Markov chains

A common assumption about probabilities for many sources of data is that the conditional dependence between random variables is limited to a small neighborhood. For example, in English text, the probability of the next letter (or next word) depends strongly on the previous few letters (or words) but only weakly on what occured 50 letters ago.

We say that a probability function $p(X_1, \ldots, X_n)$ is k^{th} order Markov for $k > 0$, if

$$p(X_{j+k}|X_1, X_2, \ldots, X_{j+k-1}) = p(X_{j+k}|X_j, X_{j+1}, \ldots, X_{j+k-1})$$

for any j such that $j + k \leq n$. That is, the probability of X_{j+k} conditioned on all previous elements in the sequence is identical to the probability of X_{j+k} conditioned on the previous k elements only. No matter how you specify the values of $X_1, \ldots, X_j$, you get the same result on the left side.

We can also talk about a "0^{th} order Markov model." In this case, we mean $k = 0$. Notice that the notation in the above equation makes no sense for $k = 0$. A zeroth order model just means that the variables X_i are independent.

0^{th} order: $p(X_j|X_1, X_2, \ldots, X_{j-1}) = p(X_j)$.

1^{st} order: $p(X_{j+1}|X_1, X_2, \ldots, X_j) = p(X_{j+1} \mid X_j)$.

2^{nd} order: $p(X_{j+2}|X_1, X_2, \ldots, X_j) = p(X_{j+2}|X_{j+1}, X_j)$.

k^{th} order: etc

Consider a 1^{st} order Markov model. Take a particular sequence $(i_1, i_2, \ldots, i_n)$. Let us rewrite the probabilities $p(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n)$ in terms of conditional probabilities.

$$\begin{aligned}
& p(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \\
= \; & p(X_n = i_n \mid X_1 = i_1,\ X_2 = i_2, \ldots, X_{n-1} - i_{n-1})\ p(X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}) \\
= \; & p(X_n = i_n \mid X_{n-1} = i_{n-1})\ p(X_{n-1} = i_{n-1} \mid X_1 = i_1,\ X_2 = i_2, \ldots, X_{n-2} = i_{n-2}) \\
& \qquad \cdot\ p(X_1 = i_1, \ldots, X_{n-2} = i_{n-2}) \\
= \; & \text{etc.} \\
= \; & p(X_n = i_n|X_{n-1} = i_{n-1})\ p(X_{n-1} = i_{n-1}|X_{n-2} = i_{n-2}) \ldots p(X_2 = i_2|X_1 = i_1)\ p(X_1 = i_1)
\end{aligned}$$

We can do the same with higher order models. For example, consider a $2nd$ order model.

$$\begin{aligned}
& p(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \\
= \; & p(X_n = i_n|X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}) p(X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}) \\
= \; & p(X_n = i_n|X_{n-1} = i_{n-1},\ X_{n-2} = i_{n-2})\ p(X_1 = i_1, X_2 = i_2, \ldots, X_{n-2} = i_{n-2}) \\
= \; & \text{etc.} \\
= \; & p(X_n = i_n|X_{n-1} = i_{n-1},\ X_{n-2} = i_{n-2})\ p(X_{n-1} = i_{n-1}|X_{n-2} = i_{n-2},\ X_{n-3} = i_{n-3})\ \cdot \\
& \quad \ldots,\ p(X_3 = i_3 \mid X_1 = i_1,\ X_2 = i_2)\ p(X_1 = i_1, X_2 = i_2)
\end{aligned}$$

The variables $X_1, X_2, \ldots, X_n$ are said to form a *Markov chain*. Markov chains gives us a way of calculating the probability of any sequence, assuming we have the conditional probability function.

Transition probability matrix

Consider a first order Markov model, with conditional probability function $p(X_{j+1}|X_j)$ on symbols $\{1, 2, \ldots, N\}$. Let this function be represented by the $N \times N$ matrix P, such that

$$P_{ik}^{(j)} = p(X_{j+1} = i | X_j = k)$$

where the row indicates the next symbol and the column indicates the previous symbol. The superscript j indicates that this function depends on j. (In the stationary case – see last lecture – we would drop this superscript.)

The matrix $P^{(j)}$ is sometimes called the *transition probability matrix* since it describes the probability that X_{j+1} is i given that X_j is k, that is, the probability of a "transition" from one symbol to another. (Such models are often used in probabilistic finite state machines, where you can look at the probability of going to the next state, given that you are a particular state.)

Here is an example transition matrix for $N = 2$,

$$P^{(j)} = \begin{bmatrix} \frac{31}{32} & \frac{1}{2} \\ \\ \frac{1}{32} & \frac{1}{2} \end{bmatrix}$$

and here is an example for $N = 3$,

$$P^{(j)} = \begin{bmatrix} \frac{2}{12} & \frac{3}{12} & \frac{4}{12} \\ \\ \frac{9}{12} & \frac{3}{12} & \frac{5}{12} \\ \\ \frac{1}{12} & \frac{6}{12} & \frac{3}{12} \end{bmatrix}$$

In the $N = 3$ example, $p(X_{j+1} = 2 \mid X_j = 1) = \frac{9}{12}$.

Notice in the above matrices how the probabilities within each column sum to 1, that is,

$$1 = \sum_{i=1}^{N} P_{ik}^{(j)} = \sum_{i=1}^{N} p(X_{j+1} = i \mid X_j = k) \ .$$

The reason is that, if $X_j = k$, the next symbol must be something, and so the sum of the probabilities of those things is 1. By contrast, the sum within a row need to be equal to 1. (It might be, or it might not be..)

Let's explore the relationship between condition, joint, and marginal probabilities and their role in Markov chains to make sure we really understand it. Again we consider 1^{st} order Markov models only. Take

$$p(X_{j+1} = i, \ X_j = k) \ = \ p(X_{j+1} = i | \ X_j = k) \ p(X_j = k) \tag{1}$$

and sum over k. The left side gives

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

and the right side gives

$$p(X_{j+1} = i) \;=\; \sum_{k=1}^{N} p(X_{j+1} = i|\; X_j = k)\; p(X_j = k)\;.$$

This can be represented as a matrix multiplication,

$$p(X_{j+1}) \quad = \quad p(X_{j+1}|\; X_j) \quad p(X_j) \tag{2}$$

where the function $p(X_{j+1}|\; X_j)$ is represented by an $N \times N$ matrix $P^{(j)}$, and the marginals $p(X_{j+1})$ and $p(X_j)$ are both $N \times 1$ vectors.

Now let's put $p(X_n)$ on the left side of Eq. (2). Expanding the right side recursively, we get

$$p(X_n) = p(X_n|X_{n-1})\; p(X_{n-1}|X_{n-2}) \ldots p(X_2|X_1)\; p(X_1)$$

where the condition probability functions each $N \times N$ matrices. Viewed in the opposite direction, if we are given the marginal probabilities $p(X_1)$ of the first element in the sequence, and we are given the conditional probabilities functions $p(X_{j+1}|\; X_j)$ for all j, then we can compute the marginal $p(X_n)$, or indeed any $p(X_j)$.

[ASIDE: Be careful to understand the notation here. This matrix multiplication above reminiscent of the chain of multiplications that we saw at the beginning of the lecture, i.e.

$$p(X_1 = i_1, \ldots, X_n = i_n) \;=\; p(X_n = i_n|X_{n-1} = i_{n-1}) \ldots p(X_2 = i_2|X_1 = i_1)\; p(X_1 = i_1).$$

But the meaning of the two statements is not the same. One represents a chain of vector and matrix multiplications on marginal and conditional probability *functions*. The other represents a chain of *scalar* multiplications on marginal and conditional probabilities.]

Stationary case

We have not yet mentioned the property of stationarity. Equation 2 does *not* assume that the sequence $X_1, X_2, \ldots, X_n$ is stationary – we could have a different matrix for different j's. However, if we *do* assume stationarity, then we get an interesting property, namely that

$$p(X_j) = p(X_{j+1})$$

for all j . For this property to hold, the $N \times 1$ vectors $p(X_j)$ must be the eigenvector of the conditional probability matrix P, with eigenvalue of 1.

Recall the two examples from earlier in the lecture. For the $N = 2$ example, we can compute that

$$p(X_j) = \begin{bmatrix} \frac{16}{17} \\ \\ \frac{1}{17} \end{bmatrix}$$

and for the $N = 3$ example, we get:

$$p(X_j) = \begin{bmatrix} \frac{51}{200} \\ \\ \frac{86}{200} \\ \\ \frac{63}{200} \end{bmatrix}$$

You can then compute the joint probability using Equation (1). If you do so, you will find that the matrix representing the joint probability functions is not symmetric. (They could be, but this would a special case.) The intuition of this is best seen with an example: in English text, the probability of pair `qu` is not the same as the probability of `uq`. (Similarly, the matrix represented the conditional probability function is not symmetric. In English, the probability of a `q` given that the previous letter was `u` is not the same as the probability of a `u` given that the previous letter was a `q`.)

Shannon's Example of English Text

[I did not discuss this in class. I include it here for your interest only. Do check it out!]

One well-known experiment using stationary Markov models was carried out by Claude Shannon in the early 1950's. Shannon considered an alphabet of 27 symbols, namely the 26 letters {`A,B,...,Z` } plus a space character. He used the probabilities of the letters (tabulated by others) to *generate* a sequence directly from a Markov model.

Here is an example generated from zeroth-order Markov model. i.e. The letters are chosen independently of each other, but with frequencies matching a large body of English text.

```
OCRO HLO RGWR NMIELWIS EU LL NBNESEBYATH EEI ALHENHTTPA OOBTTVA NAH BRL
```

Here is an example from a first-order model *i.e.* letters chosen from $p(X_n|X_{n-1})$:

```
ON IE ANTSOUTINYS ARE T INCTORE ST BE SDEAMY ACHIN D ILONASIVE TUCOOWE ATTEASONARE
FUSO TIZIN ANDY TOBE SEACE CTISBE IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID
PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE
```

and from a third-order model (starting to look more English-like...)

```
THE GENERATED JOB PROVIDUAL BETTER TRAND THE DISPLAYED CODE, ABOVERY UPONDULTS WELL
THE CODERST N THESTICAL IT DO HOCK BOTHE BERG. INSTATES CONS ERATIONS. NEVER ANY OF
PUBLIE AND TO THEORY. ENVTIAL CALLENGAND TO ELAST BENERATED IN WITH PIES AS IS WITH
THE
```

Shannon also carried out experiments where the symbols were words, rather than letters. (Probabilities were compiled from various texts.) As above, he generated a random source from these probability functions on words. Note that the alphabet here is much larger than 27 symbols. The number of words in English is in the tens of thousands.

0^{th}-order word model:

```
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE
A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE
```

1^{st} order word model:

```
THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS
POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD
THE PROBLEM FOR AN UNEXPECTED
```

This looks very much like English ! When you read it to yourself, it sounds very much like English. The reason it does is that our English language processing modules are "recognizing" the probabilistic similarity to real English.