Lecture 5 – Feb 3 – Markov chains

$$X = [\texttt{H,H,H,L,L,M,M,M,L,L,L,L,L,L,L}]$$

where we are presuming that these refer to high, medium, and low positions in a tree.

Discrete time Markov chains

We have a state set $\mathcal{S} \in \{H, L, M\}$, and a sequence of events. In a k-th order Markov process, the probabilities depend on the the previous k steps in the chain. So in a first order Markov process, the current state (i) affects the next state $(i+1)$, but the next state is independent of the previous state $(i-1)$ *conditional* on state for time i.

So if x_t is the state at time t, then:

$$\mathbb{P}(x_{t+1} \mid x_t, x_{t-1}, \dots x_1) = \mathbb{P}(x_{t+1} \mid x_t)$$

So $\mathbb{P}(x_{t+1} \mid x_t)$ is the transition probability for the Markov chain (or the transition kernel). To describe a Markov transition probability, you need to describe the from state and the to state:

$$\begin{aligned} P &= \begin{bmatrix} p(H \to H) & p(H \to M) & p(H \to L) \\ p(M \to H) & p(M \to M) & p(M \to L) \\ p(L \to H) & p(L \to M) & p(L \to L) \end{bmatrix} \\ &= \begin{bmatrix} p_{HH} & p_{HM} & p_{HL} \\ p_{MH} & p_{MM} & p_{ML} \\ p_{LH} & p_{LM} & p_{LL} \end{bmatrix} \end{aligned}$$

So

$$\mathbb{P}(X) = \epsilon_H p_{HH} p_{HH} p_{HL} p_{LL} p_{LM} p_{MM} p_{MM} p_{ML} p_{LL}^6$$

Where we have a power of 6 at the end because the last 6 transitions in the data are L to L.

$$\begin{aligned} \mathbb{P}(X \mid \epsilon, p) &= \epsilon_{x_1} \prod_{t=2}^{n} \mathbb{P}(x_t \mid x_{t-1}) \\ &= \epsilon_{x_1} \prod_{t=2}^{n} p_{x_{t-1}, x_t} \\ \ln L(\epsilon, p) &= \ln[\epsilon_{x_1}] \sum_{t=2}^{n} \ln[p_{x_{t-1}, x_t}] \\ &= \ln[\epsilon_{x_1}] \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} n_{ij} \ln[p_{ij}] \end{aligned}$$

where n_{ij} is the count of the number of times in your data in which you observed an $i \to j$ transition.

We can take the derivative with respect to each parameter. but we find that we can maximize the likelihood by maximizing each of the ϵ and p parameters, but this violates our constraints that the rows of the transition matrix must be probabilities that sum to 1.

So we can reparameterize such that p_{i1} is not a parameter:

$$p_{i1} = 1 - \sum_{j=2}^{|\mathcal{S}|} p_{ij}$$

So for all $j \geq 2$:

$$\frac{\partial \ell}{\partial p_{ij}} = \frac{n_{ij}}{p_{ij}} - \frac{n_{i1}}{p_{i1}} \tag{1}$$

$$0 = \frac{n_{ij}}{\hat{p}_{ij}} - \frac{n_{i1}}{\hat{p}_{i1}} \tag{2}$$

$$\frac{n_{ij}}{\hat{p}_{ij}} = \frac{n_{i1}}{\hat{p}_{i1}} \tag{3}$$

$$\frac{n_{ij}}{n_{i1}} = \frac{\hat{p}_{ij}}{\hat{p}_{i1}} \tag{4}$$

A formal argument would deal with the situations in which n_{i1} is 0 (which poses problems when it is in the denominator), but that is a bit tedious. In the end it boils down eqn (4), which states that the way to maximize the likelihood is to set each p_{ij} according to the relative frequency of n_{ij} among other events that started in state i. So,

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}.$$

This makes sense. The MLE for the probability of being in state j in the next step given that you are currently in i is simply proportion of times that $i \rightarrow j$ occurred in your data set out of all of the transitions that started in state i.