Hidden Markov Models, I. Examples

Steven R. Dunbar

Hidden Markov Models, I. Examples

Steven R. Dunbar

February 17, 2017

Outline

Examples

A good stack of examples, as large as possible, is indispensable for a thorough understanding of any concept, and when I want to learn something new, I make it my first job to build one. – Paul Halmos

The Variable Factory 1

Production process in a factory

- If in good state 0 then, independent of the past, with probability 0.9 it will be in state 0 during the next period,
- with probability 0.1 it will be in bad state 1,
- once in state 1 it remains in that state forever.

Classic Markov chain:

$$A = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} 0.9 & 0.1 \\ 0 & 1 \end{pmatrix} \end{array}.$$

The Variable Factory 2

- Each item produced in state 0 is acceptable quality with probability 0.99.
- Each item produced in state 1 is acceptable quality with probability 0.96.

$$\mathbb{P}\left[u \mid 0\right] = 0.01 \qquad \mathbb{P}\left[a \mid 0\right] = 0.99$$
$$\mathbb{P}\left[u \mid 1\right] = 0.04 \qquad \mathbb{P}\left[a \mid 1\right] = 0.96$$

The Variable Factory 3

Probability of starting in state 0 is 0.8.

A sequence of three observed articles are (a, u, a). Then given the possible state sequence, the probability of the observed sequence

000	(0.8)(0.99)(0.9)(0.01)(0.9)(0.99)	0.006351
001	(0.8)(0.99)(0.9)(0.01)(0.1)(0.96)	0.0006843
010	(0.8)(0.99)(0.1)(0.04)(0.0)(0.99)	0
011	(0.8)(0.99)(0.1)(0.04)(1.0)(0.96)	0.0030413
100	(0.2)(0.96)(0.0)(0.01)(0.9)(0.99)	0
101	(0.2)(0.96)(0.0)(0.01)(0.1)(0.96)	0
110	(0.2)(0.96)(1.0)(0.96)(0.0)(0.99)	0
111	(0.2)(0.96)(1.0)(0.04)(1.0)(0.96)	0.007373

The Variable Factory 4

Although we can't observe the state of the factory (expensive), we can observe the product quality.

$$\mathbb{P}\left[X_3 = 0 \,|\, (a, u, a)\right] = \frac{0.006351 + 0 + 0 + 0}{\sum_{\text{state seq}} \mathbb{P}\left[\text{seq}\right]} = 0.3640.$$

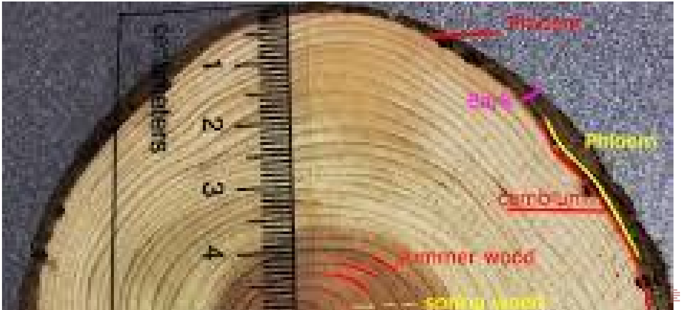
Notice that even without the 0 table entries there are calculations which can combine to make the calculations more efficient.

A Paleontological Temperature Model 1

Determine the average annual temperature at a particular location over a sequence of years in the distant past.

Only 2 annual average temperatures, "hot" and "cold".

Use correlation between the size of tree growth rings and temperature.



Temperature Model 2

- Probability of a hot year followed by hot year is 0.7 and
- probability of a cold year followed by cold year is 0.6,
- independent of the temperature in prior years.

Classic Markov Chain:

$$\begin{array}{cc} & \begin{array}{cc} H & C \end{array} \\ \begin{array}{c} H \\ C \end{array} & \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix} \end{array}.$$

Temperature Model 3

- 3 tree ring sizes, small S, medium M, and large L,
- the observable *signal* of the average annual temperature.
- the probabilistic relationship between annual temperature and tree ring sizes is

$$\begin{array}{c c} & \begin{array}{ccc} S & M & L \end{array} \\ \begin{array}{c} H \\ C \end{array} & \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{pmatrix}. \end{array}$$

Temperature Model 4

The annual average temperatures are hidden states, since we can't directly observe the temperature in the past.

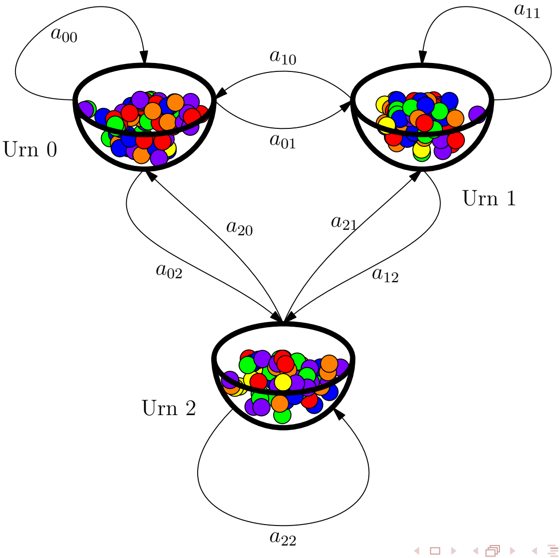
Although we can't observe the state or temperature in the past, we can observe the sequence of tree ring sizes.

From this evidence, we would like to determine the most likely temperature states in past years.

Ball and Urn Model 1

- N urns, each with a number of colored balls;
- M possible colors for the balls;
- initially choose one urn, from an initial probability distribution;
- select a ball, record color, replace the ball;
- choose a new urn according to a transition probability distribution;
- the signal or observation is the color of the selected ball;
- the hidden states are the urns.

Ball and Urn Model 2



Coin Flip Model

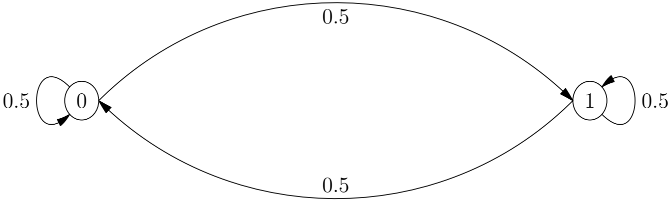
Coin flipping experiment:

- You are in a room with a curtain through which you cannot see what is happening.
- On the other side of the curtain is a person flipping a coin, or maybe one of several coins.
- The other person will not tell us exactly what is happening, only the result of each coin flip.
- How do we build a Hidden Markov Model to best explain the observed sequence of heads and tails?

Simplest one coin model

- Special case: proportion of heads in the observation sequence is 0.5
- Two states, each state solely associated with heads or tails.
- This model is not hidden because the observations directly define the state.
- This is a degenerate example (standard Bernoulli trials) of a hidden Markov model

Simplest one coin model



$P[H] = 1$

$P[T] = 0$

$P[H] = 0$

$P[T] = 1$

Two-fair-coins Model

- There are again two states in the model.
- special case: each state has a fair coin, so the probability of a head in each state is 0.5.

The probabilities associated with remaining in or leaving each of the two states form a probability transition matrix whose entries are unimportant because the observable sequences from the 2-fair coins model are identical in each of the states.

- That means this special case of the 2-fair-coin model is indistinguishable from the 1-fair-coin model in a statistical sense
- so this is another degenerate example of a Hidden Markov Model

2-biased-coins Model

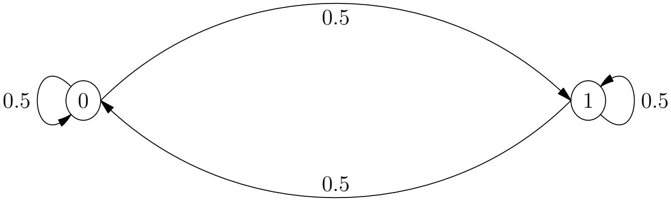
The person behind the curtain has two biased coins and a third fair coin with the biased coins associated to the faces of the fair coin respectively.

The person behind the barrier flips the fair coin to decide which biased coin to use and then flips the chosen biased coin to generate the observed outcome.

If $p_0 = 1 - p_1$ (e.g. $p_0 = 0.6$ and $p_1 = 0.4$) the long term averages of heads would be statistically indistinguishable from either the 1-fair-coin model or the 2-fair-coin model.

Other higher-order statistics of the 2-biased-coins model, such as the probability of runs of heads, *should be distinguishable* from the 1-fair-coin or the 2-fair coin model.

2-biased-coins Model



$P[H \mid 0] = p$

$P[T \mid 0] = 1 - p$

$P[H \mid 1] = 1 - p$

$P[T \mid 1] = p$

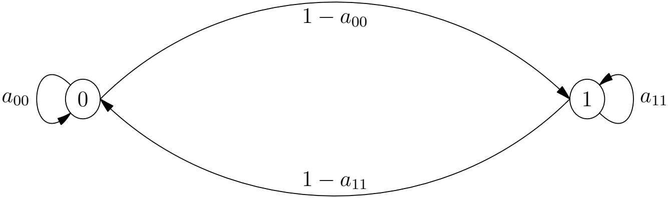
2-biased-coins Model

Now suppose observed sequence has a very long sequence of heads, then followed by another long sequence of tails of some random length, interrupted by a head, followed again by yet another long random sequence of tails.

2-coin model with two biased coins, with biased switching between the states of the coins would be a possible model for the observed sequence.

Such a sequence of many heads followed by many tails could conceivably come from one fair coin. The choice between a 1-fair-coin or 2-biased-coins model would be a choice justified by the likelihoods of the observations under the models, or possibly by other external modeling considerations.

3-biased coin Model



$P[H \mid 0] = p_0$ $\qquad$ $P[H \mid 1] = p_1$

$P[T \mid 0] = 1 - p_0$ $\qquad$ $P[T \mid 1] = 1 - p_1$

3-biased coin Model

Another Hidden Markov Model could be the "3-biased-coins model".

- First state, the coin is slightly biased to heads,
- Second state the coin is slightly biased toward tails, in the
- Third state the coin is some other distribution, maybe fair, maybe

biased.

- A Markov chain for transition probabilities among the three states.

The simple statistics and higher-order statistics of the observations would be correspondingly influenced and

Modeling Considerations 1

No reason to stop at 1, 2, or even 3 coin models.

Part of the modeling process is to decide on the number of states N for the model.

Without some *a priori* information, this choice is often difficult to make and may involve trial and error before settling on the appropriate model size.

Modeling Considerations 2

One goal is optimal estimation of the model parameters from the observations, that is, the probabilities of heads and tails in each state and the transition probabilities between states.

The choice of what "optimal" means is a mathematical modeling choice:

- maximize the number of expected states that are correctly predicted (forward-backward algorithm)
- Choose the sequence of states whose conditional probability, given the observations, is maximal (Viterbi algorithm)

After choice of optimal, parameters estimation is a statistical problem.

Modeling Considerations 3

The length of the observation sequence.

With a too short observation sequence, we may not be able to reliably estimate the model parameters or even the number of states.

With insufficient data, some Hidden Markov Models may not be statistically different.

Modeling Considerations 4

Finally, this emphasizes the title of the subject as Hidden Markov *Models*.

If the Hidden Markov Model is completely specified, then one might as well make a larger-state ordinary Markov process from it.

- The completely specified 2-coin model biased, could be a four-state Markov process.
- The three-coin model could be a 6-state Markov process
- the classic results of Markov processes would apply

Modeling Considerations 5

Here we have *only* the observations or signals, *not* all the necessary information, even the number of states. From that, we wish to best determine the underlying states and probabilities. The word "best" indicates choices of measures of optimality. So this is a modeling and statistical problem.

That accounts for calling these Hidden Markov *Models*, not considering them from the viewpoint of Markov processes.

CpG Islands 1

In the human genome the dinucleotide CpG (written CpG to distinguish it from the CG base-pair across two strands) is rarer than would be expected from the independent probabilities of C and G, for reasons of chemistry.

CpG Islands 2

For biologically important reasons, the chemistry is suppressed in short regions of the genome, such as around the promoters or start regions of many genes. In these regions, we see many more `CpG` dinucleotides than elsewhere.

Such regions are called `CpG` islands. They are typically a few hundred to a few thousand bases long.

CpG Islands 3

Given a short stretch of genomic sequence, how would we decide if it comes from a `CpG` island or not?

Second, given a long piece of sequence, how would we find the `CpG` islands in it, if there are any?

Note the similarity to the coin toss example.

Language Analysis and Translation

Language translation is a classic application of Hidden Markov Models,

You obtain a large body of English text, such as the "Brown Corpus" with 500 samples of English-language text, totaling roughly one million words, compiled from works published in the United States in 1961.

With knowledge of Hidden Markov Models, but no knowledge of English, you would like to determine some basic properties of this mysterious writing system. Can you partition the characters into sets so that characters in each set are "different" in some statistically significant way?

Language Analysis 2 (Experiment)

- Remove punctuation, numbers; convert all letters to lower case.
- Leaves 26 distinct letters and the space, total of 27 symbols.
- Test the hypothesis that English text has an underlying Markov chain with two states.
- For each of these two hidden states, assume that the 27 symbols appear according to a fixed probability distribution (not same).

This sets up a Hidden Markov Model with $N = 2$ and $M = 27$ where the state transition probabilities and the observation probabilities from each state are unknown.

Language Analysis 3

	State 0	State 1
a	0.13845	0.00075
b	0.00000	0.02311
c	0.00062	0.05614
d	0.00000	0.06937
e	0.21404	0.00000
f	0.00000	0.03559
g	0.00081	0.02724
h	0.00066	0.07278
i	0.12275	0.00000
j	0.00000	0.00365
k	0.00182	0.00703
l	0.00049	0.07231
m	0.00000	0.03889
n	0.00000	0.11461
o	0.13156	0.00000
p	0.00040	0.03674
q	0.00000	0.00153
r	0.00000	0.10225
s	0.00000	0.11042
t	0.01102	0.14392
u	0.04508	0.00000
v	0.00000	0.01621
w	0.00000	0.02303
x	0.00000	0.00447
y	0.00019	0.02587
z	0.00000	0.00110
space	0.33211	0.01298

Language Analysis 4

Results of a case study using the first $50{,}000$ observations of letters from the Brown Corpus.

Without having any assumption about the nature of the two states, the probabilities tell us that the one hidden state contains the vowels while the other hidden state contains the consonants.

Space is more like a vowel and "y" is a consonant.

The Hidden Markov Model "deduces" the statistically significant difference between vowels and consonants without knowing anything about the English language.

Speech Recognition 1

Hidden Markov Models were developed in the 1960s and 1970s for satellite communication (Viterbi, 1967). They were later adapted for language analysis and translation and speech recognition in the 1970s and 1980s (Bell Labs, IBM). Interest in HMMs for speech recognition seems to have peaked in the late 1980s.

It is *not* clear if current (2017) speech recognition relies on HMMs at any level or in any way.

Speech Recognition 2

1. Feature analysis – spectral or temporal analysis of speech signal
2. Unit matching – the speech signal parts are matched to words or phonemes with an HMM
3. Lexical analysis – if units are phonemes, combine into recognized words.
4. Syntactic analysis – with a grammar, group words into proper sequences (if single word like "yes" or "no", or digit sequences, this is minimal)
5. Semantic analysis – interpret for the task model.

Speech Recognition 3

Example: Isolated word recognition: With Viterbi Algorithm, a vocabulary of $V = 100$ words with an $N = 5$ state model, and $T = 40$ observations, it takes about 10^5 computations (addition/multiplication) for a single word recognition.

Speech Recognition 4

It is hard to determine what current (2017) speech recognition is based on. Explanations are clouded with buzzwords and hype, with no theory.

$$\begin{aligned}\text{Artificial Intelligence} &\supseteq \text{Machine Learning} \\ &\supseteq \text{Neural Networks} \supseteq \text{Deep Learning}\end{aligned}$$

In a nutshell, neural networks take a high dimensional problem and turn it into a low-dimensional problem:

`.\Kernel_Machine.pdf`

There does not seem to be a connection to HMMs.