STAT 534
Lecture 9
Markov Chains
April 30 2019

©2019 Marina Meilă
mmp@stat.washington.edu
Scribes:

1 Background

(Discrete time) **Markov chains**: a term used to represent a stochastic process

Stochastic process $X_{t t \in Z}$

- A collection of random variables t (t is an index) contained in some state space S
  - E.g. t = time
- $\mathbb{Z}$ is the set of positive integers

Going to assume the state space S is finite

- Is a set of states which X_t can take for $X_t \in S$

We're going to assume we can describe this situation by the Markov property

- Says for any $t \in Z$, we have that $X_t \perp\!\!\!\perp X_s | X_{t-1}$, where $s \in 1, , t - z$
- Equivalent to $X_t \perp\!\!\!\perp X_s | X_r$
  - Where $s < r < t$
  - If you give me any event in the past, then it doesn't matter what happened before that event
  - Can prove this from the first statement with induction
- Probability distribution at X_t is completely defined by previous state

Markov chain is one of the simplest types of graphical models

2 State space S

State space S, e.g. $S = \{1, 2, 3\}$

- These are the nodes/states

$p_i(t) = P(X_t = i)$

- At time t our chain is in state i
- If I'm in state t, assuming I started at t=0, weve gone through t transitions to get here
- Even if transition probabilities are well-defined, might not be easy to think about where I am after t transitions

Full probability vector $= p(t) = (p_i(t), ..., p_m(t))$

- m is the cardinality of the state space S; $m = |S|$

$$\sum_{i=1}^{m} p_i(t) = 1$$

- = probability X_t is in the state space

$p(0) =$ initial probability

- Our model involves looking at what happened before, and coming up with a probability distribution based on that
- But nothing happens before the initial point, so this gets a special name
- To estimate this, need several chains, assume independence of the chains, etc.

Transition probabilities $= a_{ij}(t) = P(X_t = j | X_{t-1} = i)$

- How would I write $P(X_1, X_2, X_3)$ as a conditional (in general, not for Markov chains)
  - $= P(X_2|X_0, X_1)P(X_1|X_0)P(X_0)$

Can define entire probability distribution based on 1 step transitions b/c this is a Markov chain

3 Transition probability matrix (TPM)

Usually denoted as A

$[a_{ij}]_{i,j=1}^{m}$

- TPM= matrix of all of these a_{ij}s, from 1 to m

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

- The sum over all js of $a_{ij} = 1$
- This is how you know its the rows that represent probability distributions
  - Each row sums to 1 if we sum over the columns
  - Rows = i, columns = j

$a_{ij} \geq 0$

- Any matrix that has the above form and property and the additional constraint that all $a_{ij} \geq 0$ is called a stochastic matrix
  - b/c it defines a set of probability distributions
- No negatives: cant have negative probabilities

Rows = initial state, columns = state you wind up in

4 (Ir)reducibility

Reducible if $\exists\ i, j \in SxS\ s.t. P(X_t = j|X_s = i) = 0 \forall s < t, i \neq j$

- Means if I'm ever in state i, at any time, I can't get to state j, no matter what
  - b/c this is for all s <t, this isn't for t-1
  - means theres no path in the graph from i to j

When a chain is reducible, we have two disjoint subsets of the state space

- if we start in one set we stay in that set forever

5 Periodicity

A Markov chain is periodic if $\exists\ i \in S\ s.t. P(X_{t+t} = i|X_t = i)$ for $t > 0, t \neq kt_0, k = 1, 2, ,$ and $t_0 > 1$

- t_0 and k are each any positive integer > 1, t is any integer > 0
  - $t \neq kt_0$ means holds for all t' that aren't multiples of t_0
- starting at state i, probability of being back in state i t units later
- in a general situation maybe I can get back to state I in 3 steps, or 4 steps, or whatever
- but maybe I have a weird chain w/ periodic behavior, where I cant get back to I in an odd number of steps, ever; that's a periodic chain
- the transition probalbity matrix A defines a periodic chain:
  - $A = \{010\}\{001\}\{100\}$
  - Is just a loop; if I start at state 1, I can only get back there in 3 steps

Aperiodic chain is not periodic

A chain that is aperiodic and irreducible and homogeneous is ergodic

- We like Markov chains to be Eergodic
- Homogeneous chains:
  - Notation were using is $a_{ij}(t)$, but we can drop the t b/c these transition probabilities dont depend on t; transition matrix A isnt indexed by t

6 Ergodicity

$p(t+1)$

- Probability distribution of all of the states at time t+1
- Lets consider only p(t)
- What is $p(t+1)$ in terms of $p(t)$?
  - We know state distribution at p(t), using induction we figure out p(t+1), and do that for all t, using the Markov property

How do I build up to $p_j(t+1)$

- Probability Im in state j at time t+1
- $= P(X_{t+1} = j)$
- $= \sum_{i=1}^{m} P(X_i)a_{ij} = \sum_{i=1}^{m} p_i(t)a_{ij}$
  - $i \neq j$
  - Is saying I could be in any of the m states at time t
  - For any of those m states $p_i(t)$ is probability Im in that state
  - (a_{ij}) = probability I go from that state to j
    * If I sum over all those possibilities I get the probability I'm in state j at time t
  - Probability of going from i to j, times probability of starting in i
  - Summed over all m states that you could go to j from

Can write this in vector notation, assuming these are row vectors

- $p_i(t+1) = p(t)A$
  - if $p(t)$ is a probability distribution so is $p(t+1)$ given the TPM
  - A= TPM

What does this imply?

- Lets say I know p_0, what can I do now? I can keep multiplying by the matrix A over and over again
  - Notation $= p(0)A^{t+1}$
- For any generic time t, the probability vector at time t = initial probability * transition probability matrix raised to t
  - $p(t) = p(0)A^t$

Next step is to think about what happens when t is really big? Is there some kind of limit? Does this converge to something?

- p(0) matters a lot at the beginning, but after many cycles it shouldn't matter much
- What happens when we take t to infinity?
- This is where the ergodic property comes in, in the form of the Ergodic theorem

6.1 Ergodic theorem

For any ergodic Markov chain defined by transition probability matrix A, we have that p(t) goes to some limit well call p^{∞} as t goes to ∞ This is built up

through a series of facts

1. $P^{\infty}A = p^{\infty}$
   - p^{∞} is an eigenvector of A
2. For any stochastic matrix A, a vector of all 1s as a column vector, $\mathbf{1} = (1,,1)^T$
   - We have A$\mathbf{1}$ = $\mathbf{1}$
3. Any eigenvalue has magnitude ≤ 1 of this stochastic matrix A
   - $|\lambda| \leq 1$
4. If we can write A as XDX^{-1}
   - Where D is a diagonal matrix of eigenvalues
   - Then $A^t = XD^tX^{-1}$
     - Power of a matrix by t; $A^2 = AA$
     - Lowercase t isnt transpose!
   - If we can describe the limit of D^t as t goes to infity, then we describe the limit A^t
     - And D is a diagonal matrix so possibly easier to work with
5. D^t goes to a limit D^{∞}
   - where D^{∞} is a matrix with 1 at the top left and 0 everywhere else
   - if max of all of the eigenvalues $(\lambda_i) = 1$ and this is unique
     - An assumption
   - $A \in \mathbb{R}^{mxm}$
     - A is an m x m matrix, so is X and D
   - This tells us $A^t \to A^{\infty}$
     - Where LHS is a matrix with all of its rows defined as p^{∞}
       * P^{∞} is a row vector, so this is still a matrix

And we call p^{∞} the stationary distribution of this Markov chain

- We get here no matter where we start from; if we run the chain long enough we converge to this distribution