Markov Processes

Daniel Wysocki

March 12, 2015

1. Stochastic Processes
2. Markov Chains
3. Hidden Markov Models

Stochastic Processes

Definition

A stochastic process is a collection of random variables $\{X(t), t \in T\}$ defined on a common probability space indexed by the index set T which describes the evolution of some system.

(Resnick, 1992)

Properties

- the evolution of the system is non-deterministic, even if all initial variables are known
- the system may evolve in many (possibly infinite) ways
- any given sequence of events may be assigned a probability
- the probability of all possible sequences must sum to 1

History

- early motivation for the theory of stochastic processes lies in Brownian motion
- Brownian motion was a phenomenon observed by Robert Brown in 1827
    - he observed under a microscope the motions of pollen in a fluid
    - the movements seemed unpredictable
- Brownian motion is the outcome of many unpredictable or unobservable events, each imparting a negligible influence on an observed phenomenon, but collectively having a significant influence

(Paul and Baschnagel, 2013)

History (cont.)

- Louis Bachelier made his PhD thesis in 1900, applying stochastic theory to the price of financial assets
    - the significance of his work was not realized until much later
    - it contained many of the results of stochastic theory used today
    - his advisor, Henri Poincaré, when trying to solve the Brownian motion problem years later, did not realize Bachelier had already solved it

(Paul and Baschnagel, 2013)

History (cont.)

- in his 1906 article in *Annals of Physics*, Albert Einstein described the mechanisms which drive Brownian molecular motion
- this helped strengthen the evidence for atoms and molecules
- presented a method for measuring the size of the atom

(Einstein, 1956)

Markov Processes

- a special case of stochastic processes
- satisfy the property that the future is dependent only on the present, not the past
- may not be a perfect model of the underlying stochastic process, but often serves as a good approximation, and reduces the complexity of the model

(Resnick, 1992)

History

- Andrey Markov, a Russian mathematician, developed a technique, now known as the Markov chain
- he used it to discover patterns in the vowels and consonants of Alexander Pushkin's novel, *Eugene Onegin*
    - published a summary of his findings in an address to the Imperial Academy of Sciences in St. Petersberg in 1913
- though his work did little to understand the works of Pushkin, it started a new branch of probability theory, in wide use today

(Hayes and others, 2013)

Markov Chains

Definition

- Markov chains are Markov processes with a discrete index set T, and a countable or finite state space
- the Markov chain can be described by only two features: a 1-dimensional initial probability distribution, $a(k)$, and a 2-dimensional transition probability distribution, $p(i,j)$
    - $a(k)$ is the probability that the first element of the chain is in state k
    - $p(i,j)$ is the probability that a system in state i will transition into j

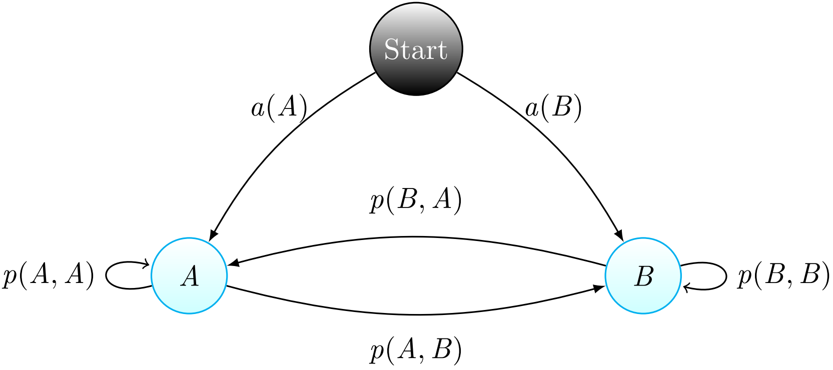
$$\sum_{k=1}^{N} a(k) = \sum_{j=1}^{N} p(i,j) = 1,$$

where N is the number of possible states

(Resnick, 1992)

Graph Representation

A simple Markov process, with two states A and B, can be represented by this graph



Training a Model

- a Markov chain may be constructed by observing a sequence of events, and constructing a transition frequency matrix and an initial frequency vector
- then we normalize the two, such that the rows sum to 1
  - row stochastic
- consider the observed sequences (A, A, B, A, B) and (B, A, B, B, A)

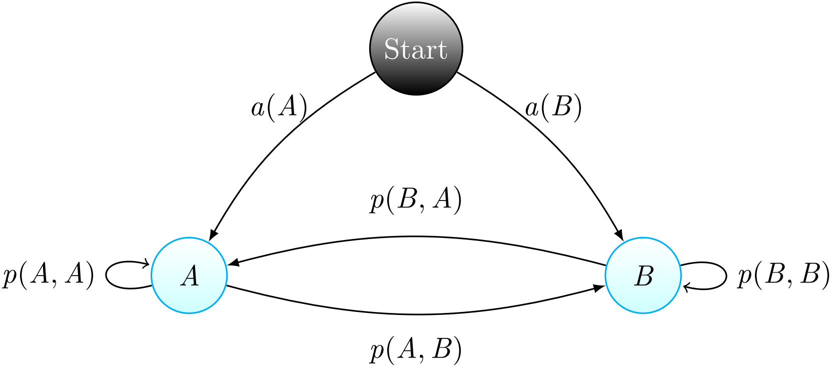
$$\tilde{a} = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \tilde{p} = \begin{bmatrix} 1 & 3 \\ 1 & 3 \end{bmatrix} \implies a = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix}, \quad p = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix}$$

Generating Sequences

- once we have trained a model, we may use it to generate random sequences, which we shall denote as X
- we utilize a random number generator, and with probability $a(k)$ we choose initial state k, so $X_0 = k$
- subsequent states can be found recursively, such that we choose state X_n with probability $p(X_{n-1}, X_n)$

Graphical Sequence Generation

We can envision the process of generating a sequence as traversing this graph. At each node, we use the associated transition probabilities to decide where to go next



Hidden Markov Models

Definition

In a hidden Markov model, the states in the Markov chain are not directly observable. However, there is a set of observables variables, which are somehow correlated with the underlying (or hidden) Markov chain.

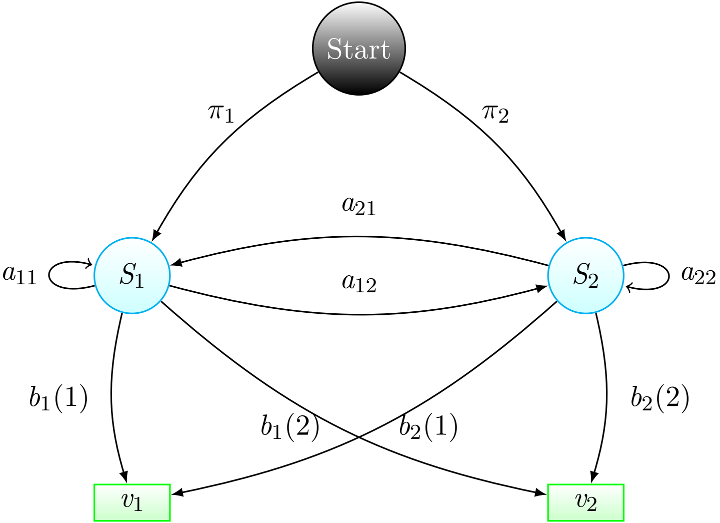
(Alpaydin, 2010)

History

- early theory of HMMs developed by Leonard E. Baum and colleagues at the Institute for Defense Analyses (IDA) in the 1960's
- Jim Baker applied HMMs to speech recognition in the 1970's, but was not 100% successful
- Jack Ferguson and colleagues at IDA gave a classical series of lectures in 1980, which boosted the popularity of HMMs
- continued usage in speech recognition
- now widely used in biological sequencing

(Rabiner, 2015; Yoon, 2009)

Graph Representation



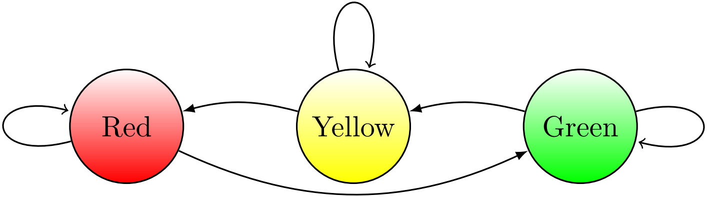
Parameters

- a hidden Markov process can be parameterized by the following variables
  - the states $S = \{S_1, S_2, \ldots, S_N\}$
  - the observation symbols $V = \{v_1, v_2, \ldots, v_M\}$
  - the state transition probabilities $\mathbf{A} = [a_{ij}]$, where $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$
  - the observation probabilities $\mathbf{B} = [b_j(m)]$, where $b_j(m) = P(O_t = v_m | q_t = S_j)$
  - the initial state probabilities $\mathbf{\Pi} = [\pi_i]$ where $\pi_i = P(q_1 = S_i)$
- the model itself is denoted $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$
- the observation sequence is denoted $O = (O_1, O_2, \ldots, O_T)$
- the hidden state sequence is denoted $Q = (q_1, q_2, \ldots, q_T)$

Traffic Light Markov Chain

Suppose we want to model a traffic light as a Markov process. This can be done using a very simple Markov chain.

There are 3 states, red, yellow, and green. Each state has 2 non-zero transition probabilities, either it remains the same, or changes in the progression green → yellow → red → green . . .

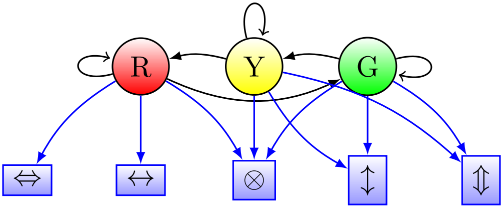


Traffic Light HMM

Suppose we want to model this same traffic light, but we only have a video stream from a satellite. The satellite cannot see the light itself, but can resolve the velocity of the moving cars, both in the same direction as the light, and the intersecting direction.

The hidden states are red, yellow, and green. The observables are fast same direction ($\Updownarrow$), slow same direction ($\updownarrow$), fast intersecting direction ($\Leftrightarrow$), slow intersecting direction ($\leftrightarrow$), all stopped ($\otimes$).

Assuming cars don't run red lights, the HMM might look something like.



Common HMM Problems

Some common problems occur in HMMs, which have standard solutions.

1. Given the model, λ, and a sequence of observations, O, determine the likelihood of observing that sequence, $P(O|\lambda)$
    - this is solved by the forward algorithm
    - might be used to detect unusual traffic
2. Given the model, λ, and a sequence of observations, O, find an optimal state sequence Q
    - this is solved by the backward algorithm
    - this will tell us the most likely sequence of traffic light states
3. Given an observation sequence, O, and the dimensions of the model, N and M, find the model λ that maximizes the probability of O
    - this is solved by the Baum–Welch algorithm
    - this is how we initially train the model

(Stamp, 2012)

Forward Algorithm

1. Let $\alpha_0(i) = \pi_i b_i(O_0)$, for $i = 0, 1, \ldots, N-1$
2. For $t = 1, 2, \ldots, T-1$ and $i = 0, 1, \ldots, N-1$, compute

$$\alpha_t(i) = \left[\sum_{j=0}^{N-1} \alpha_{t-1}(j) a_{ji} \right] b_i(O_t)$$

3. Since $\alpha_t(i) = P(O_0, O_1, \ldots, O_t, x_t = q_i | \lambda)$, it follows that

$$P(O|\lambda) = \sum_{i=0}^{N-1} \alpha_{T-1}(i)$$

(Stamp, 2012)

Backward Algorithm

1. Let $\beta_{T-1}(i) = 1$, for $i = 0, 1, \dots, N-1$.
2. For $t = T-2, T-3, \dots, 0$ and $i = 0, 1, \dots, N-1$, compute

$$\beta_t(i) = \sum_{j=0}^{N-1} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$$

3. For $t = 0, 1, \dots, T-2$ and $i = 0, 1, \dots, N-1$, define $\gamma_t(i) = P(x_t = q_i | O, \lambda)$, which can be written as

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$$

The most likely state at time t is the state q_i with maximum $\gamma_t(i)$.

(Stamp, 2012)

Baum–Welch Algorithm

- define "di-gamma" as

$$\gamma_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}$$

- this is related to gamma by

$$\gamma_t(i) = \sum_{j=0}^{N-1} \gamma_t(i,j)$$

- given γ and di-gamma, we can re-estimate a model using the following procedure

(Stamp, 2012)

Baum–Welch Algorithm

1. for $i = 0, 1, \ldots, N-1$, let

$$\pi_i = \gamma_0(i)$$

2. for $i = 0, 1, \ldots, N-1$ and $j = 0, 1, \ldots, N-1$, compute

$$a_{ij} = \frac{\sum_{t=0}^{T-2} \gamma_t(i, j)}{\sum_{t=0}^{T-2} \gamma_t(i)}$$

3. for $j = 0, 1, \ldots, N-1$ and $k = O_0, O_1, \ldots, O_{M-1}$, compute

$$b_j(k) = \frac{\sum_{\substack{t \in \{0,1,\ldots,T-2\} \\ O_t = k}} \gamma_t(j)}{\sum_{t=0}^{T-2} \gamma_t(j)}$$

(Stamp, 2012)

Baum–Welch Algorithm

Now that $\gamma_t(i,j)$ and re-estimation have been defined, we can write the entire algorithm as

1. initialize $\lambda = (A, B, \pi)$.
2. compute $\alpha_t(i)$, $\beta_t(i)$, $\gamma_t(i,j)$ and $\gamma_t(i)$
3. re-estimate the model according the the method just defined
4. if $P(O|\lambda)$ increases by some threshold, or max iterations reached, repeat from step 2

(Stamp, 2012)

Hidden Markov Music

- MIDI events (or some invertible function of them) are the observations
- a musical piece is an observation sequence O
- a model, λ, is trained on some set of musical pieces, $\mathbf{O}$
- the hidden states are completely determined by the observations, and might correspond to anything
  - the only human intervention is in choosing the number of states, N
- the forward algorithm can be used to find $P(O|\lambda)$ for a given piece, which may be interpreted as how well it matches the model (classification)
- new pieces may be generated from the model, by walking through the model, making choices based on transition and emission probabilities

[1] E. Alpaydin. "Introduction to Machine Learning, MIT Press". In: *Cambridge, MA, USA* (2010).

[2] A. Einstein. *Investigations on the Theory of the Brownian Movement.* Courier Corporation, 1956.

[3] B. Hayes and others. "First links in the Markov chain". In: *American Scientist* 101.2 (2013), p. 92.

[4] W. Paul and J. Baschnagel. *Stochastic Processes: From Physics to Finance.* Springer, 2013. ISBN: 9783319003276. .

[5] L. Rabiner. *First-Hand:The Hidden Markov Model.* 2015. (visited on 03/11/2015).

[6] S. I. Resnick. *Adventures in stochastic processes.* Springer Science & Business Media, 1992.

[7] M. Stamp. *A Revealing Introduction to Hidden Markov Models.* 2012. (visited on 02/24/2015).

[8] B. Yoon. "Hidden Markov Models and their Applications in Biological Sequence Analysis". In: *Current Genomics* 10.6 (2009), pp. 402-415.