Algorithm Foundations of Data Science

Lecture 1: Markov Chain

MING GAO

DaSE@ECNU
(for course related communications)
mgao@dase.ecnu.edu.cn

Mar. 7, 2018

Outline

1. Markov Chain and Random Walk
   - Reminder on Conditional Probability
   - Markov Chain
2. Graphical models
   - Directed Model
   - Undirected Model

Outline

1. Markov Chain and Random Walk
   - Reminder on Conditional Probability
   - Markov Chain
2. Graphical models
   - Directed Model
   - Undirected Model

Conditional probability

Let E, F, and C be events,

- $P(E|F) = \frac{P(E\cap F)}{P(F)}$ (well defined only if $P(F) > 0$)
- $$\begin{aligned} P(E\cap F|C) &= \frac{P(E\cap F\cap C)}{P(C)} \\ &= \frac{P(E\cap F\cap C)}{P(F\cap C)}\frac{P(F\cap C)}{P(C)} = P(E|F\cap C)\cdot P(F|C). \end{aligned}$$

Conditional probability

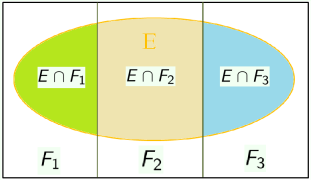
Let E, F, and C be events,

- $P(E|F) = \frac{P(E\cap F)}{P(F)}$ (well defined only if $P(F) > 0$)
-

$$\begin{aligned} P(E\cap F|C) &= \frac{P(E\cap F\cap C)}{P(C)} \\ &= \frac{P(E\cap F\cap C)}{P(F\cap C)}\frac{P(F\cap C)}{P(C)} = P(E|F\cap C)\cdot P(F|C). \end{aligned}$$

Let X be a discrete r.v.,

- $\sum_k P(X = x_k|F) = 1$;
- $P(E) = \sum P(E|X = x_k)P(X = x_k)$;
- $\sum P(X = x_k, F|E) = P(F|E)$;
- But, $\sum P(E|X = x_k) \neq 1$.



Random process

A random process is a collection of r.v.s indexed by some set I, taking values in a set S.

- I is the index set, usually time, e.g., $\mathcal{Z}^+$, $\mathcal{R}$, and $\mathcal{R}^+$, etc.
- S is the state space, e.g., $\mathcal{Z}^+$, $\mathcal{R}^n$, and $\{a, b, c\}$, etc.

Random process

A random process is a collection of r.v.s indexed by some set I, taking values in a set S.

- I is the index set, usually time, e.g., $\mathcal{Z}^+$, $\mathcal{R}$, and $\mathcal{R}^+$, etc.
- S is the state space, e.g., $\mathcal{Z}^+$, $\mathcal{R}^n$, and $\{a, b, c\}$, etc.

We classify random processes according to

- Index set can be discrete or continuous;
- State space can be finite, countable or uncountable (continuous).

Outline

Markov property

More formally, $X(t)$ is Markovian if the following property holds:

$$\begin{aligned} P(X(t_n) = j_n | X(t_{n-1}) = j_{n-1}, \cdots, X(t_1) = j_1) \\ = P(X(t_n) = j_n | X(t_{n-1}) = j_{n-1}) \end{aligned}$$

for all finite sequence of times $t_1 < \cdots < t_n \in I$ and of states $j_1, \cdots, j_n \in S$.

Markov property

More formally, $X(t)$ is Markovian if the following property holds:

$$\begin{aligned} P(X(t_n) = j_n | X(t_{n-1}) = j_{n-1}, \cdots, X(t_1) = j_1) \\ = P(X(t_n) = j_n | X(t_{n-1}) = j_{n-1}) \end{aligned}$$

for all finite sequence of times $t_1 < \cdots < t_n \in I$ and of states $j_1, \cdots, j_n \in S$.

- A random process is called a **Markov Process** if conditional on the current state, its future is independent of its past.

Markov property

More formally, $X(t)$ is Markovian if the following property holds:

$$P(X(t_n) = j_n | X(t_{n-1}) = j_{n-1}, \cdots, X(t_1) = j_1)$$
$$= P(X(t_n) = j_n | X(t_{n-1}) = j_{n-1})$$

for all finite sequence of times $t_1 < \cdots < t_n \in I$ and of states $j_1, \cdots, j_n \in S$.

- A random process is called a **Markov Process** if conditional on the current state, its future is independent of its past.
- 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 $x_i \in \Omega$.

Markov property

More formally, $X(t)$ is Markovian if the following property holds:

$$\begin{aligned} P(X(t_n) = j_n | X(t_{n-1}) = j_{n-1}, \cdots, X(t_1) = j_1) \\ = P(X(t_n) = j_n | X(t_{n-1}) = j_{n-1}) \end{aligned}$$

for all finite sequence of times $t_1 < \cdots < t_n \in I$ and of states $j_1, \cdots, j_n \in S$.

- A random process is called a **Markov Process** if conditional on the current state, its future is independent of its past.
- 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 $x_i \in \Omega$.
- The term Markov property refers to the memoryless property of a stochastic process.

Discrete time Markov Chain

A stochastic process $X_0, X_1, \cdots$ of discrete time and discrete space is a Markov chain if it satisfies the Markov property, i.e.,

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

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

Discrete time Markov Chain

A stochastic process $X_0, X_1, \cdots$ of discrete time and discrete space is a Markov chain if it satisfies the Markov property, i.e.,

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

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

- A Markov chain $\{X_t\}$ is said to be time homogeneous if $P(X_{s+t} = j | X_s = i)$ is independent of s. When this holds, putting $s = 0$ gives

$$P(X_{s+t} = j | X_s = i) = P(X_t = j | X_0 = i).$$

Discrete time Markov Chain

A stochastic process $X_0, X_1, \cdots$ of discrete time and discrete space is a Markov chain if it satisfies the Markov property, i.e.,

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

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

- A Markov chain $\{X_t\}$ is said to be time homogeneous if $P(X_{s+t} = j|X_s = i)$ is independent of s. When this holds, putting $s = 0$ gives
$$P(X_{s+t} = j|X_s = i) = P(X_t = j|X_0 = i).$$
- If moreover $P(X_{n+1} = j|X_n = i) = P_{ij}$ is independent of n, then X is said to be time homogeneous Markov chain.

Examples

- $\Omega = \{A, B\}$; $\pi(A, B) = q$, $\pi(A, A) = 1 - q$, $\pi(B, A) = r$, $\pi(B, B) = q$ for some $0 < q, r < 1$ and $\mu(A) = 1, \mu(B) = 0$, so that at time 0 we always start at A.

Examples

- $\Omega = \{A, B\}$; $\pi(A, B) = q$, $\pi(A, A) = 1 - q$, $\pi(B, A) = r$, $\pi(B, B) = q$ for some $0 < q, r < 1$ and $\mu(A) = 1, \mu(B) = 0$, so that at time 0 we always start at A.
- $\Omega = \mathbb{Z}$; $\pi(a, a-1) = \frac{1}{2}$, $\pi(a, a+1) = \frac{1}{2}$, $\pi(a, b) = 0$ if $b \neq a \pm 1$ for every $a \in \mathbb{Z}$, $\mu(0) = 1$ and $\mu(a) = 0$ if $a \neq 0$, so at time 0 we always start at 0.

Examples

- $\Omega = \{A, B\}$; $\pi(A, B) = q$, $\pi(A, A) = 1 - q$, $\pi(B, A) = r$, $\pi(B, B) = q$ for some $0 < q, r < 1$ and $\mu(A) = 1, \mu(B) = 0$, so that at time 0 we always start at A.
- $\Omega = \mathbb{Z}$; $\pi(a, a-1) = \frac{1}{2}$, $\pi(a, a+1) = \frac{1}{2}$, $\pi(a, b) = 0$ if $b \neq a \pm 1$ for every $a \in \mathbb{Z}$, $\mu(0) = 1$ and $\mu(a) = 0$ if $a \neq 0$, so at time 0 we always start at 0.
- A graph $G = (V, E)$ consists of a vertex set V and an edge set E, where the elements of E are unordered pairs of vertices: $E \subset \{(x, y) : x, y \in V, x \neq y\}$. The degree $deg(x)$ of a vertex x is the number of neighbours of x. In this case, $\Omega = V$ and is finite, $\pi(a, b) = \frac{1}{deg(a)}$ if b is a neighbour of a and $\pi(a, b) = 0$ otherwise. $\mu(a) = \frac{1}{|V|}$ for every $a \in V$, so at time 0 we start from the uniform distribution on V

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 $\forall x_i \in \Omega$

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

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, $P^{(t+1)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}$

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 $P_{1,3}^{(t+1)} = P[X_{t+1} = 3|X_t = 1] = 0$

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 $\pi_x^{(t)} = P[X_t = x]$.

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, $\pi^{(t)}$ 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

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

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, $\pi^{(t)}$ 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

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

- Let $\pi^{(t)} = (0.4, 0.6, 0, 0)$ be a state distribution, then $\pi^{(t+1)} = (0.4, 0.6, 0, 0)$

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, $\pi^{(t)}$ 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

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

- Let $\pi^{(t)} = (0.4, 0.6, 0, 0)$ be a state distribution, then $\pi^{(t+1)} = (0.4, 0.6, 0, 0)$
- Let $\pi^{(t)} = (0, 0, 0.5, 0.5)$ be a state distribution, then $\pi^{(t+1)} = (0, 0, 0.5, 0.5)$

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, $\pi^{(t)}$ 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

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

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

Stationary distributions

Definition

A stationary distribution of a finite Markov chain with transition matrix P is a probability distribution π such that $\pi P = \pi$.

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

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.

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

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

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

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.

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 $x, \bigcirc, \bigcirc, \Box, \bigcirc, \bigcirc, \Box, \cdots$, only the squares are possible to be x.

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 $x, \bigcirc, \bigcirc, \Box, \bigcirc, \bigcirc, \Box, \cdots,$ 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.

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 $x, \bigcirc, \bigcirc, \square, \bigcirc, \bigcirc, \square, \cdots,$ 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 $\lim_{t\to\infty} \pi^{(0)} P^t = \pi$.

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

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 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

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 $x_1, x_2, \cdots, x_n \geq 0$

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

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 $A\overrightarrow{x} = \overrightarrow{x}$

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 $B = 0.85A + 0.15$(matrix with every entry $\frac{1}{n}$), where B is also a Markov chain
  - A page's rank is the probability the random user will end up on that page, equivalently

The curse of dimensionality

- Modern machine learning is usually concerned with high-dimensional objects

The curse of dimensionality

- Modern machine learning is usually concerned with high-dimensional objects
- Consider learning a distribution over $x \in \{0, 1\}^N$

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)

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

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

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

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

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

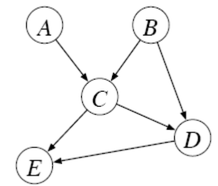
- The special structure graphical models assume is conditional independence
- If you would like to guess the value of some variable x_i, then once you know the values of some "neighboring" variables $x_{\mathcal{N}(i)}$, then you get no additional benefit from knowing all other variables
- Turns out,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

Conditional independence

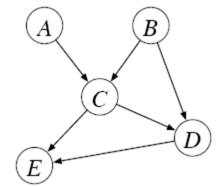
- X is independent of Y if "knowing Y doesn't help you to guess X": $X \perp Y \leftrightarrow P(X, Y) = P(X)P(Y)$
- X is independent of Y given Z if "once you know Z, knowing Y doesn't help you to guess X"

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

Representing knowledge through graphical models



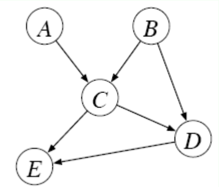
Representing knowledge through graphical models



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

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

Representing knowledge through graphical models



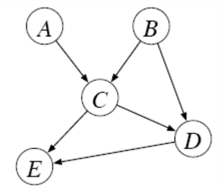
A graphical model is a probability distribution written in a factorized form. For example

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

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

Representing knowledge through graphical models



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

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

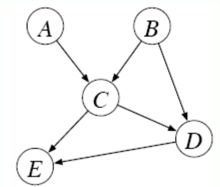
Graph

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

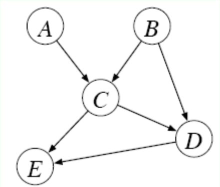
- Vertices correspond to random variables
- Edges represent statistical dependencies between the variables

Outline

Directed acyclic graphical models



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?

How to do learning

Maximum likelihood

Given a fixed graph, how to do learning?

- Natural criterion

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

How to do learning

Maximum likelihood

Given a fixed graph, how to do learning?

- Natural criterion

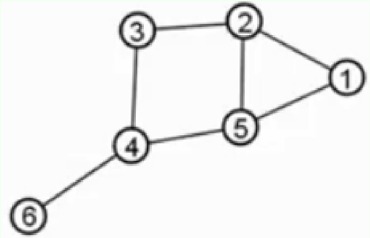
$$\arg\max_{\theta} L(\theta), L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log P(x_d | x_{\pi(d)}; \theta)$$

- Solution is empirical conditionals

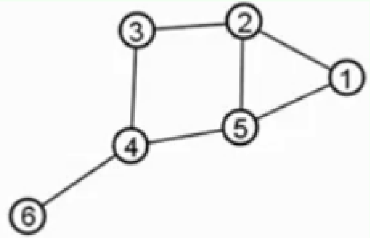
$$P(X_i = x_i | X_{\pi(i)} = x_{\pi(i)}, \theta) = \frac{\#[X_i = x_i, X_{\pi(i)} = x_{\pi(i)}]}{\#[X_{\pi(i)} = x_{\pi(i)}]}$$

Outline

Undirected graphs



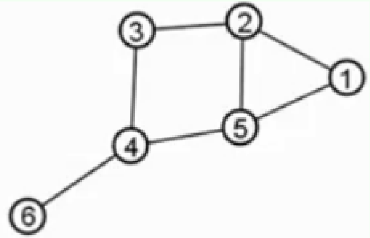
Undirected graphs



Which is true

a. $x_1 \perp x_3 | x_2$

Undirected graphs

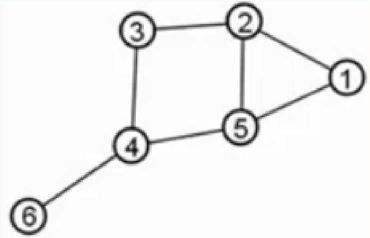


Which is true

a. $x_1 \perp x_3 | x_2$

b. $x_1 \perp x_3 | x_{2,4}$

Undirected graphs



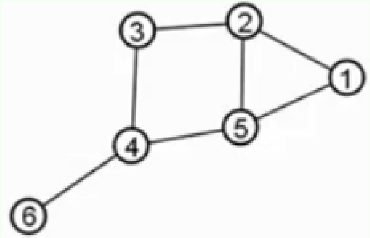
Which is true

a. $x_1 \perp x_3 | x_2$

b. $x_1 \perp x_3 | x_{2,4}$

c. $x_1 \perp x_3 | x_{2,5}$

Undirected graphs



Which is true

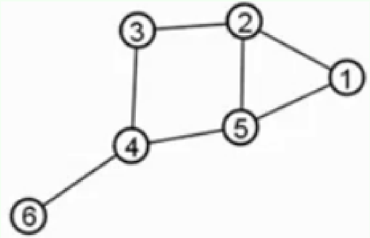
a. $x_1 \perp x_3 | x_2$

b. $x_1 \perp x_3 | x_{2,4}$

c. $x_1 \perp x_3 | x_{2,5}$

d. $x_1 \perp x_6 | x_{2,3,4,5}$

Undirected graphs



Which is true

a. $x_1 \perp x_3 | x_2$

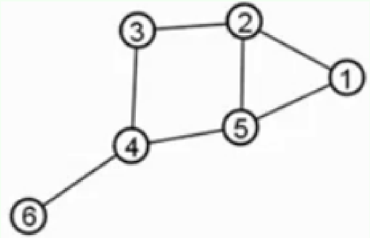
b. $x_1 \perp x_3 | x_{2,4}$

c. $x_1 \perp x_3 | x_{2,5}$

d. $x_1 \perp x_6 | x_{2,3,4,5}$

e. $x_1 \perp x_6 | x_{2,4}$

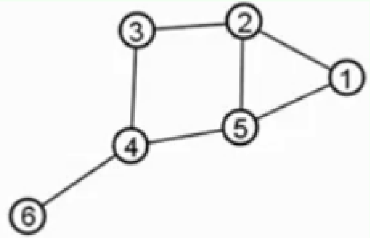
Undirected graphs



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. $x_1 \perp x_6 | x_2$

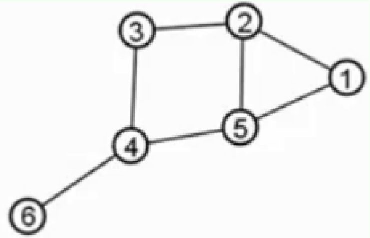
Undirected graphs



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. $x_1 \perp x_6 | x_4$

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. $x_1 \perp x_6 | x_{2,3,4,5}$

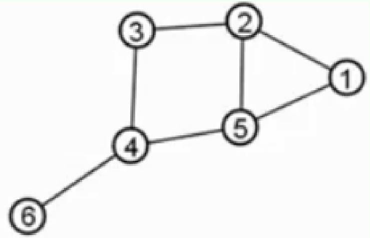
e. $x_1 \perp x_6 | x_{2,4}$

f. $x_1 \perp x_6 | x_2$

g. $x_1 \perp x_6 | x_4$

h. $x_{1,6} \perp x_{3,5} | x_4$

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 $P(x)$?

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

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"

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 ψ

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 ψ
- 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 $\theta = \log \frac{p}{1-p}, \phi(x) = x, A(\theta) = \log\left(\frac{1}{1-p}\right)$

Examples

Examples for exponential family

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

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

Examples

Examples for exponential family

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

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

- Bernoulli, Gaussian, Binomial, Poisson, Exponential, Weibull, Laplace, Gamma, Beta, Multinomial, Wishart distributions are all exponential families

Maximum likelihood learning

MLE

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

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

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 $\theta_i := \theta_i + \lambda \frac{\partial}{\partial \theta_i} L(\theta)$

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 $\theta_i := \theta_i + \lambda \frac{\partial}{\partial \theta_i} L(\theta)$
- $\frac{\partial}{\partial \theta_i} A(\theta) = E(\phi(x)(i))$

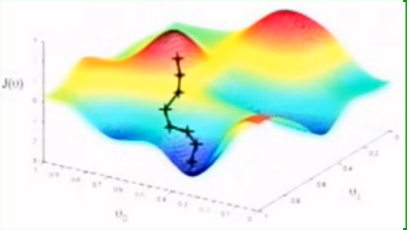
Maximum likelihood learning

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 $\frac{1}{D}\sum_{d=1}^{D} \phi(x_d) = \widehat{E}_{\theta}(\phi(x))$



Maximum likelihood learning Cont'd

MLE

For a distribution of the exponential family in Equation **??**, given data $D = (x_1, \cdots, x_n)$ with i.i.d $x_i \in \mathbb{R}^d$, our goal is to compute the value θ_{MLE} such that

$$\theta_{MLE} = \arg\max_{\theta \in \Theta} f(D|\theta),$$

Maximum likelihood learning Cont'd

MLE

For a distribution of the exponential family in Equation **??**, given data $D = (x_1, \cdots, x_n)$ with i.i.d $x_i \in \mathbb{R}^d$, our goal is to compute the value θ_{MLE} such that

$$\theta_{MLE} = \arg\max_{\theta \in \Theta} f(D|\theta),$$

Maximum likelihood learning Cont'd

MLE

For a distribution of the exponential family in Equation **??**, given data $D = (x_1, \cdots, x_n)$ with i.i.d $x_i \in \mathbb{R}^d$, our goal is to compute the value θ_{MLE} such that

$$\theta_{MLE} = \arg\max_{\theta \in \Theta} f(D|\theta),$$

where likelihood $f(D|\theta)$ can be computed as

$$\begin{aligned} f(D|\theta) &= \prod_{i=1}^{n} f(x;\theta) = \prod_{i=1}^{n} \exp\left(\theta^T S(x_i) - z(\theta)\right) h(x_i) \\ &= \exp\left(\theta^T \sum_{i=1}^{n} S(x_i) - nz(\theta)\right) \prod_{i=1}^{n} h(x_i) \\ &\equiv \exp\left(\theta^T S(D) - nz(\theta)\right) h(D). \end{aligned}$$

Maximum likelihood learning Cont'd

MLE

Thus we obtain the log-likelihood as $\log f(D|\theta) = -nz(\theta) + \eta(\theta)^T S(D) + h(D)$.

Maximum likelihood learning Cont'd

MLE

Thus we obtain the log-likelihood as $\log f(D|\theta) = -nz(\theta) + \eta(\theta)^T S(D) + h(D)$. We can obtain the estimator after we take derivatives of the log-likelihood:

$$\frac{\partial}{\partial \theta_j} \log f(D|\theta) = -n \frac{\partial}{\partial \theta_j} z(\theta) + S_j(D).$$

Maximum likelihood learning Cont'd

MLE

Thus we obtain the log-likelihood as $\log f(D|\theta) = -nz(\theta) + \eta(\theta)^T S(D) + h(D)$. We can obtain the estimator after we take derivatives of the log-likelihood:

$$\frac{\partial}{\partial \theta_j} \log f(D|\theta) = -n \frac{\partial}{\partial \theta_j} z(\theta) + S_j(D).$$

In terms of maximum likelihood theory for an exponential family [**?**], we have

$$\nabla z(\theta) = E_\theta S(x) \Rightarrow n E_\theta S(X) = S(D) = \sum_{i=1}^{n} S(x_i). \tag{2}$$

Maximum likelihood learning Cont'd

MLE

Thus we obtain the log-likelihood as $\log f(D|\theta) = -nz(\theta) + \eta(\theta)^T S(D) + h(D)$. We can obtain the estimator after we take derivatives of the log-likelihood:

$$\frac{\partial}{\partial \theta_j} \log f(D|\theta) = -n \frac{\partial}{\partial \theta_j} z(\theta) + S_j(D).$$

In terms of maximum likelihood theory for an exponential family [**?**], we have

$$\nabla z(\theta) = E_\theta S(x) \Rightarrow n E_\theta S(X) = S(D) = \sum_{i=1}^{n} S(x_i). \tag{2}$$

Finally, we have

$$E_{\theta_{MLE}} S(X) = \frac{1}{n} \sum_{i=1}^{n} S(x_i). \tag{3}$$

Example of four vertex undirected model

Factorization

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

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

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

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

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

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

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

Undirected models

Exponential family

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

$$p(x;\theta) = \exp(\theta^T\phi(x) - A(\theta)) \tag{4}$$

$$\phi(x) = \{\mathbb{I}_{x_c=x_c^*} | c \in \mathcal{C}, \text{all possible } x_c^*\} \tag{5}$$

Undirected models

Exponential family

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

$$p(x;\theta) = \exp(\theta^T \phi(x) - A(\theta)) \tag{4}$$

$$\phi(x) = \{\mathbb{I}_{x_c = x_c^*} | c \in \mathcal{C}, \text{all possible } x_c^*\} \tag{5}$$

- An undirected model is an E.F. where $\phi(x)$ has indicator functions for every configuration of every clique

Undirected models

Exponential family

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

$$p(x;\theta) = \exp(\theta^T\phi(x) - A(\theta)) \tag{4}$$

$$\phi(x) = \{\mathbb{I}_{x_c=x_c^*}|c\in\mathcal{C}, \text{all possible } x_c^*\} \tag{5}$$

- 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	$P(x_i \mid x_{pa(i)}; \theta) = \widehat{P}(x_i \mid x_{pa(i)})$	$P(x_c; \theta) = \widehat{P}(x_c)$

Take-home messages

- Markov chain
- Graphical model
  - Directed model
  - Undirected model