Markov Chains and Link Analysis

Shan-Hung Wu
shwu@cs.nthu.edu.tw

Department of Computer Science,
National Tsing Hua University, Taiwan

NetDB-ML, Spring 2015

Outline

Outline

1. Markov Chains
   - Random Processes
   - Markov Processes and Markov Chains
   - Ergodic Chains
   - Reversible Chains
2. Link Analysis
   - Web as A Graph
   - The PageRank
   - Personalized PageRank

Random Processes (1/2)

Definition (Random Process)

Given a probability space $(\Omega, \mathcal{F}, P)$, a ***random process*** (or ***stochastic process***), denoted by $\{X^{(t)} : t \in \mathcal{T}\}$, is a collection of random variables defined over $(\Omega, \mathcal{F}, P)$ and indexed by elements t of a set $\mathcal{T}$.

- Typically, we think $t = 1, 2, \cdots$ as ***time***
  - t could also be space or position on a DNA string, etc.
- The values of $X^{(t)}$ are called the ***states***
- $X^{(t)}$ is a function of both the outcome ω and time t
  - Fixing the outcome $\omega \in \Omega$, $X^{(t)}$ is a deterministic function of t
  - Fixing t, $X^{(t)}$ is a random variable
- The distribution of $X^{(t)}$, called ***state distribution***, changes with t

Random Processes (2/2)

- A random process is said to be ***discrete*** or ***continuous in time*** depending on whether $\mathcal{T}$ is discrete (finite or infinitely countable) or continuous
- A random process is ***discrete*** or ***continuous in state*** depending on whether X is discrete or continuous
- Random processes vs graphical models (e.g., Bayesian networks and Markov random fields)?

Random Processes (2/2)

- A random process is said to be ***discrete*** or ***continuous in time*** depending on whether $\mathcal{T}$ is discrete (finite or infinitely countable) or continuous
- A random process is ***discrete*** or ***continuous in state*** depending on whether X is discrete or continuous
- Random processes vs graphical models (e.g., Bayesian networks and Markov random fields)?
    - A Bayesian network degenerates into a discrete-state-and-time random process when its random variables are indexed by time

Why Random Process?

- So far, we assume that instances in a dataset are i.i.d.
  - This simplifies the calculation of the likelihood $P[\mathcal{X}|\theta]$
- However, in practice instances may come in order and successive instances may be dependent
  - E.g.,. letters in a word, phonemes in speech utterance, page visits in the Web, etc.
- The sequence can be characterized as being generated by a parametric random process
  - We want to estimate the parameter of a random process and then make predictions

Outline

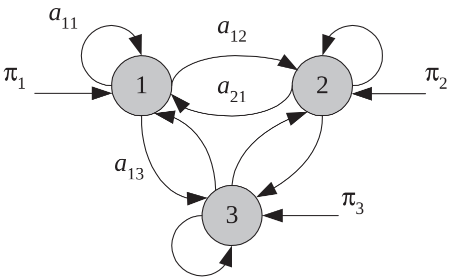
1. Markov Chains
   - Random Processes
   - Markov Processes and Markov Chains
   - Ergodic Chains
   - Reversible Chains
2. Link Analysis
   - Web as A Graph
   - The PageRank
   - Personalized PageRank

Markov Processes

- A random process is called the ***Markov process*** if it satisfies the ***Markov property***:
  $P\left(X^{(t_0+t_1)} \leqslant x | X^{(t_0)} = x_0, X^{(t)} = x_t, -\infty < t < t_0\right) =$
  $P\left(X^{(t_0+t_1)} \leqslant x | X^{(t_0)} = x_0\right)$
  - We can think t_0 and t_1 be the present and future time respectively
  - This is a mathematical version of the saying "today is the first day of the rest of your life"

Markov Chains (1/2)

- A Markov process which is discrete in both time and state is called the first-order ***Markov Chain***
- Let $S_1, \cdots, S_K$ be the K distinct states and $t = 1, 2, \cdots$, the word "first-order" means:
  $P\left(X_{t+1} = S_j | X_t = S_i, X_{t-1} = S_k, \cdots\right) = P\left(X_{t+1} = S_j | X_t = S_i\right)$
- Denote the unconditional ***state probability*** at time t by a row vector $\boldsymbol{\pi}^{(t)} = [\pi_0^{(t)}, \cdots, \pi_K^{(t)}]$, where $\pi_i^{(t)} = P\left(X^{(t)} = S_i\right)$ and $\sum_{i=1}^{K} \pi_i^{(t)} = 1$



Markov Chains (2/2)

- We say a Markov chain is ***time homogeneous*** iff $P\left(X^{(t+1)}=S_j|X^{(t)}=S_i\right)=P\left(X^{(t+m+1)}=S_j|X^{(t+m)}=S_i\right)$
  - That is, the ***transition probability*** $a_{i,j}=P\left(X^{(t+1)}=S_j|X^{(t)}=S_i\right)$ from state i to j does not change with time t
  - $\sum_{j=1}^{N}a_{i,j}=1$ for any i
- A time homogeneous Markov chain has a fixed ***transition matrix*** $\boldsymbol{A}$, whose each row sum to 1

State Probabilities

- Given an initial state probability $\boldsymbol{\pi}^{(1)}$, we have $\boldsymbol{\pi}^{(2)} = \boldsymbol{\pi}^{(1)}\boldsymbol{A}$, $\boldsymbol{\pi}^{(3)} = \boldsymbol{\pi}^{(2)}\boldsymbol{A}$, and so on...
- $\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^{(t-1)}\boldsymbol{A} = \boldsymbol{\pi}^{(1)}\boldsymbol{A}^{t-1}$ for $t \geqslant 2$
  - $\pi_i^{(t)}$ is the sum of transition probabilities multiplied along all paths of length $t-1$ from initial states j to i and weighted by the initial state probabilities $\pi_j^{(1)}$
- In practice, we are often interested in the ***steady state probability*** $\boldsymbol{\pi} = \lim_{t\to\infty}\boldsymbol{\pi}^{(t)}$
  - $\boldsymbol{\pi}$ indicates how much chance we see a particular state in the long run
- However, the steady state probability may not exist

Outline

1. Markov Chains
   - Random Processes
   - Markov Processes and Markov Chains
   - Ergodic Chains
   - Reversible Chains
2. Link Analysis
   - Web as A Graph
   - The PageRank
   - Personalized PageRank

Ergodic Markov Chains (1/2)

Definition (Ergodic Markov Chain)

A Markov chain is said to be ***ergodic*** if there exists a positive integer T such that for all pairs of states i, j in the Markov chain, if it is started at time 0 in state i then for all $t > T$, the probability of being in state j at time t is greater than 0.

- An ergodic Markov chain cannot have
  - Absorbing state: once entering this state, there is no way to get out
  - Transient state: starting from this state, there is a positive probability that the chain will never return
  - Periodic state: the number of steps required to return to this state are always multiples of some integer larger than 1
- An ergodic chain constitutes of a single set of positive recurrent states
  - Positive recurrent state: starting from this state, the probability of eventual return is 1 and the expected number of steps required is finite (a positive integer)

Ergodic Markov Chains (2/2)

- Most real world chains are ergodic if we can define proper transition probability between any two states

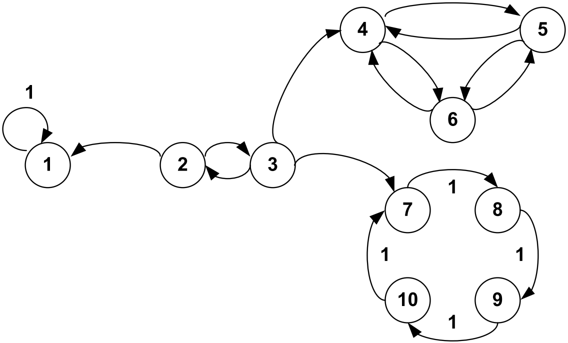


Figure : A non-ergodic Markov chain. Transition probabilities equal to 1 are marked explicitly. States 1 is an absorbing state; states 2 and 3 are transient states; states 7, 8, 9, and 10 are periodic states; states 4, 5, and 6 are positive recurrent states.

Steady State Probability and Random Walk

Theorem

For any ergodic Markov chain, there is a unique steady state probability vector $\boldsymbol{\pi}$ *such that if* $\eta(i,t)$ *is the number of visits to state* i *in* t *steps, then* $\lim_{t\to\infty}\frac{\eta(i,t)}{t} = \pi_i$.

- If a Markov chain is ergodic, the steady state $\boldsymbol{\pi}$ exists, and its value is independent with the initial state
- But how to find $\boldsymbol{\pi}$?

Steady State Probability and Random Walk

Theorem

For any ergodic Markov chain, there is a unique steady state probability vector $\boldsymbol{\pi}$ *such that if* $\eta(i,t)$ *is the number of visits to state* i *in* t *steps, then* $\lim_{t\to\infty}\frac{\eta(i,t)}{t} = \pi_i$.

- If a Markov chain is ergodic, the steady state $\boldsymbol{\pi}$ exists, and its value is independent with the initial state
- But how to find $\boldsymbol{\pi}$?
  - Since $\boldsymbol{\pi}\boldsymbol{A} = \boldsymbol{\pi}$, $\boldsymbol{\pi}$ is the (left) eigenvector of $\boldsymbol{A}$ corresponding to the eigenvalue 1
  - Or we can multiply $\boldsymbol{\pi}^{(1)}$ enough times by $\boldsymbol{A}$ to get an approximation
- The steady state probability $\boldsymbol{\pi}$ also implies the (probability of the) final stop of a ***random walk***
  - Insensitive to the initial state where the walk begins

Outline

Reversible Markov Chains (1)

Definition (Reversible Markov Chain)

A Markov chain is said to be ***reversible*** if there exists a probability distribution over states, π, such that the ***detailed balanced equation***

$$\pi_i P\left(X^{(t+1)} = S_j | X^{(t)} = S_i\right) = \pi_j P\left(X^{(t+1)} = S_i | X^{(t)} = S_j\right)$$

holds for all times t and states i and j.

- If the Markov chain has finite states and is time-homogeneous, we have $\pi_i A_{i,j} = \pi_j A_{j,i}$

Reversible Markov Chains (2)

- Summing the detailed balanced equations over all states gives $\sum_i \pi_i A_{i,j} = \sum_i \pi_j A_{j,i} = \pi_j \sum_i A_{j,i} = \pi_j$
  - Valid for all j, which implies $\boldsymbol{\pi A} = \boldsymbol{\pi}$
- So, for reversible Markov chains, ***$\boldsymbol{\pi}$ is always the steady state probability vector***
- If, the chain begins with the steady state distribution (i.e., $\pi_i^{(1)} = P(X^{(1)} = S_i) = \pi_i$), then we have $\pi_i^{(t)} = P(X^{(t)} = S_i) = \pi_i$ for all t, and the detailed balanced equation be written as $P\left(X^{(t+1)} = S_j, X^{(t)} = S_i\right) = P\left(X^{(t+1)} = S_i, X^{(t)} = S_j\right)$

Outline

Web as A Graph

- Nodes (vertexes): web pages
- Links: hyperlinks (e.g., `<a href="http://www.ibm.com">A computer service company</a>`)
  - Directed
  - Annotated by anchor text
- How can this structure help in, for example, searching a web page?

Web as A Graph

- Nodes (vertexes): web pages
- Links: hyperlinks (e.g., `<a href=“http://www.ibm.com”>A computer service company</a>`)
  - Directed
  - Annotated by anchor text
- How can this structure help in, for example, searching a web page?
  - When scoring a page for a term query, a search engine can additionally calculate the (cosine) similarity between the query terms and the the tokens in the anchor text pointing to that page
- Idea: descriptions to a website from others are usually more precise than those from site holders themselves
  - Evidence: despite that everyone knows IBM is a computer company, we seldom see the word “computer” in its homepage

Outline

1. Markov Chains
   - Random Processes
   - Markov Processes and Markov Chains
   - Ergodic Chains
   - Reversible Chains
2. Link Analysis
   - Web as A Graph
   - The PageRank
   - Personalized PageRank

The PageRank Algorithm (1/2)

- In addition to the anchor text, the structure of the graph also helps in scoring a page
- PageRank is a score derived from the structure of the graph
- The idea is to model the graph as a Markov chain
  - Pages as the states; hyperlinks as the transitions
- In the simplest form, the ***random walk model*** can be used to calculate the transition matrix $\boldsymbol{A}$
  - If a page p_i has two links to p_j and p_k, then $a_{i,j} = a_{i,k} = 1/2$
- The state probability $\pi_i^{(t)}$ denotes the probability that a random walking user stay at the page i at time t
- If the chain is ergodic, then the PageRank of page i can be simply the steady state probability π_i denoting the "popularity"

The PageRank Algorithm (2/2)

- What if a page has no outward link?
  - It becomes an absorbing state and prevents the chain from being ergodic
- Solution?

The PageRank Algorithm (2/2)

- What if a page has no outward link?
  - It becomes an absorbing state and prevents the chain from being ergodic
- Solution? Observe that users may teleport between pages (e.g., enter URLs directly in the browser)
- We can associate each $a_{i,j}$ with a teleport probability:
  - $a_{i,j} = 1/K$ if page i has no outward link;
  - $a_{i,j} = (1-\alpha)/o_i + \alpha/K$ otherwise, where o_i is number of outward links from page i and α is a user-specified parameter

Remarks

- PageRank is one of the commonly used scoring functions in major search engines today
  - Often contributes to a portion of the final score of a page
- Why is PageRank so popular?

Remarks

- PageRank is one of the commonly used scoring functions in major search engines today
  - Often contributes to a portion of the final score of a page
- Why is PageRank so popular?
- It is language independent
  - Exploits the link structure and provides an alternative view to the content-based scores (e.g., cosine similarity)
  - Agnostic to queries (i.e., the PageRank of page remains the same for all queries)
- It can be computed in a scalable manner
  - We can multiply $\boldsymbol{\pi}^{(1)}$ enough times by $\boldsymbol{A}$ to get an approximation of the steady state $\boldsymbol{\pi}$
  - Each multiplication task can be divided into parts and done in parallel (e.g., using the MapReduce framework)
- It is spam-resistant
  - You need to fake "inward links" from others' pages to increase the visibility of your site

Outline

Personalized PageRank (1/2)

- The PageRank is static to different users
- Can we personalize PageRank based a particular user's interests?

Personalized PageRank (1/2)

- The PageRank is static to different users
- Can we personalize PageRank based a particular user's interests?
- Recall the random walk model assumes that all links of a page are clicked with the equal probability
- We can instead estimate the $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{A}$ by tracing the sequence of page visits from a user

Personalized PageRank (2/2)

- Given N sample sequences of length T, where $X^{(n,t)}$ is the state at time t in sequence n
- If we assume that $\pi_i^{(1)}$ and $a_{i,j}$ are with Binomial distributions (or Multinomial distributions with constrains), we can easily obtain their ML estimators:
  - $\widehat{\pi}_i^{(1)} = \frac{\sum_{n=1}^{N} \zeta(X^{(n,1)} = s_i)}{N}$, where $\zeta(x)$ equals 1 if x is true and 0 otherwise
  - $\widehat{a}_{i,j} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T-1} \zeta(X^{(n,t)} = s_i \text{ and } X^{(n,t+1)} = s_j)}{\sum_{n=1}^{N} \sum_{t=1}^{T-1} \zeta(X^{(n,t)} = s_i)}$ [Proof]
- To ensure an ergodic chain, we still assume the teleport probabilities
  - We can further estimate the teleport probabilities of a particular user

What's Next?

- Markov chains are the most simple type of Markov processes
  - Lay out the fundamentals for other processes commonly used in machine learning:

	States are fully observable	**States are partially observable**
Transition is autonomous	Markov chains	Hidden Markov models
Transition is controlled	Markov decision processes	Partially observable Markov decision processes