Introduction to Operating Systems

Lecture 3: Advanced Probabilistic Topics

MING GAO

SE@ecnu
(for course related communications)
mgao@sei.ecnu.edu.cn

Sep. 30, 2016

Outline

1. Markov chain and random walk
2. Graphical models
   - Directed Model
   - Undirected Model
3. Tail Bounds

Markov chain

A stochastic processes $\{X_t | t \in T\}$ is a collection of random variables. The index t is often called time, as the process represents the value of a random variable changing over time. Let Ω be the set of values assumed by the random variables X_t. We call each element of Ω a state, as X_t represents the state of the process at time t.

Definition of Markov property

A process $X_0, X_1, \cdots$ satisfies the Markov property if

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, \cdots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

for all n and all $x_i \in \Omega$.

Definition of Markov chain

A stochastic process $X_0, X_1, \cdots$ of discrete time and discrete space is a Markov chain

A random walk on a graph can be modeled as a Markov chain

Transition matrix

Definition

Let a Markov chain have $P_{x,y}^{(t+1)} = P[X_{t+1} = y | X_t = x]$, and the finite state space be $\Omega = [n]$. This gives us a transition matrix $P^{(t+1)}$ at time t. The transition matrix is an $N \times N$ matrix of nonnegative entries such that the sum over each row of $P^{(t)}$ is 1, since $\forall n$ and $\forall x_i \in \Omega$

$$\sum_y P_{x,y}^{(t+1)} = \sum_y P[X_{t+1} = y | X_t = x] = 1$$

- For example, $P^{(t+1)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}$
- $P_{1,2}^{(t+1)} = P[X_{t+1} = 2 | X_t = 1] = \frac{1}{2}$ and $P_{1,3}^{(t+1)} = P[X_{t+1} = 3 | X_t = 1] = 0$
- $\sum_{i=1}^{4} P_{1,i}^{(t+1)} = 1$

State distribution

Definition

Let $\pi^{(t)}$ be the state distribution of the chain at time t, that $\pi_x^{(t)} = P[X_t = x]$.

For a finite chain, $\pi_x^{(t)}$ is a vector of N nonnegative entries such that $\sum_x \pi_x^{(t)} = 1$. Then, it holds that $\pi^{(t+1)} = \pi^{(t)} P^{(t+1)}$. We apply the law of total probability

$$\pi_y^{(t+1)} = P[X_{t+1} = y] = \sum_x P[X_{t+1} = y | X_t = x] P[X_t = x] = \sum_x \pi_x^{(t)} P_{x,y}^{(t+1)}$$

- Let $\pi_x^{(t)} = (0.4, 0.6, 0, 0)$ be a state distribution, then $\pi_x^{(t+1)} = (0.4, 0.6, 0, 0)$
- Let $\pi_x^{(t)} = (0, 0, 0.5, 0.5)$ be a state distribution, then $\pi_x^{(t+1)} = (0, 0, 0.5, 0.5)$
- Let $\pi_x^{(t)} = (0.1, 0.9, 0, 0)$ be a state distribution, then $\pi_x^{(t+1)} = (0.35, 0.65, 0, 0)$

Stationary distributions

Definition

A stationary distribution of a finite Markov chain with transition matrix P is a probability distribution π such that

$$\pi P = \pi$$

- For some Markov chains, no matter what the initial distribution is, after running the chain for a while, the distribution of the chain approaches the stationary distribution
- E.g., $P^{20} = \begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}$. The chain could converge to any distribution which is a linear combination of $(0.4, 0.6, 0, 0)$ and $(0, 0, 0.5, 0.5)$. We observe that the original chain P can be broken into two disjoint Markov chains, which have their own stationary distributions. We say that the chain is **reducible**

Irreducibility

Definition

State y is accessible from state x if it is possible for the chain to visit state y if the chain starts in state x, in other words, $P^n(x, y) > 0, \forall n$. State x **communicates with** state y if y is accessible from x and x is accessible from y. We say that the Markov chain is **irreducible** if all pairs of states communicates.

- y is accessible from x means that y is connected from x in the transition graph, i.e., there is a directed path from x to y
- x communicates with y means that x and y are strongly connected in the transition graph
- A finite Markov chain is irreducible if and only if its transition graph is strongly connected
- The Markov chain associated with transition matrix P is not irreducible

Irreducibility

Definition

State y is accessible from state x if it is possible for the chain to visit state y if the chain starts in state x, in other words, $P^n(x,y) > 0, \forall n$. State x **communicates with** state y if y is accessible from x and x is accessible from y. We say that the Markov chain is **irreducible** if all pairs of states communicates.

- y is accessible from x means that y is connected from x in the transition graph, i.e., there is a directed path from x to y
- x communicates with y means that x and y are strongly connected in the transition graph
- A finite Markov chain is irreducible if and only if its transition graph is strongly connected
- The Markov chain associated with transition matrix P is not irreducible

Aperiodicity

Definition

The period of a state x is the greatest common divisor (gcd), such that $d_x = gcd\{n|(P^n)_{x,x} > 0\}$. A state is aperiodic if its period is 1. A Markov chain is aperiodic if all its states are aperiodic.

- For example, suppose that the period of state x is $d_x = 3$. Then, starting from state x, chain $x, \bigcirc, \bigcirc, \square, \bigcirc, \bigcirc, \square, \bigcirc, \bigcirc, \square, \cdots$, only the squares are possible to be x.
- In the transition graph of a finite Markov chain, $(P^n)_{x,x} > 0$ is equivalent to that x is on a cycle of length n. Period of a state x is the greatest common devisor of the lengths of cycles passing x.

Theorem

1. If the states x and y communicate, then $d_x = d_y$.
2. We have $(P^n)_{x,x} = 0$ if $n \bmod(d_x) \neq 0$

Convergence of Markov chain

Fundamental theorem of Markov chain

Let $X_0, X_1, \cdots$, be an irreducible aperiodic Markov chain with finite state space Ω, transition matrix P, and arbitrary initial distribution $\pi^{(0)}$. Then, there exists a unique stationary distribution π such that $\pi P = \pi$, and $\lim_{t\to\infty} \pi^{(0)} P^t = \pi$.

- Existence: there exists a stationary distribution
- Uniqueness: the stationary distribution is unique
- Convergence: starting from any initial distribution, the chain converges to the stationary distribution
- In fact, any finite Markov chain has a stationary distribution. Irreducibility and aperiodicity guarantee the uniqueness and convergence behavior of the stationary distribution

Google's PageRank

Problem definition

Given n interlinked webpages, rank them in order of "importance" in terms of importance scores $x_1, x_2, \cdots, x_n \geq 0$

- Key insight: use the existing link structure of the web to determine importance. A link to a page is like a vote for its importance
  - Given a web with n pages, construct $n \times n$ matrix A as: $a_{ij} = \frac{1}{n_j}$ if page j links to page i, 0 otherwise
  - Sum of $j-$th column is 1, so A is a Markov matrix.
  - The ranking vector $\overrightarrow{x}$ solves $A\overrightarrow{x} = \overrightarrow{x}$
- Possible issues?
  - Replace A with $B = 0.85A + 0.15(\text{matrix with every entry } \frac{1}{n})$, where B is also a Markov chain
  - A pages rank is the probability the random user will end up on that page, OR, equivalently

The curse of dimensionality

- Modern machine learning is usually concerned with high-dimensional objects
- Consider learning a distribution over $x \in \{0, 1\}^N$
- If $N = 100$, $p(x)$ has 1267650600228229401496703205375 free parameters

Why do we need graphical models?

- Graphs are an intuitive way of representing and visualising the relationships between many variables. (Examples: family trees, electric circuit diagrams, neural networks)
- A graph allows us to abstract out the conditional independence relationships between the variables from the details of their parametric forms. Thus we can answer questions like: "Is A dependent on B given that we know the value of C?" just by looking at the graph
- Graphical models allow us to define general message-passing algorithms that implement probabilistic inference efficiently. Thus we can answer queries like "What is $p(A|C = c)$?" without enumerating all settings of all variables in the model

Graphical models = statistics $\times$ graph theory $\times$ computer science

Conditional independence

- The special structure graphical models assume is conditional independence
- If you would like to guess the value of some variable x_i, then once you know the values of some "neighboring" variables $x_{\mathcal{N}(i)}$, then you get no additional benefit from knowing all other variables
- Turns out, this leads to factorized distributions

Conditional independence

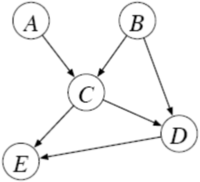
- X is independent of Y if "knowing Y doesn't help you to guess X"

$$X \perp Y \leftrightarrow P(X, Y) = P(X)P(Y)$$

- X is independent of Y given Z if "once you know Z, knowing Y doesn't help you to guess X"

$$X \perp Y|Z \leftrightarrow P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Representing knowledge through graphical models



A graphical model is a probability distribution written in a factorized form. For example

$$p(x) \propto \psi(x_1, x_3)\psi(x_2, x_3)\psi(x_3, x_4)$$

Graph

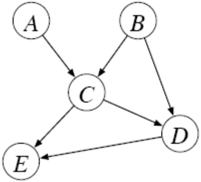
The two most common forms of graphical model are *directed graphical models* and *undirected graphical models*, based on directed acylic graphs and undirected graphs, respectively.

Let $G = (V, E)$ be a graph, where V and E represent the sets of vertices and edges, respectively

- Vertices correspond to random variables
- Edges represent statistical dependencies between the variables

Outline

Directed acyclic graphical models



Bayesian network

A DAG Model or Bayesian network corresponds to a factorization of the joint probability distribution

$$P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B, C)P(E|C, D)$$

In general $P(X_1, X_2, \cdots, X_n) = \Pi_{i=1}^{n} P(X_i | X_{pa(i)})$, where $pa(i)$ denotes the parents of vertex i.

How to do learning

Maximum likelihood

Given a fixed graph, how to do learning?

- Natural criterion

$$\arg\max_{\theta} L(\theta), L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log P(x_d|\theta)$$

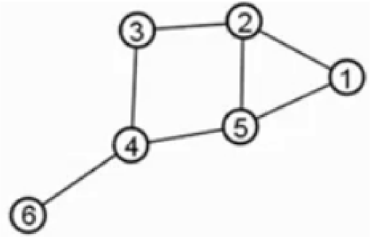
- Solution is empirical conditionals

$$P(X_i = x_i | X_{\pi(i)} = x_{\pi(i)}, \theta) = \frac{\#[X_i = x_i, X_{\pi(i)} = x_{\pi(i)}]}{\#[X_{\pi(i)} = x_{\pi(i)}]}$$

Outline

1. Markov chain and random walk
2. Graphical models
   - Directed Model
   - Undirected Model
3. Tail Bounds

Undirected graphs



Which is true

a. $x_1 \perp x_3 | x_2$
b. $x_1 \perp x_3 | x_{2,4}$
c. $x_1 \perp x_3 | x_{2,5}$
d. $x_1 \perp x_6 | x_{2,3,4,5}$
e. $x_1 \perp x_6 | x_{2,4}$
f. $x_1 \perp x_6 | x_2$
g. $x_1 \perp x_6 | x_4$
h. $x_{1,6} \perp x_{3,5} | x_4$
i. $x_{1,6} \perp x_{3,5} | x_{2,4}$

Undirected graphs Cont.

Equivalently (when $P(x) > 0$), a graph asserts that $P(x_i|x_{-i}) = P(x_i|x_{N(i)})$, but what's the formula for $P(x)$?

Hammersley-Clifford theorem

- A positive distribution $P(x) > 0$ obeys the conditional independencies of a graph G when $P(x)$can be represented as

$$P(x) = \frac{1}{Z}\Pi_{c\in\mathcal{C}}\psi_c(x_c)$$

  where $\mathcal{C}$ is the set of all cliques, and $Z = \sum_x \Pi_{c\in\mathcal{C}}\psi_c(x_c)$ is the "partition function"

- This is not obvious and no direct probabilistic interpretation for ϕ
- It is easy to show that $P(x) = \frac{1}{Z}\Pi_{c\in\mathcal{C}}\psi_c(x_c)$ obeys this conditional independence assumptions of a graph

Exponential family

An exponential family is a set of distributions

$$\begin{aligned} p(x;\theta) &= \frac{1}{Z(\theta)} Exp(\theta^T \phi(x)) \\ &= Exp(\theta^T \phi(x) - A(\theta)) \end{aligned}$$

parameterized by $\theta \in \Theta \subset \mathbb{R}^d$, $Z(\theta) = \sum_x Exp(\theta^T \phi(x))$ and $A(\theta) = \log Z(\theta)$ is the "log-partition function". We care because: (1) Many interesting properties; (2) Undirected models are an exponential family

Examples

Examples for exponential family

- Bernoulli distribution: r.v. $X \sim p^x(1-p)^{1-x}$, where $x \in \{0, 1\}$. We have $\theta = \log \frac{p}{1-p}, \phi(x) = x, A(\theta) = \frac{1}{1-p}$
- Gaussian distribution: r.v. $X \sim N(\mu, \sigma^2)$, in terms of canonical form of exponential family, we have

$$S(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \theta = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix}, z(\theta) = \frac{\mu^2}{2\sigma^2} + \log \sigma\sqrt{2\pi} \quad (1)$$

- Bernoulli, Gaussian, Binomial, Poisson, Exponential, Weibull, Laplace, Gamma, Beta, Multinomial, Wishart distributions are all exponential families
- We will discuss how to learn parameters for undirected models

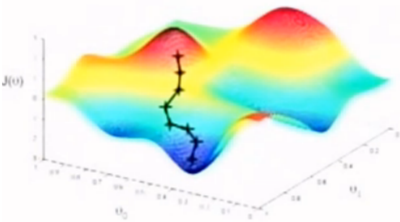
Maximum likelihood learning

MLE

Given $x_1, x_2, \cdots, x_D$, we want to solve

$$\arg\max_{\theta} L(\theta), L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log P(x_d|\theta)$$

- Simple approach: gradient descent, repeatedly set $\theta_i := \theta_i + \lambda \frac{\partial}{\partial \theta_i} L(\theta)$
- $\frac{\partial}{\partial \theta_i} A(\theta) = E(\phi(x)(i))$
- Notice that $\sum_{d=1}^{D} \phi(x_d) = E_\theta(\phi(x))$



Example of four vertex undirected model

Factorization

Assume x is binary, $P(x) = \frac{1}{Z}\psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{34}(x_3, x_4)$

Rewite

Equivalent to $p(x;\theta) = \frac{1}{Z(\theta)}Exp(\theta^T\phi(x))$ with

- $\phi(x) = [\mathbb{I}_{x_1=0,x_2=0}, \mathbb{I}_{x_1=0,x_2=1}, \cdots, \mathbb{I}_{x_3=1,x_4=1}]$
- $\theta = [\theta(x_1 = 0, x_2 = 0), \theta(x_1 = 0, x_2 = 1), \cdots, \theta(x_3 = 1, x_4 = 1)]$
- $\frac{\partial A(\theta)}{\partial\theta} = [P(x_1 = 0, x_2 = 0), P(x_1 = 0, x_2 = 1), \cdots, P(x_3 = 1, x_4 = 1)]$

Undirected models

Exponential family

- Typically written as $P(x) = \frac{1}{Z}\Pi_{c \in \mathcal{C}} \phi_c(x_c)$
- Rewrite as

$$p(x;\theta) = \exp(\theta^T \phi(x) - A(\theta)) \tag{2}$$

$$\phi(x) = \{\mathbb{I}_{x_c = x_c^*} | c \in \mathcal{C}, \text{all possible } x_c^*\} \tag{3}$$

- An undirected model is an E.F. where $\phi(x)$ has indicator functions for every configuration of every clique
- Recall also that at the maximum likelihood solution, $\sum_{i=1}^{D} \phi(x_d) = E^{\theta}(\phi(X))$

Comparisons of directed and undirected models

Summary

Directed and undirected models stem from similar conditional independence assumptions

	directed	undirected
Assumption	$P(X_i \mid X_{i-1}, \cdots, X_1) = P(X_i \mid X_{pa(i)})$	$P(X_i \mid X_{-i}) = P(X_i \mid X_{N(i)})$
Likelihood	$P(x) = \Pi_i P(X_i \mid X_{pa(i)})$	$P(x) = \frac{1}{Z} \Pi_{c \in \mathcal{C}} \phi_c(x_c)$
Learning	$P(x_i \mid x_{pa(i)}; \theta) = \widehat{P}(x_i \mid x_{pa(i)})$	$P(x_c; \theta) = \widehat{P}(x_c)$

Tail bounds

Question

Consider the experiment of tossing a fair coin n times. What is the probability that the number of heads exceeds $\frac{3n}{4}$.

Note

The tail bounds of a r.v. X are concerned with the probability that it deviates significantly from its expected value $E(X)$ on a run of the experiment

Markov inequality

Markov inequality

If X is any r.v. and $0 < a < +\infty$, then

$$P(X > a) \le \frac{E(X)}{a} \text{ or } P(X > aE(X)) \le \frac{1}{a}$$

Proof

$$P(X > a) = \int_{X>a} dx \le \int \frac{X}{a} dx = \frac{E(X)}{a} \tag{4}$$

Example

$$P(X > \frac{3n}{4}) \le \frac{n/2}{3n/4} = \frac{2}{3} \tag{5}$$

Chebyshevs inequality

Chebyshevs inequality

If r.v. X has mean and variance $\mu = E(X)$ and $\sigma^2 = E[(X-\mu)^2]$, then

$$P(|X-\mu| > a) \le \frac{\sigma^2}{a^2} \text{ or } P(|X-\mu| > aE(X)) \le \frac{\sigma^2}{a^2 E(X)^2}$$

Proof

Let $Y = |X-\mu|^2$ in Markov's inequality, then

$$P(|X-\mu| > a) = P(Y > a^2) \le \frac{E(Y)}{a^2} = \frac{\sigma^2}{a^2} \tag{6}$$

For Example,

$$P(X > \frac{3n}{4}) < P(|X - \frac{n}{2}| > \frac{n}{4}) \le \frac{Var(X)}{(\frac{n}{4})^2} = \frac{4}{n}$$

Chernoff bound

Deriving Chernoff bound

Let X_i be a sequence of independent r.v.s with $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$. r.v. $X = \sum_{i=1}^{n} X_i$, then $P(X_i = k) = \binom{n}{k} p_i^k (1 - p_i)^{n-k}$.

- $P(X < (1-\delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$, where $\mu = \sum_{i=1}^{n} p_i$
- $P(X < (1-\delta)\mu) < \exp\left(-\mu\delta^2/2\right)$

Proof

- For $t > 0$, $P(X < (1-\delta)\mu) = P\big(\exp(-tX) > \exp(-t(1-\delta)\mu)\big) < \frac{\Pi_{i=1}^{n} E(\exp(-tX_i))}{\exp(-t(1-\delta)\mu)}$
- $E(\exp(-tX_i)) = p_i e^{-t} + (1 - p_i) = 1 - p_i(1 - e^{-t}) < \exp(p_i(e^{-t} - 1))$ $(1 - x < e^{-x})$
- $\Pi_{i=1}^{n} E(\exp(-tX_i)) < \Pi_{i=1}^{n} \exp(p_i(e^{-t} - 1)) = \exp(\mu(e^{-t} - 1))$

Proof of Chernoff bound Cont.

Proof Cont.

- $P(X < (1-\delta)\mu) < \frac{\exp(\mu(e^{-t}-1))}{\exp(-t(1-\delta)\mu)} = \exp(\mu(e^{(-t)} + t - t\delta - 1))$
- Now its time to choose t to make the bound as tight as possible. Taking the derivative of $\mu(e^{(-t)} + t - t\delta - 1)$ and setting $-e^{(-t)} + 1 - \delta = 0$. We have $t = \ln(1/1-\delta)$
- $P(X < (1-\delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$

Proof of second statement

To get the simpler form of the bound, we need to get rid of the clumsy term $(1-\delta)^{(1-\delta)}$

- $(1-\delta)\ln(1-\delta) = (1-\delta)(\sum_{i=1} -\frac{\delta^i}{i}) > -\delta + \frac{\delta^2}{2}$
- $(1-\delta)^{(1-\delta)} > \exp(-\delta + \frac{\delta^2}{2})$
- $P(X < (1-\delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} < \left(\frac{e^{-\delta}}{\exp(-\delta+\frac{\delta^2}{2})}\right)^{\mu} = \exp(-\mu\delta^2/2)$

Chernoff bound (Upper tail)

Theorem

Let X_i be a sequence of independent r.v.s with $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$. r.v. $X = \sum_{i=1}^{n} X_i$ and $\mu = \sum_{i=1}^{n} p_i$, then $P(X_i = k) = \binom{n}{k} p_i^k (1 - p_i)^{n-k}$.

- $P(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$
- $P(X > (1+\delta)\mu) < \exp\left(-\mu\delta^2/4\right)$

Example

Let X be the number of heads in n tosses of a fair coin, then $\mu = \frac{n}{2}$ and $\delta = \frac{1}{2}$, we have

$$P(X > \frac{3n}{4}) = P(X > (1 + \frac{1}{2})\frac{n}{2}) < \exp\left(-\frac{n}{2}\delta^2/4\right) = \exp\left(-8n\right)$$

If we toss the coin 100 times, the probability is less than $\exp -800$

Hoeffding inequality

Theorem

Let $X_1, X_2, \cdots, X_n$ be i.i.d. observations such that $E(X_i) = \mu$ and $a \leq X_i \leq b$. Then, for any $\epsilon > 0$,

$$P(|\overline{X} - \mu| > \epsilon) < 2\exp\left(-2n\epsilon^2/(b-a)^2\right)$$

Example

If $X_1, X_2, \cdots, X_n \sim Bernoulli(p)$

- In terms of Hoeffding inequality, we have

$$P(|\overline{X} - p| > \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$$

- If $p = 0.5$,

$$P(\overline{X} - 0.5 > \frac{1}{4}) < P(|\overline{X} - 0.5| > \frac{1}{4}) \leq 2\exp\left(-32n\right)$$

Take-home messages

- Markov chain
- Graphical model
    - Directed model
    - Undirected model
- Tail bounds