

1 Markov Chains

Have a sequence of random variables. Each random variable takes its value from a set of states. Markov property is that the distribution of values for each random variable is affected only by the outcome of the value of the most recent random variable. That is

$$P(X_i = s_a | X_{i-1} = x, X_{i-2} = y, \dots, X_1 = z) = P(X_i = s_a | X_{i-1} = x)$$

Example: weather in the land of Oz. Three states: Rain, Nice, Snow.

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$$

If we know a particular initial distribution of states according to state vector $\mathbf{u} = [1/31/31/3]$, we can get the probability of being in each state after n steps by $\mathbf{u}^{(n)} = \mathbf{u}\mathbf{P}^n$. (EXAMPLE OF HOW STATE VECTORS EVOLVE) More generally, we can multiply this matrix by itself \mathbf{P}^n to get probability of getting from state i to state j in n steps. (EXAMPLE)

2 Stationary Distributions for Regular (Ergodic) Markov Chains

Definition. A Markov chain is regular iff for there is some n such that \mathbf{P}^n has only positive elements (i.e., it's possible to get from any state to any other)

Theorem. For regular Markov Chain with transitional probability matrix \mathbf{P}

1. there is a unique long term state vector x^* , called the stationary distribution such that $x^*\mathbf{P} = x^*$
2. No matter what the starting distribution x_0 , the sequence of state vectors

$$x_0, x_1, x_2, \dots$$

will converge to x^* .

How do we compute x^* ? Let's write x^* as (w, x, y) . Thus, we need to solve the following equation:

$$(abc) \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} = (abc)$$

This leads to the following equations in 3 unknowns:

$$\begin{aligned} (1/2)a + (1/2)b + (1/4)c &= a \\ (1/4)a + (1/4)c &= b \\ (1/4)a + (1/2)b + (1/2)c &= c \end{aligned}$$

We need only solve this (almost). But we need to guarantee that

$$a + b + c = 1$$

Solve this using Matlab!

Another example: Google Page rank (probability in the stationary distribution to be at page p_i given the following transitional probability)

$$P(p_k|p_i) = \begin{cases} \frac{1-q}{k_i} + \frac{q}{N} & \text{if } (p_j, p_k) \in L \\ \frac{q}{N} & \text{o.w.} \end{cases}$$

where N is the number of web pages, k_i is the number of links out of p_i , and q is a parameter (taken to be 0.15 according to Wikipedia).

3 Absorbing Markov Chains

Definition. A state s_i of a Markov chain is absorbing if it is impossible to leave it. (i.e., $p_{ii} = 1$). A Markov chain is absorbing if it has at least one absorbing state, and if it is possible to go to an absorbing state from every state.

Definition. In an absorbing Markov chain, a state which is not absorbing is called transient.

Example: Flip a fair coin forever, stopping when you get heads. What's the probability that the first H is on an even numbered flip? Represent this as a Markov chain with 4 states:

1. an even number of Ts and no H
2. an odd number of Ts and no H.
3. H after an even number of T flips
4. H after an odd number of T flips

DRAW diagram. This transition probabilities will be as follows:

$$\begin{pmatrix} 0 & .5 & .5 & 0 \\ .5 & 0 & 0 & .5 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

We might wonder about a number of properties of this chain: what's the probability that the process will end up in a given absorbing state? On average, how long will it take for the process to be absorbed? On average, how many times will the process pass through each transient state? Answers depend on starting state as well as transition probabilities.

Canonical form of a transition matrix for an absorbing Markov chain:

$$\begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ 0 & \mathbf{I} \end{pmatrix}$$

where \mathbf{Q} is transition matrix among transient states, and \mathbf{R} is transition matrix from transient states to absorbing states.

Theorem. In an absorbing Markov chain, the probability that the process will be absorbed is 1. (i.e., $Q^n \rightarrow 0$ as $n \rightarrow \infty$).

Theorem. For an absorbing Markov chain, the matrix $I - Q$ has an inverse N and $N = I + Q + Q^2 + \dots$. The ij -entry n_{ij} of matrix N is the expected number of times the chain is in state s_j given that it starts in state s_i . The initial state is counted if $i = j$.

For the coin flip example,

$$N = \begin{pmatrix} 1.3333 & 0.6667 \\ 0.6667 & 1.3333 \end{pmatrix}$$

Theorem. Let t_i be the expected number of steps before the chain is absorbed, given that the chain starts in s_i , and let \mathbf{t} be the column vector whose i th entry is t_i . Then

$$\mathbf{t} = N\mathbf{c}$$

where \mathbf{c} is a column vector all of whose entries are 1.

For the coin flip example,

$$\mathbf{t} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Theorem. Let b_{ij} be the probability that an absorbing chain will be absorbed in the absorbing state s_j if it starts in the transient state s_i . Let \mathbf{B} be the matrix with entries b_{ij} . Then \mathbf{B} is a t -by- r matrix and

$$\mathbf{B} = \mathbf{N}\mathbf{R}$$

where \mathbf{N} is the fundamental matrix and \mathbf{R} is as in the canonical form.

For the coin flip example,

$$\mathbf{B} = \begin{pmatrix} 0.6667 & 0.3333 \\ 0.3333 & 0.6667 \end{pmatrix}$$

For a cognitive example, see Niyogi and Berwick (1996) on parameter setting.

4 Markov chains and Ngrams

What is the goodness of a sequence of phonemes? Associate individual phonemes with states in a Markov chain. The probability of a sequence of phonemes is then the likelihood of that sequence of states in the Markov chain.

Because of Markov assumption, transition probabilities are bigram probabilities. So, the probability of a sequence will be the product of the bigram probabilities. (Question: how can this be generalized to n -gram probabilities?)

How do we estimate these probabilities?

1. *Maximum Likelihood Estimate (MLE)*

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Problems with data sparseness.

2. *Laplace/Add-1 Smoothing*

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n) + 1}{C(w_{n-1}) + V}$$

3. *Good-Turing discounting*: use frequency of things you've seen once as a re-estimate of the frequency of zero-count bigrams

$$N_c = \sum_{x:\text{count}(x)=c} 1$$

Under MLE, count for N_c is c . According to Good-Turing, we replace this with:

$$c^* = (c + 1) \frac{N_{c+1}}{N_c}$$

This leaves some missing mass, which we reassign to the zero occurrence items:

$$P_{GT}^*(\text{things with frequency zero in training}) = \frac{N_1}{N}$$

Divide this probability mass among the zero count items.