STAT 534

Lecture 7

Markov chains

May 2, 2018

Marina Meilă
mmp@stat.washington.edu

1 What is a Markov chain?

The **Markov chain** model or shortly the Markov model is a process where the outcomes of consecutive trials depend on each other.

For example, if the sample space S represents the set of rooms in a building, then from any room you can get only to neighboring rooms. Thus the *state* (i.e which room you are in) at time t will depend on the state at the previous time step. If it does not depend on anything else, this is a Markov chain.

Formally, we call S the **state space** and we have a **time** variable t taking values from $0, 1, 2, \ldots T$. At each moment t, we denote by X_t the outcome of the t-th experiment on S – which in this framework is called the **state** at time t. The state X_t depends on the previous state X_{t-1} for $t > 0$ and given X_{t-1} is independent of all the other states in the past.

$$X_t \perp X_{t-k} \, | X_{t-1} \text{ for all } k > 1 \tag{1}$$

and for all $t > 1$. In words this means: “the present makes the future independent on the past” and it is known as **the Markov property**.

From the Markov property we can derive (by induction) more general independence relationships of the form

$$X_r \perp X_t \mid X_s \text{ if } r < s < t \tag{2}$$

We can represent all the dependencies and independencies in a Markov chain (or **Markov model**) by a graph like the one figure 1.

2 Parametrization

Let us now construct a probability model that describes the Markov chain, that is a probability distribution over all sequences of states up to time T (for any T!). For this we start by introducing some notation.

Let $S = \{1, \ldots m\}$ and

$$p_i[t] = P_{X_t}(i) \tag{3}$$

the probability that at time t the chain is in state i for $i \in S$. Then

$$p[t] = [p_1[t] \ \ldots \ p_m[t]] \tag{4}$$

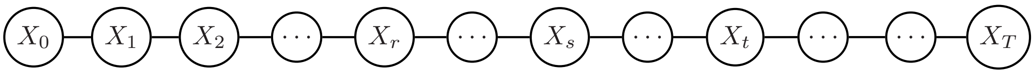

Figure 1: Graphical representation of a Markov chain. States X_r, X_t are (marginally) dependent, but they are independent conditioned on state X_t with $r < s < t$.

describes the probability distribution of state X_t and therefore

$$\sum_{i=1}^{m} p_i[t] \;=\; 1 \tag{5}$$

We call $p[0]$ the **initial probability**. Then we define the *transition probabilities* as

$$a_{ij}[t] \;=\; P_{X_t|X_{t-1}}(j|i) \tag{6}$$

In words, $a_{ij}[t]$ is the probability of transitioning to state j at t given that the chain is in state i at $t-1$. We assume that this probability doesn't depend on t and from now on write

$$a_{ij}[t] \;=\; a_{ij} \;\; \text{for all } t > 1 \tag{7}$$

The matrix $A = [a_{ij}]_{i,j=1}^{m}$ is called the **transition matrix**. It has the property that

$$\sum_{j=1}^{m} a_{ij} \;=\; 1 \tag{8}$$

A matrix with non-negative elements that has the property (8) is called a **stochastic** matrix.

Example 1

$$A \;=\; \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0.25 & 0.25 \\ 0.25 & 0.75 & 0 \end{bmatrix} \tag{9}$$

is a stochastic matrix. It represents the transition matrix for a Markov chain with 3 states.

A Markov chain is called **reducible** if there are 2 states $i, j \in S$ such that i cannot be reached from j in finite time. Intuitively, a reducible chain can be decomposed into two chains over disjoint subsets of S. A chain that is not reducible is *irreducible*. The chain in example 1 is irreducible.

Example 2 *A reducible chain:*

$$A \;=\; \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.75 & 0.25 \\ 0 & 0.5 & 0.5 \end{bmatrix} \tag{10}$$

In this chain state 1 cannot be reached from 2 or 3.

A Markov chain is called periodic if there exists at least one state i for which $P_{X_{t+t'}|X_t}(i|i) = 0$ for $t' > 0,\ t' \neq kt_0,\ k = 1, 2, \ldots,\ t_0 > 1$. The state i is called a **periodic state**. A chain that is not periodic is **aperiodic**. The chain in example 1 is aperiodic.

Example 3 *A periodic chain:*

$$A \;=\; \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \tag{11}$$

This chain cycles through the states in the order 1–2–3–1–. . . .

A chain that is aperiodic and irreducible is **ergodic**.

3 State distributions over time. The ergodic theorem

In this section we assume that a Markov chain is ergodic. We want to describe how the state distribution $p[t]$ evolves in time, given the transition matrix A and an initial distribution $p[0]$.

The first question is how does $p[t+1]$ depend on $p[t]$. We have

$$\begin{aligned} p_j[t+1] &= \sum_{i=1}^{m} P_{X_t X_{t+1}}(i,j) && (12) \\ &= \sum_{i=1}^{m} P_{X_t}(i) P_{X_{t+1}|X_t}(i|j) && (13) \\ &= \sum_{i=1}^{m} p_i[t] a_{ij} && (14) \end{aligned}$$

In compact form (assuming $p[t+1], p[t]$ are row vectors) the above becomes:

$$p[t+1] \;=\; p[t]A \tag{15}$$

Hence, advancing one step in the chain corresponds to multiplying the state distribution by the transition matrix A. You can verify that if $p[t]$ is a probability distribution, then $p[t+1]$ is also a probability distribution.

By induction

$$p[t+k] \;=\; p[t]A^k \tag{16}$$

and

$$p[t] \;=\; p[0]A^t \tag{17}$$

Thus, the distribution over the states at any time t can be described as a simple function of the initial distribution and the transition matrix! This suggests another question: as $t \to \infty$, should the state distribution still depend on the starting state? Common sense makes us expect that the chain "forgets" its origin as time goes by. This intuition is supported by mathematical proof if the chain is ergodic.

The ergodic theorem If a Markov chain defined by transition matrix A is ergodic, then

$$p[t] \longrightarrow p^{\infty} \tag{18}$$

for any initial distribution $p[0]$.

The following sequence of simple facts helps understand why this theorem is true. Most of them can be proved by verification.

1. p^{∞} is an eigenvector of A^T with eigenvalue 1.

$$p^{\infty} A = p^{\infty} \tag{19}$$

2. Denote by $\mathbf{1}$ the vector $[1\ 1\ 1\ \ldots\ 1]$. Any stochastic matrix has an eigenvalue equal to 1 with $\mathbf{1}$ as eigenvector.

$$\mathbf{1}A^T = \mathbf{1} \tag{20}$$

3. If λ is an eigenvalue for A then $|\lambda| \leq 1$. Hence $\lambda = 1$ is the largest eigenvalue of A.

4. If A diagonalizes as

$$A = XDX^{-1} \tag{21}$$

then

$$A^t = XD^tX^{-1} \tag{22}$$

Here D is the diagonal matrix of the eigenvalues, with 1 in the upper left corner. Note: this is a simplification, A does not always diagonalize.

5. Assume all $|\lambda|$'s but the first are smaller than 1, we have that, when $t \to \infty$

$$D^t \longrightarrow; D^{\infty} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ & & \ldots & \\ 0 & 0 & \ldots & 0 \end{bmatrix} \tag{23}$$

6. Hence A^t also converges to some matrix A^{∞}. This matrix has identical rows. Moreover, each row is equal to p^{∞}. Or rather, having arrived here we define p^{∞} to be a row of A.

$$A^t \longrightarrow A^{\infty} = \begin{bmatrix} p^{\infty} \\ p^{\infty} \\ \ldots \\ p^{\infty} \end{bmatrix} \tag{24}$$

Example 4 *The Markov chain defined by example 1 starts from state 1 (hence $p[0] = [1\ 0\ 0]$). The following table gives the succesive values for $p[t]$.*

0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1.0000	*0*	*0.3750*	*0.2812*	*0.2812*	*0.2988*	*0.2856*	*0.2922*	*0.2898*	*0.2904*	*0.2904*	*0.2903*
0	*0.5000*	*0.5000*	*0.4062*	*0.4766*	*0.4414*	*0.4546*	*0.4513*	*0.4513*	*0.4519*	*0.4514*	*0.4517*
0	*0.5000*	*0.1250*	*0.3125*	*0.2422*	*0.2598*	*0.2598*	*0.2565*	*0.2589*	*0.2577*	*0.2582*	*0.2581*

4 Maximum Likelihood parameter estimation

As usual, we set up the estimation problem as: given some data $\mathcal{D}$, estimate the parameters so as to maximize the likelihood of the data.

In this case, the data set consists of s sequences of observations

$$\mathcal{D} \;=\; \{\bar{x}^{(1)} = (x_0^{(1)}, x_1^{(1)}, \dots x_{T_1}^{(1)}),\ \bar{x}^{(2)} = (x_0^{(2)}, x_1^{(2)}, \dots x_{T_2}^{(2)}),\ \dots\dots \bar{x}^{(s)} = (x_0^{(s)}, x_1^{(s)}, \dots x_{T_s}^{(s)})\}$$

The sequences are independent of each other, but the states within each sequence are not. The parameters of the Markov chain are the initial probabilities $p_i[0],\ i = 1, \dots m$ and the transition matrix elements $a_{ij},\ i, j = 1, \dots\ m$.

To write the expression of the likelihood we need the probability of a sequence. This is

$$P(\bar{x}) \;=\; \underbrace{P(x_0)}_{p_{x_0}[0]} P(x_1|x_0)P(x_2|x_1)\dots \underbrace{P(x_t|x_{t-1})}_{a_{X_{t-1}X_t}} \dots P(x_T|x_{T-1} \tag{25}$$

$$\;=\; p_{x_0}[0] \prod_{t=1}^{T} a_{X_{t-1}X_t} \tag{26}$$

$$\;=\; p_{x_0}[0] \prod_{i,j=1}^{m} a_{ij}^{n_{ij}} \tag{27}$$

where n_{ij} is the number of times transition $i \to j$ occurs in the sequence.

For example, for the chain

$$(\bar{x}) = 1, 1, 3, 1, 2, 2, 2, 3, 1, 3 \tag{28}$$

we have $n_{11} = 1,\ n_{12} = 1,\ n_{13} = 2,\ n_{22} = 2,\ n_{23} = 1,\ n_{31} = 2$ and the other n_{ij}'s equal 0. Of course the sum

$$\sum_{i,j=1}^{m} n_{ij} \;=\; T. \tag{29}$$

With this, the likelihood of s independent sequences is:

$$L(p[0], A) \;=\; P(\mathcal{D}) \tag{30}$$

$$\;=\; \prod_{k=1}^{s} p_{x_0}^{(k)}[0] \prod_{i,j=1}^{m} a_{ij}^{n_{ij}^{(k)}} \tag{31}$$

$$\begin{aligned} &= \prod_{i=1}^{m} p_i^{n_i^0}[0] \prod_{i,j=1}^{m} a_{ij}^{\sum_k n_{ij}^{(k)}} && (32) \\ &= \prod_{i=1}^{m} p_i^{n_i^0}[0] \prod_{i,j=1}^{m} a_{ij}^{n_{ij}} && (33) \end{aligned}$$

In the above n_i^0 is the number of sequences that start in state i (with $\sum_i n_i^0 = s$) and $n_{ij} = \sum_k n_{ij}^{(k)}$ is the total number of $i \to j$ transitions occuring in all the sequences in the dataset. It is probably clear by now that $n_{ij},\ i,j = 1, \ldots m,\ n_i^0,\ i = 1, \ldots m$ are the sufficient statistics of the Markov chain model.

To find the solution to the estimation problem, note that $p[0]$ and $[\,a_{ij} = P(i \to j|i),\ j = 1, \ldots m\,]$ for each i are $m+1$ separate probability distribution over S. Hence their parameters can be estimated separately (i.e one can maximize L separately over each of these distribution's parameters) just like the parameters of any other discrete distribution. We get:

$$\begin{aligned} p_i[0] &= \frac{n_i^0}{s} && (34) \\ a_{ij} &= \frac{n_{ij}}{\underbrace{\sum_{j=1}^{m} n_{ij}}_{n_i}} && (35) \end{aligned}$$

Example 5 *For the dataset*

$$\mathcal{D} = \{\bar{x}^{(1)} = (1,1,3,1,2,2,2,3,1),\ \bar{x}^{(2)} = (2,2,2,3,1,1,3,1,2)\} \qquad (36)$$

we have $s = 2,\ n_{11} = 2,\ n_{12} = 2,\ n_{13} = 2,\ n_{21} = 0,\ n_{22} = 4,\ n_{23} = 2,\ n_{31} = 3,\ n_{32} = 0,\ n_{33} = 0$. *The total numbers of times in each state are* $n_1 = 6,\ n_2 = 6,\ n_3 = 4$*; note that here we have not counted the last states in each sequence, from which no transition occurs. Thus we obtain*

$$\begin{aligned} p[0] &= \begin{bmatrix} \frac{1}{2}\ \frac{1}{2}\ 0 \end{bmatrix} && (37) \\ A &= \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} \\ 1 & 0 & 0 \end{bmatrix} && (38) \end{aligned}$$