification. You can then train your model on the data and test its accuracy.