Probabilistic Graphical Models

Lectures 22

Introduction to Learning

Learning Bayesian Networks

Sampling vs. Learning

$$P_\theta(X) = P(X;\theta) = P(X|\theta)$$

Sampling $\quad P_\theta(X) = \overset{\text{given}}{\checkmark} \implies$ generate data $X^1, X^2, \dots, X^m$

i.i.d samples

Learning $\quad$ data $X^1, X^2, \dots, X^m = \overset{\text{given}}{\checkmark} \implies$ find $P_\theta(X)$

$$\text{find } P_\theta(X) \begin{cases} \text{find parameters } \theta \\ \text{find structure (connections in BN/MRF graph)} \end{cases}$$

Likelihood function

K. N. Toosi
University of Technology

given data $X^1, X^2, \dots, X^m$ what is the probability of occurrence of $X^1, \dots, X^m$?

$$Pr(X^1, X^2, \dots, X^m) \underset{\substack{\text{independet} \\ \text{samples}}}{=} \prod_{i=1}^{m} Pr(X^i) \underset{\substack{\text{all are} \\ \text{samples} \\ \text{of } P_\theta(X)}}{=} \prod_{i=1}^{m} P_\theta(X^i)$$

$$\ell(\theta) = \prod_{i=1}^{m} P_\theta(X^i)$$ likelihood

Maximum Likelihood Solution

K. N. Toosi
University of Technology

$i=1$

One Solution: choose θ that maximized the likelihood

$$\theta^* = \underset{\theta}{\arg\max}\ \ell(\theta) = \underset{\theta}{\arg\max} \prod_{i=1}^{m} P_\theta(X^i)$$

maximum-likelihood solution

Maximum Likelihood Solution

K. N. Toosi
University of Technology

$i=1$

One Solution: choose θ that maximized the likelihood

$$\theta^* = \underset{\theta}{\arg\max}\ \ell(\theta) = \underset{\theta}{\arg\max} \prod_{i=1}^{m} P_\theta(X^i)$$

maximum-likelihood solution

Example: tossing a coin

(unfair) coin $\quad P_\theta(X) \quad X \in \{H, T\}$

$$\theta = Pr(X=H) = P_\theta(H) \qquad Pr(X=T) = 1-\theta = P_\theta(T)$$

$$X^1, X^2, \dots, X^m$$

H H T H T T .

Example: tossing a coin

27

$$\ell(\theta) = \prod_{i=1}^{m} P_\theta(x^i) \qquad P_\theta(x^i) = \begin{cases} \theta & x^i = H \\ 1-\theta & x^i = T \end{cases}$$

$$= \prod_{i=1}^{m} \left(\theta\, \mathbb{1}(x^i = H) + (1-\theta)\, \mathbb{1}(x^i = T)\right)$$

n_H = no. of H

n_T = no. of T

$m = n_H + n_T$

$$\Longrightarrow \underline{\ell(\theta) = \theta^{n_H} (1-\theta)^{n_T}}$$

Example: tossing a coin

$$\ell\left(\theta = \frac{n_H}{m}\right) = \left(\frac{n_H}{m}\right)^{n_H} \left(1 - \frac{n_H}{m}\right)^{n_T}$$

$$= \left(\frac{n_H}{m}\right)^{n_H} \left(\frac{n_T}{m}\right)^{n_T} \quad (>0 \text{ if } n_H, n_T > 0)$$

$$\cancel{\ell(\theta)} = \ell(\theta = 0) = \ell(\theta = 1) = 0 \quad (\text{if } n_H, n_T > 0)$$

$$\theta^* = \frac{n_H}{n_H + n_T} = \frac{n_H}{m}$$

maximum-likelihood soluti

log-likelihood $= \ell\ell(\theta)$

Example: tossing a coin - log-likelihood

K. N. Toosi University of Technology

$n_H + n_T$ $\qquad m$

$$\theta^* = \arg\max_\theta \ell(\theta) = \arg\max_\theta \underbrace{\log \ell(\theta)}_{\text{log-likelihood} = \ell\ell(\theta)} = \arg\max_\theta \log \prod_{i=1}^{m} P_\theta(X^i)$$

$$= \arg\max_\theta \sum_{i=1}^{m} \log P_\theta(X^i)$$

$$= \arg\max_\theta \; n_H \log\theta + n_T \log(1-\theta)$$

$$\max \ell\ell(\theta) \Rightarrow \frac{d}{d\theta}\ell\ell(\theta) = \frac{d}{d\theta} n_H \log\theta + n_T \log(1-\theta) \Rightarrow \frac{n_H}{\theta} - \frac{n_T}{1-\theta}$$

$$= \frac{n_H(1-\theta) - n_T\theta}{\theta(1-\theta)} = 0 \Longrightarrow \boxed{\theta^* = \frac{n_H}{n_H + n_T}}$$

Sufficient statistics

data $X^1, X^2, \dots, X^m \Rightarrow \ell(\theta) = \theta^{n_H} (1-\theta)^{n_T}$ 27

$$\boxed{\begin{aligned} n_H &= \sum 1(X^i = H) \\ n_T &= \sum_{i=1}^{m} 1(X^i = T) \end{aligned}}$$

$$\ell\ell(\theta) = n_H \log\theta + n_T \log(1-\theta)$$

for calculating the likelihood we only need n_H, n_T (not the complete data)

n_H, n_T: sufficient statistics

Example 2: normal distribution

$$P_\theta(x) = P_{(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Example 2: Normal Distribution

Example 2: normal distribution

$$P_\theta(x) = P_{(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

data $x^1, x^2, \ldots, x^m$

find maximum-likelihood solution for μ, σ

Example 2: Normal Distribution

$$ll(\theta) = \sum_{i=1}^{m} \log P_\theta(x^i) = \sum_{i=1}^{m} \left(-\log\sqrt{2\pi} - \log\sigma - \frac{1}{2}\frac{(x^i-\mu)^2}{\sigma^2} \right)$$

$$ll(\theta) = C - m\log\sigma - \frac{1}{2}\,\frac{\sum_{i=1}^{m}\left[(x^i)^2 + \mu^2 - 2\mu x^i\right]}{\sigma^2}$$

$$= C - m\log\sigma - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{m}(x^i)^2 - 2\mu\sum_{i=1}^{m} x^i + m\mu^2 \right)$$

sufficient statistics $\left(\sum (x^i)^2, \sum x^i, m \right)$

Example 2: Normal Distribution

$$\frac{\partial}{\partial\mu} \cancel{\theta}\, ll(\theta) = \frac{\partial}{\partial\mu} ll(\mu,\sigma) = -\frac{1}{2\sigma^2}\left(-2\sum_{i=1}^{m} x^i + 2m\mu\right) = 0$$

$$\Rightarrow 2m\mu = 2\sum_{i=1}^{m} x^i \Rightarrow \boxed{\mu^* = \frac{1}{m}\sum_{i=1}^{m} x^i}$$

$$\frac{\partial}{\partial\sigma} ll(\theta) = \frac{-m}{\sigma} - \frac{1}{2}\times(-2)\frac{\sum (x^i-\mu^*)^2}{\sigma^3} = 0$$

$$m\sigma^2 = \sum_{i=1}^{m} (x^i-\mu^*)^2$$

$$\Rightarrow \boxed{\sigma^2 = \frac{1}{m}\sum (x^i-\mu^*)^2} \qquad \sigma = \sqrt{\sigma^2}$$

Example 3: Tossing a dice

K. N. Toosi
University of Technology

$P_{\theta}(X)$ $\quad X \in \{1, 2, \dots, q\}$ $\quad$ 27 IV

$X \in \{1, 2, 3, 4, 5, 6\}$ $\quad \theta = (\theta_1, \theta_2, \dots, \theta_q)$

$\theta_1 = Pr(X = 1)$

$\theta_2 = Pr(X = 2)$

$\vdots$

$\theta_q = Pr(X = q) = 1 - \theta_1 - \theta_2 - \cdots - \theta_{q-1}$

$\sum_{j=}^{q} \theta_j = 1$ $\quad$ data $= X^1, X^2, \dots, X^m$

$l(\theta) = l(\theta_1 - \theta_q) = \prod_{i=1}^{m} P_{\theta}(X) = \theta_1^{n_1} \, \theta_2^{n_2} \cdots \theta_q^{n_q}$

Example 3: Tossing a dice

$$\ell(\theta) = \ell(\theta_1 - \theta_q) = \prod_{i=1}^{m} P_\theta(X) = \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_q^{n_q}$$

suff. stat: $n_j = \sum_{i=1}^{m} 1(X^i = j) = \# X^i = j$

$$\ell\ell(\theta) = \sum_{i=1}^{m} n_i \log \theta_i$$

Example 3: Tossing a dice

$$\theta^* = \underset{\theta_1, \theta_2 \ldots \theta_q}{\operatorname{argmax}} \sum_{j=1}^{q} n_j \log \theta_j \quad \text{subject to} \quad \sum_{j=1}^{q} \theta_j = 1$$

lagrange multipliers

or

or $\theta_q = 1 - \sum_{j=1}^{q-1} \theta_j$

$$\Longrightarrow \theta_j^* = \frac{n_j}{\sum_{k=1}^{q} n_k} = \frac{n_j}{m}$$

PGM problems

PGM $P_\theta(X) = P_\theta(X_1, X_2, \ldots, X_n)$ n: large

data: $X^1 = (X_1^1, X_2^1, X_3^1, \ldots, X_n^1)$

$X^2 = (X_1^1, X_2^1, \ldots, X_n^1)$

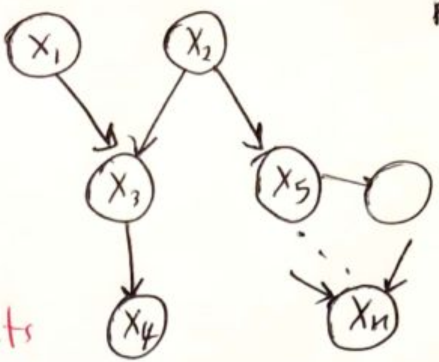
$\vdots$

$X^m = (X_1^1, X_2^1, \ldots, X_n^1)$

Bayes Nets Parameter Learning

K. N. Toosi
University of Technology

pgm 22 III



$$P_\theta(X) = P_\theta(X_1, X_2, \dots, X_n)$$

$$= \prod_{i=1}^{n} P_\theta(X_i \mid \underbrace{X_{P_i}}_{\text{parents of } X_i})$$

Data: $X^1, X^2, \dots, X^m = (x_1^1, x_2^1, \dots, x_n^1), (x_1^2, x_2^2, \dots, x_n^2), \dots, (x_1^m, x_2^m, \dots, x_n^m)$

$$\ell\ell(\theta) = \sum_{k=1}^{m} \log P_\theta(X^k) = \sum_{k=1}^{m} \log \prod_{i=1}^{n} P_\theta(X_i^k \mid X_{P_i}^k)$$

$$= \sum_{k=1}^{m} \sum_{i=1}^{n} \log P_\theta(X_i^k \mid X_{P_i}^k)$$

No shared parameters

1: Each CPD has its own parameters

$$\theta = (\theta_1, \theta_2, \ldots, \theta_n)$$

$$P_\theta(X_i \mid X_{P_i}) = P_{\theta_i}(X_i \mid X_{P_i})$$

$$\Rightarrow ll(\theta) = \sum_{k=1}^{m} \sum_{i=1}^{n} \log P_\theta(X_i^k \mid X_{P_i}^k)$$

$$= \sum_{k=1}^{m} \sum_{i=1}^{n} \log P_{\theta_i}(X_i^k \mid X_{P_i}^k)$$

$$= \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{m} \log P_{\theta_i}(X_i^k \mid X_{P_i}^k)}_{\text{local log-likelihood}}$$

No shared parameters

K. N. Toosi
University of Technology

$$\Rightarrow ll(\theta) = \sum_{k=1}^{m} \sum_{i=1}^{n} \log P_{\theta}(X_i^k \mid X_{P_i}^k)$$

$$= \sum_{k=1}^{m} \sum_{i=1}^{n} \log P_{\theta_i}(X_i^k \mid X_{P_i}^k)$$

$$= \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{m} \log P_{\theta_i}(X_i^k \mid X_{P_i}^k)}_{\text{local log-likelihood}}$$

$$\frac{\partial}{\partial \theta_j} ll(\theta) = \frac{\partial}{\partial \theta_j} ll(\theta_1, \theta_2, \dots, \theta_n)$$

$$= \sum_{k=1}^{m} \frac{\partial}{\partial \theta_j} \log P_{\theta_j}(X_i^k \mid X_{P_i}^k)$$

$$= \sum_{k=1}^{m} \frac{\frac{\partial}{\partial \theta_j} P_{\theta_j}(X_i^k \mid X_{P_i}^k)}{P_{\theta}(X_i^k \mid X_{P_i}^k)}$$

gradient w.r.t. θ_j

$\Rightarrow$ each θ_j can be found independently of other θ_i-s

No shared parameters - table representation

K. N. Toosi University of Technology

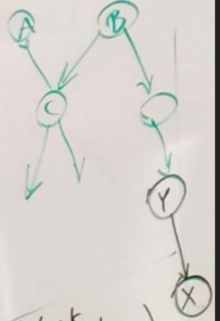
$$\sum_{i=1}^{n} \underbrace{\sum_{k=1}^{m} \log P_{\theta_i}(X_i^k \mid X_{P_i}^k)}_{\text{local log-likelihood}}$$

table representation

Eample $P(X_i \mid Y_i)$ $X_i, X_i \in \{0,1\}$

X_i	Y_i	$P(X_i \mid Y_i)$
1	0	θ_0
0	0	$1-\theta_0$
1	1	θ_1
0	1	$1-\theta_1$

local likelihood



$$\sum_{k=1}^{m} \log P(X_i^k \mid Y_i^k)$$

$$\frac{\partial}{\partial \theta_0} \sum_{k=1}^{m} \log P(X_i^k \mid Y_i^k)$$

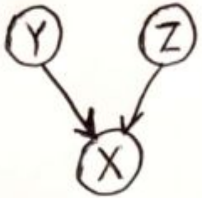
$$\frac{\partial}{\partial \theta_0} \underbrace{\sum_{\substack{k=1 \\ Y_i^k=0}}^{m} \log P(X_i^k \mid 0)}_{\text{a function of } \theta_0} + \underbrace{\cancel{\sum_{\substack{k=1 \\ Y_i^k=1}}^{m} \log P(X_i^k \mid 1)}}_{\text{a function of } \theta_1}$$

$$= \frac{\partial}{\partial \theta_0}\left(n_{10}\log\theta_0 + n_{00}\log(1-\theta_0)\right) \Rightarrow \theta_0^* = \frac{n_{10}}{n_{10}+n_{11}}$$

$$n_{10} = \#(X_i^k = 1, Y_i^k = 0)$$

$$n_{00} = \#(X_i^k = 0, Y_i^k = 0)$$

No shared parameters - table repres



$X \in \{1, 2, \dots, c\}$

$Y, Z \in \{0, 1\}$

$\gamma^i_{00} =$

$\gamma^i_{01} = Pr(X = i \mid Y = 0, Z = 1)$

$\gamma^i_{10} = Pr(X = i \mid Y = 1, Z = 0)$

$\gamma^i_{11} = Pr(X = i \mid Y = 1, Z = 1)$

$\sum_i \gamma^i_{10} = 1$

local likelihood $\sum_{k=1}^{m} \log P_\gamma(X^k \mid Y^k, Z^k)$

$$\sum_{k=1}^{m} \log \gamma^{x^k}_{y^k, z^k}$$

No shared parameters - table representation

local likelihood $\sum_{k=1}^{m} \log P_{\gamma}(X^k \mid Y^k, Z^k)$

$$\sum_{k=1}^{m} \log \gamma^{X^k}_{Y^k, Z^k}$$

$$\sum_{y=0}^{1} \sum_{z=0}^{1} \sum_{x=1}^{c} \left(\log \gamma^{x}_{yz}\right) \cdot \underbrace{\#\left(X^k = x, Y^k = y, Z^k = z\right)}_{\text{sufficient statistics}}$$

$$\boxed{\sum_{x} \gamma^{x}_{yz} = 1}$$

$$\gamma^{x}_{yz} = \frac{\#(X^k = x, Y^k = y, Z^k = z)}{\#(Y^k = y, Z^k = z)}$$

No shared parameters - table representation

$$\sum_{y=0}^{1}\sum_{z=0}^{1}\sum_{x=1}^{c}\left(\log \gamma_{yz}^{x}\right)\cdot \underbrace{\left[\#\left(X^k=x,\ Y^k=y,\ Z^k=z\right)\right]}_{\text{sufficient statistics}}$$

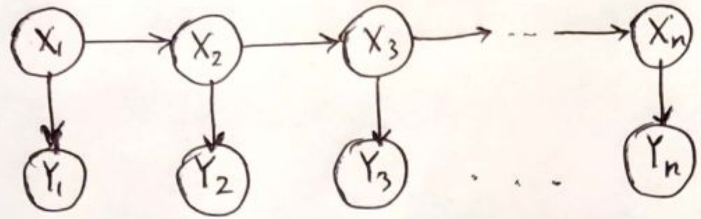
$$\boxed{\sum_{x} \gamma_{yz}^{x} = 1}$$

$$\gamma_{yz}^{x} = \frac{\#\left(X^k=x,\ Y^k=y,\ Z^k=z\right)}{\#\left(Y^k=y,\ Z^k=z\right)}$$

$$= \frac{\sum_{k=1}^{m} 1\left(X^k=x,\ Y^k=y,\ Z^k=z\right)}{\sum_{k=1}^{m} 1\left(Y^k=y,\ Z^k=z\right)}$$

shared parameters

K. N. Toosi
University of Technology



$$P_\theta(X_1 - X_n, Y_1 - Y_n) = P_\alpha(X_1) \prod_{i=2}^{n} P_\beta(X_i | X_{i-1}) \prod_{i=1}^{n} P_\gamma(Y_i | X_i)$$

$$\theta = (\alpha, \beta, \gamma)$$

$$ll(\theta) = \sum_{k=1}^{m} \log P_\alpha(X_1^k) + \sum_{k=1}^{m} \sum_{i=2}^{n} \log P_\beta(X_i^k | X_{i-1}^k)$$
$$+ \sum_{k=1}^{m} \sum_{i=1}^{n} \log P_\gamma(Y_i^k | X_i^k)$$

$$\frac{\partial ll(\theta)}{\partial \beta} = \sum_{k=1}^{m} \sum_{i=2}^{n} \frac{\partial}{\partial \beta} \log P_\beta(X_i^k | X_{i-1}^k)$$

shared parameters - table representation

table representation $X_i \in \{0,1\}$

$\beta_0 = \beta_{00}, \beta_{10}, \beta_{11}, \beta_{01}$

$\beta_{01} = P(X_i = 0 \mid X_{i-1} = 1)$ independent of i

* ML solution

$$\beta_{01}^* = \frac{\#(X_i^k = 0, X_{i-1}^k = 1)}{\#(X_{i-1}^k = 1)}$$

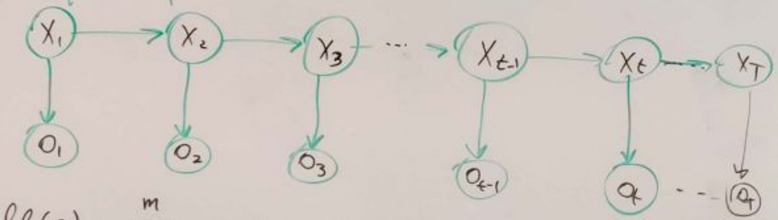
$$= \frac{\sum_{i=2}^{n} \sum_{k=1}^{m} 1(X_i^k = 0, X_{i-1}^k = 1)}{\sum_{i=2}^{n} \sum_{k=1}^{m} 1(X_{i-1}^k = 1)}$$

shared parameters - table representation

Shared parameters $X^1, X^2, \dots, X^m$

$$ll(\theta) = \sum_{k=1}^{m} \sum_{i=1}^{n} \log P(X_i^k \mid X_{P_i}^k)$$

Example X_1



$$ll(\theta) = \sum_{k=1}^{m} \log P(X_1^k) + \sum_{k=1}^{m} \sum_{i=2}^{T} \log P(X_i^k \mid X_{i-1}^k) + \sum_{k=1}^{m} \sum_{i=1}^{T} \log P(O_i^k \mid X_i^k)$$

$$P(X_t = k \mid X_{t-1} = l) = f_\theta(k, l)$$

$$P(X_t \mid X_{t-1}) = f_\theta(X_t, X_{t-1}) =$$

$$P(O_t \mid X_t) = g_\gamma(O_t, X_t)$$

table representation:

$$P(X_t = j \mid X_{t-1} = l) = \lambda_{j,l}$$

$$P(O_t = j \mid X_t = l) = \gamma_{j,l}$$

$$\frac{\partial ll(\lambda, \gamma)}{\partial \lambda_{l,n}} = \sum_{k=1}^{m} \sum_{i=2}^{T} \log P(X_i^k \mid X_{i-1}^k)$$

shared parameters - table representation

$$P(X_t = k \mid X_{t-1} = \ell) = f_\theta(k, \ell)$$

$$P(X_t \mid X_{t-1}) = f_\theta(X_t, X_{t-1}) =$$

$$P(O_t \mid X_t) = g_\gamma(O_t, X_t)$$

table representation:

$$P(X_t = j \mid X_{t-1} = \ell) = \lambda_{j,\ell}$$

$$P(O_t = j \mid X_t = \ell) = \gamma_{j,\ell}$$

$$\frac{\partial \ell\ell(\lambda,\gamma)}{\partial \lambda_{\ell,m}} = \sum_{k=1}^{m} \sum_{i=2}^{T} \log P(X_i^k \mid X_{i-1}^k)$$

Data $X_1^1, X_2^1, \ldots, X_T^1, O_1^1, O_2^1, \ldots, O_T^1$

$X_1^2, X_2^2, \ldots, X_T^2, O_1^2, O_2^2, \ldots, O_T^2$

$\vdots$

$X_1^m, X_2^m, \ldots X_T^m, O_1^m, \ldots O_T^m$

$$X_t \in \{1, 2, \ldots, q\}$$

$$= \sum_{j=1}^{q} \sum_{\ell=1}^{q} \sum_{\substack{i=2 \\ X_i^k = j \\ X_{i-1}^k = \ell}}^{T} \log P(X_i^k \mid X_{i-1}^k)$$

$$\sum_{j=1}^{q} \sum_{\ell=1}^{q} \sum_{X_i^k = j,\, X_{i-1}^k = \ell} \log P(j \mid \ell)$$

$$\sum_{j=1}^{q} \sum_{\ell=1}^{q} \sum_{\substack{X_i^k = j \\ X_{i-1}^k = \ell}} \log \lambda_{j\ell}$$

$$\lambda_{j\ell}^{*} = \frac{\#(X_i = j,\, X_{i-1} = \ell)}{\#(X_{i-1} = \ell)}$$