Lecture 9: "Markov Chains"

Sotiris Nikoletseas
Professor

Computer Engineering and Informatics Department
2017 - 2018

Markov Chains - Stochastic Processes

- Stochastic Process: A set of random variables $\{X_t, t \in T\}$ defined on a set D, where:
  - T: a set of indices representing time
  - X_t: the state of the process at time t
  - D: the set of states

- The process is discrete/continuous when D is discrete/continuous. It is a discrete/continuous time process depending on whether T is discrete or continuous.

- In other words, a stochastic process abstracts a random phenomenon (or experiment) evolving with time, such as:
  - the number of certain events that have occurred (discrete)
  - the temperature in some place (continuous)

Markov Chains - transition matrix

- Let S a <u>state space</u> (finite or countable). A Markov Chain (MC) is at any given time at one of the states. Say it is currently at state i; with probability P_{ij} <u>it moves</u> to the state j. So:

$$0 \leq P_{ij} \leq 1 \text{ and}$$
$$\sum_j P_{ij} = 1$$

  The matrix $P = \{P_{ij}\}_{ij}$ is the <u>transition probabilities matrix</u>.

- The MC starts at an <u>initial state</u> X_0, and at each point in time it moves to a new state (including the current one) according to the transition matrix P. The resulting sequence of states $\{X_t\}$ is called <u>the history</u> of the MC.

The memorylessness property

- Clearly, the MC is a stochastic process, i.e. a random process in time.
- the defining property of a MC is its memorylessness, i.e. the random process "forgets" its past (or "history"), while its "future" (next state) only depends on the "present" (its current state). Formally:

  $\Pr\{X_{t+1} = j | X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i\} = \Pr\{X_{t+1} = j | X_t = i\} = P_{ij}$

  The initial state of the MC can be arbitrary.

t-step transitions

- For states $i, j \in S$, the t-step transition probability from i to j is:

$$P_{ij}^{(t)} = \Pr\{X_t = j | X_0 = i\}$$

  i.e. we compute the (i, j)-entry of the t-th power of transition matrix P.

- Chapman - Kolmogorov equations:

$$P_{ij}^{(t)} = \sum_{i_1, i_2, \ldots, i_{t-1} \in S} \Pr\{X_t = j, \bigcap_{k=1}^{t-1} X_k = i_k | X_0 = i\}$$

$$= \sum_{i_1, i_2, \ldots, i_{t-1} \in S} P_{ii_1} P_{i_1 i_2} \cdots P_{i_{t-1} j}$$

First visits

- The probability of <u>first visit</u> at state j after t steps, starting from state i, is:

$$r_{ij}^{(t)} = \Pr\{X_t = j, X_1 \neq j, X_2 \neq j, \ldots, X_{t-1} \neq j | X_0 = i\}$$

- The <u>expected number of steps</u> to arrive for the first time at state j starting from i is:

$$h_{ij} = \sum_{t>0} t \cdot r_{ij}^{(t)}$$

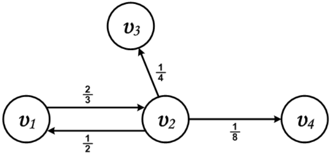
Visits/State categories

- The probability of a visit (not necessarily for the first time) at state j, starting from state i, is:
$$f_{ij} = \sum_{t>0} r_{ij}^{(t)}$$
- Clearly, if $f_{ij} < 1$ then there is a positive probability that the MC never arrives at state j, so in this case $h_{ij} = \infty$.
- A state i for which $f_{ii} < 1$ (i.e. the chain has positive probability of never visiting state i again) is a transient state. If $f_{ii} = 1$ then the state is persistent (also called recurrent).
- If state i is persistent but $h_{ii} = \infty$ it is null persistent. If it is persistent and $h_{ii} \neq \infty$ it is non null persistent.

Note. In finite Markov Chains, there are no null persistent states.

Example (I)

- A Markov Chain



- The transition matrix P:

$$P = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{1}{2} & \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- The probability of starting from v_1, moving to v_2, staying there for 1 time step and then moving back to v_1 is:
$\Pr\{X_3 = v_1, X_2 = v_2, X_1 = v_2 | X_0 = v_1\} =$
$= P_{v_1 v_2} P_{v_2 v_2} P_{v_2 v_1} = \frac{2}{3} \cdot \frac{1}{8} \cdot \frac{1}{2} = \frac{1}{24}$

Example (II)

- The probability of moving from v_1 to v_1 in 2 steps is:
  $P^{(2)}_{v_1v_1} = P_{v_1v_1} \cdot P_{v_1v_1} + P_{v_1v_2} \cdot P_{v_2v_1} = \frac{1}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot \frac{1}{2} = \frac{4}{9}$
  Alternatively, we calculate P^2 and get the (1,1) entry.

- The first visit probability from v_1 to v_2 in 2 steps is:
  $$r^{(2)}_{v_1v_2} = P_{v_1v_1} P_{v_1v_2} = \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9}$$
  while $r^{(7)}_{v_1v_2} = (P_{v_1v_1})^6 P_{v_1v_2} = \left(\frac{1}{3}\right)^6 \cdot \frac{2}{3} = \frac{2}{3^7}$
  and $\quad r^{(t)}_{v_2v_1} = (P_{v_2v_2})^{t-1} P_{v_2v_1} = \left(\frac{1}{8}\right)^{t-1} \cdot \frac{1}{2} = \frac{1}{2^{3t-2}}$
  for $t \geq 1$ (since $r^{(0)}_{v_2v_1} = 0$)

Example (III)

- The probability of (eventually) visiting state v_1 starting from v_2 is:

$$f_{v_2 v_1} = \sum_{t \geq 1} \frac{1}{2^{3t-2}} = \frac{4}{7}$$

- The expected number of steps to move from v_1 to v_2 is:

$$h_{v_1 v_2} = \sum_{t \geq 1} t \cdot r_{v_1 v_2}^{(t)} = \sum_{t \geq 1} t \cdot (P_{v_1 v_1})^{(t-1)} P_{v_1 v_2} = \frac{3}{2}$$

(actually, we have the mean of a geometric distribution with parameter $\frac{2}{3}$)

Irreducibility

- Note: A MC can naturally be represented via a directed, weighted graph whose vertices correspond to states and the transition probability P_{ij} is the weight assigned to the edge (i, j). We include only edges (i, j) with $P_{ij} > 0$.
- A state u is reachable from a state v (we write $v \to u$) iff there is a path $\mathcal{P}$ of states from v to u with $\Pr\{\mathcal{P}\} > 0$.
- A state u communicates with state v iff $u \to v$ and $v \to u$ (we write $u \leftrightarrow v$)
- A MC is called irreducible iff every state can be reached from any other state (equivalently, the directed graph of the MC is strongly connected).

Irreducibility (II)

- In our example, v_1 can be reached only from v_2 (and the directed graph is not strongly connected) so the MC is not irreducible.

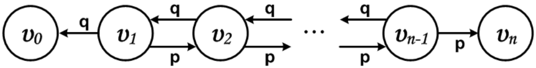
- Note: In a finite MC, either all states are transient or all states are (non null) persistent.

- Note: In a finite MC which is irreducible, all states are persistent.

Absorbing states

- Another type of states: A state i is <u>absorbing</u> iff $P_{ii} = 1$ (e.g. in our example, the states v_3 and v_4 are absorbing)

- Another example:



The states v_0, v_n are absorbing

State probability vector

- Definition. Let $q^{(t)} = (q_1^{(t)}, q_2^{(t)}, ..., q_n^{(t)})$ be the row vector whose i-th component $q_i^{(t)}$ is the probability that the MC is in state i at time t. We call this vector the state probability vector (alternatively, we call it the distribution of the MC at time t).

- Main property. Clearly
$$q^{(t)} = q^{(t-1)} \cdot P = q^{(0)} \cdot P^t$$
where P is the transition probability matrix

- Importance: rather than focusing on the probabilities of transitions between the states, this vector focuses on the probability of being in a state.

Periodicity

- Definition. A state i called periodic iff the largest integer T satisfying the property

$$q_i^{(t)} > 0 \Rightarrow t \in \{a + kT | k \geq 0\}$$

  is largest than 1 ($a > 0$ a positive integer); otherwise it is called aperiodic. We call T the periodicity of the state.

- In other words, the MC visits a periodic state only at times which are terms of an arithmetic progress of rate T.

Periodicity (II)

- Example: a random walk on a bipartite graph clearly represents a MC with all states having periodicity 2. Actually, a random walk on a graph is aperiodic iff the graph is not bipartite.
- Definition: We call aperiodic a MC whose states are all aperiodic. Equivalently, the chain is aperiodic iff (gcd: greatest common divisor):

$$\forall x, y : \ \gcd\{t : P_{xy}^{(t)} > 0\} = 1$$

Ergodicity

- Note: the existence of periodic states introduces significant complications since the MC "oscillates" and does not "converge". The state of the chain at any time clearly depends on the initial state; it belongs to the same "part" of the graph at even times and the other part at odd times.

- Similar complications arise from null persistent states.

- Definition. A state which is non null persistent and aperiodic is called <u>ergodic</u>. A MC whose states are all ergodic is called ergodic.

- Note: As we have seen, a finite, irreducible MC has only non-null persistent states.

Stationarity

- Definition: A state probability vector (or distribution) π for which
$$\pi^{(t)} = \pi^{(t)} \cdot P$$
is called stationary distribution

- Clearly, for the stationary distribution we have
$$\pi^{(t)} = \pi^{(t+1)}$$
In other words, when a chain arrives at a stationary distribution it "stays" at that distribution for ever, so this the "final" behaviour of the chain (i.e. the probability of being at any vertex tends to a well-defined limit, independent of the initial vertex). This is why we also call it equilibrium distribution or steady state distribution. We also say that the chain converges to stationarity.

The Fundamental Theorem of Markov Chains

- In general, a stationary distribution may not exist so we focus on Markov Chains with stationarity.

- Theorem. For every irreducible, finite, aperiodic MC it is:
  1. The MC is ergodic.
  2. There is a unique stationary distribution π, with $\pi_i > 0$ for all states $i \in S$
  3. For all states $i \in S$, it is $f_{ii} = 1$ and $h_{ii} = \frac{1}{\pi_i}$
  4. Let $N(i,t)$ the number of times the MC visits state i in t steps. Then

$$\lim_{t \leftarrow \infty} \frac{N(i,t)}{t} = \pi_i$$

  Namely, independently of the starting distribution, the MC converges to the stationary distribution.

Stationarity in doubly stochastic matrices

- <u>Definition:</u> A nxn matrix M is <u>stochastic</u> if all its entries are non-negative and for each row i, it is:
$$\sum_j M_{ij} = 1$$
(i.e. the entries of any row add to 1). If in addition the entries of any column add to 1, i.e. for all j it is:
$$\sum_i M_{ij} = 1$$
then the matrix is called <u>doubly stochastic</u>.
- <u>Lemma:</u> The stationary distribution of a Markov Chain whose transition probability matrix P is doubly stochastic is the uniform distribution.
<u>Proof:</u> The distribution $\pi_v = \frac{1}{n}$ for all v is stationary, since it satisfies:
$$[\pi \cdot P]_v = \sum_u \pi_u P_{uv} = \sum_u \frac{1}{n} P_{uv} = \frac{1}{n} \sum_u P_{uv} = \frac{1}{n} 1 = \pi_v \quad \square$$

Stationarity in symmetric chains

- Definition: A chain is called symmetric iff:

$$\forall u, v : \; P_{uv} = P_{vu}$$

- Lemma: If a chain is symmetric its stationary distribution is uniform.

- Proof: Let N be the number of states. From Fundamental Theorem, it suffices to check that $\pi_u = \frac{1}{N}$, $\forall u$, satisfies $\pi \cdot P = \pi$. Indeed:

$$(\pi P)_u = \sum_v \pi_v \cdot P_{vu} = \frac{1}{N} \sum_v P_{uv} = \frac{1}{N} \cdot 1 = \pi_u \qquad \square$$

Examples - Card shuffling

- Given a set of n cards, let a Markov Chain whose states are all possible permutations of the cards ($n!$) and one step transition between states defined by some card shuffling rule. For the shuffling to be effective the stationarity distribution must be the uniform one. We provide two such effective shufflings:

  (1) Random transpositions: "choose" any two cards at random and swap them e.g.

  $$\cdots\ a \cdots\ b \cdots \Rightarrow \cdots\ b \cdots\ a \cdots$$

  Note: Indeed the transition probabilities in both directions are the same $\left(\text{each one is } \frac{1}{\binom{n}{2}}\right)$ so the chain is symmetric and its stationary distribution uniform.

Examples - Card shuffling (II)

(2) Top-in-at-Random: "place the top card to a random new position of the n possible ones"

Note: There are n potential new states. Also, each state can be reached from n other states with probability $\frac{1}{n}$ from each. So the chain is doubly stochastic and its stationary distribution uniform.

On the mixing time

- Although the Fundamental Theorem guarantees that an aperiodic, irreducible finite chain converges to a stationary distribution, it does not tell us <u>how fast convergence happens</u>.

- The convergence rate appropriately close to stationarity is captured by an important measure (the <u>"mixing time"</u>).

On the mixing time (II)

- As an example, the number of shufflings needed by "Top-in-at-Random" to produce an almost uniform permutation of cards is $O(n \log n)$. Other methods are faster e.g. their mixing time is $O(\log n)$, such as in Riffle-Shuffle where the deck of cards is randomly split into two sets (left, right) which are then "interleaved".

- This convergence rate is very important in algorithmic applications, where we want to ensure that a proper sample can be obtained in fairly small time, even when the state space is very large!

Random walks on graphs

- Let $G = (V, E)$ a connected, non-bipartite, undirected graph with n vertices. We define a Markov Chain MC_G corresponding to a random walk on the vertices of G, with transition probability:

$$P_{uv} = \begin{cases} \frac{1}{d(u)}, & \text{if } uv \in E(G) \\ 0, & \text{otherwise} \end{cases}$$

  where $d(u)$ is the degree of vertex u.

- Since the graph is connected and undirected, MC_G is clearly irreducible. Also, since the graph is non-bipartite, MC_G is aperiodic.

The stationary distribution

- So (from fundamental theorem of Markov Chains) M_G has a unique stationary distribution π.

  Lemma 1: For all vertices $v \in V$ it is $\pi_v = \frac{d(v)}{2m}$, where m is the number of edges of G.

  Proof: From the definition of stationarity, it must be:

$$\pi_v = [\pi \cdot P]_v = \sum_u \pi_u P_{uv} \quad , \forall v \in V$$

  Because of uniqueness, it suffices to verify the claimed solution. Indeed, for all $v \in V$ we have (for the claimed solution value):

$$\sum_u \pi_u P_{uv} = \sum_{u:uv \in E} \frac{d(u)}{2m} \frac{1}{d(u)} = \frac{1}{2m} \sum_{u:uv \in E} 1 = \frac{1}{2m} d(v) = \pi_v$$

$\square$