Hidden Markov Model

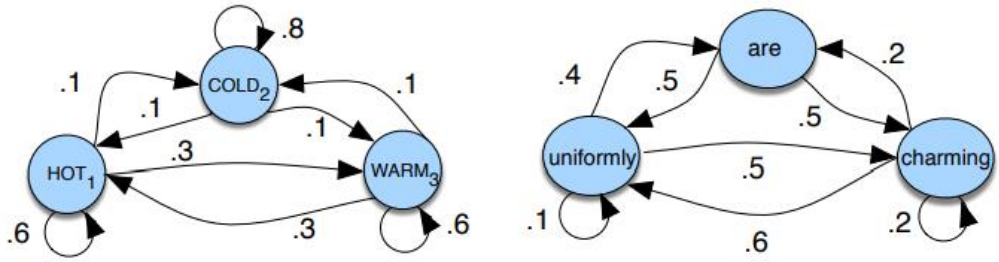
Subham Roy
Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260

1 Overview

Hidden Markov models (HMMs), named after the Russian mathematician Andrey Andreyevich Markov, who developed much of relevant statistical theory, are introduced, and studied in the early 1970s. They were first used in speech recognition and have been successfully applied to the analysis of biological sequences since the late 1980s.

2 About Hidden Markov Model

Hidden Markov Models (HMMs) are a class of probabilistic graphical model that allow us to predict a sequence of unknown (hidden) variables from a set of observed variables. The HMM is based on augmenting the Markov chain. A Markov chain is a model that tells us something about the probabilities of sequences of random variables, states, each of which can take on values from some set. These sets can be words, or tags, or symbols representing anything, like the weather. A Markov chain makes a very strong assumption that if we want to predict the future in the sequence, all that matters is the current state. The states before the current state have no impact on the future except via the current state. It's as if to predict tomorrow's weather you could examine today's weather, but you weren't allowed to look at yesterday's weather.



A Markov chain for weather (a) and one for words (b), showing states and transitions. A start distribution π is required; setting π = [0.1, 0.7, 0.2] for (a) would mean a probability 0.7 of starting in state 2 (cold), probability 0.1 of starting in state 1 (hot), etc.

A Markov chain is useful when we need to compute a probability for a sequence of observable events. In many cases, however, the events we are interested in are hidden and we don't observe them directly. A hidden Markov model (HMM) allows us to talk about both observed events (like words that we see in the input) and hidden events (like part-of-speech tags) that we think of as causal factors in our probabilistic model. An HMM is specified by the following components:

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} \ldots a_{ij} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, \ldots, v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

3 Experiment

The dataset used is Gold Price dataset. This dataset has 11151 rows and 2 variables. The features in this dataset are: 'datetime' and 'gold price (in USD)'.
'datetime': This column consists of dates between '1978-12-29' to '2021-09-24'.
'gold_price_usd': This column consists of gold price for respective dates.
This is a time series problem, and we wish to predict the future gold prices with the help of historical/past data. This dataset does not have null values for any columns. The 'gold_price_usd' column is of float type and 'datetime' is of object type. Since 'datetime' column is object type, we transformed it to datetime format. Also, we created a 'gold_price_change' column where we are taking the delta of the price change also finding a daily price change column to fit the model because it will define the change in the prices very well.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11151 entries, 0 to 11150
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   datetime        11151 non-null  object
 1   gold_price_usd  11151 non-null  float64
dtypes: float64(1), object(1)
memory usage: 174.4+ KB
```

	datetime	gold_price_usd	gold_price_change
0	1978-12-29	137.06	NaN
1	1979-01-01	137.06	0.00
2	1979-01-02	137.29	0.23
3	1979-01-03	134.01	-3.28
4	1979-01-04	136.79	2.78

	gold_price_usd	gold_price_change
count	11151.000000	11150.000000
mean	576.028117	0.121454
std	384.936088	7.205561
min	133.830000	-106.170000
25%	305.860000	-2.250000
50%	366.130000	0.000000
75%	951.325000	2.440000
max	1745.460000	71.500000

Fig : (Left) Dataset description, (Middle) few records within the dataset, and (Right) the Quantiles

For current analysis, we are restricting the data to starting from the year 2008 and onwards to cover two major Global impacts like the Great Recession of 2008 and the Pandemic from 2020.

We see from the plot on the left that there is a big dip during 2008 and heavy fluctuations during 2020. So, we consider our target variable to be 'gold_price_change' — this allows us to capture the market volatility. Daily changes in gold prices are now fitted to a Gaussian emissions model with 3 hidden states. The reason for using 3 hidden states is that we expect at the very least 3 different regimes in the daily changes — low, medium, and high volatility.

4 Analysis

Now let us analyze the HMM model prediction based on the plots below. As expected, we see the periods of high volatility correspond between 2008 to 2009, the recession of 2011–2012 and the covid-19 pandemic in 2020. Furthermore, we notice the price of gold tends to rise during times of uncertainty as investors increase their purchases of gold which is seen as a stable and safe asset.

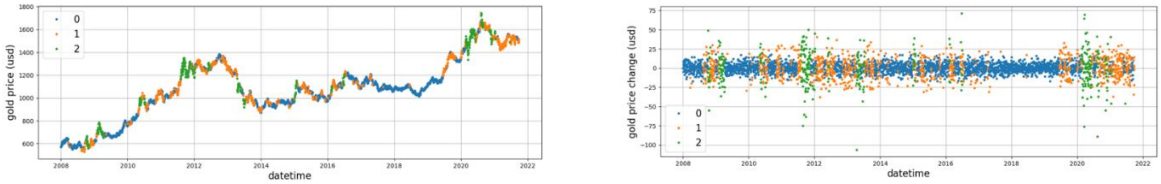


Fig:(Left) 0,1, and 2 signifies Low, Medium and High gold prices and the plot shows the day-to-day trend of the gold prices. (Right) Mean Square error fluctuations for the Gold price changes.

References

[1] Gareth James [at], Daniela Witten [aut], Trevor Hastie [aut, cre], Rob Tibshirani [aut], Balasubramanian Narasimhan [ctb], *Introduction to Statistical Learning, Second Edition*.

[2] Tom M. Mitchell, *Machine Learning: A multistrategy approach, 1997 Edition*.

[3] URL - https://github.com/cchangyou/hmmlearn/tree/main/examples ; https://medium.com/@postsanjay/hidden-markov-models-simplified-c3f58728caab