Ch22 MA4270

Last Lecture

MLE → learn the parameters (Given the graph structure)

Given a dataset $\mathcal{D} = \{\underline{x}_1, \dots, \underline{x}_t\} \subset [r]^d$ $\underline{x}_i$

$\updownarrow$

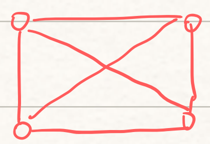
$\underline{x}_i \in \{1, \dots, r\}^d \rightarrow \begin{cases} x_{i1} \\ \vdots \\ x_{id} \end{cases}$

$\underline{x}_t$ is drawn from a BN or MRF $P(\cdot)$

↓

$P(\cdot)$ is unknown ! ⇒ its structure and parameters are completely unknown

Example

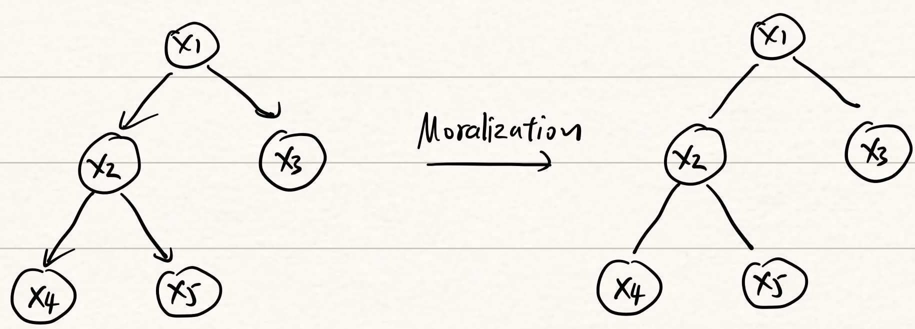


$d=4$

→ no information !

Assume that $P^*(\cdot)$ is a BN in which ① each variable/node x_i has at most 1 parent and ② graph contains no loop.

Eg.



$$P^*(\underline{x}) = \prod_{i=1}^{d} P^*(x_i \mid x_{pa(i)})$$ $\boxed{\text{Last lecture}}$

$\theta_{x_i \mid x_{pa(i)}}$

$$= \prod_{i=1}^{d} \theta_{i|pa(i)}(x_i \mid x_{pa(i)})$$

Claim: For this class of BNs, we can write p^*

$$p^*(\underline{x}) = \prod_{i=1}^{d} \theta_i(x_i) \prod_{\{i,j\} \in E} \frac{\theta_{ij}(x_i, x_j)}{\theta_i(x_i)\,\theta_j(x_j)}$$

E: set of undirected edges in the MRF

$$\begin{cases} \theta_i(x_i) = p_i^*(x_i) \\ \theta_{ij}(x_i, x_j) = p_{ij}^*(x_i, x_j) \end{cases}$$

Pf: $p^*(\underline{x}) = \theta_1(x_1)\, \theta_{2|1}(x_2|x_1)\, \theta_{3|1}(x_3|x_1)\, \theta_{4|2}(x_4|x_2)\, \theta_{5|2}(x_5|x_2)$

$$= \theta_1(x_1) \cdots \theta_5(x_5) \frac{\theta_{2|1}(x_2|x_1)}{\theta_2(x_2)} \frac{\theta_{3|1}(x_3|x_1)}{\theta_3(x_3)} \frac{\theta_{4|2}(x_4|x_2)}{\theta_4(x_4)} \frac{\theta_{5|2}(x_5|x_2)}{\theta_5(x_5)}$$

$$= \prod_{i \in V} \theta_i(x_i) \prod_{\{i,j\} \in E} \frac{\theta_{ij}(x_i, x_j)}{\theta_i(x_i)\,\theta_j(x_j)}$$

$E = \{ \{1,3\}, \{1,2\}, \{2,4\}, \{2,5\} \}$

Let T_d be the set of all tree-stuctured MRFs.

- connected
- acyclic
- undirected graph

(at most 1 parent)

We want to find:

E

$\uparrow$

$$p^* = \underset{p \in T_d}{\operatorname{argmax}} \log P(\mathcal{D} \mid \theta, G)$$

$$P(\underline{x}) = \prod_{i \in V} \theta_i(x_i) \prod_{\{i,j\} \in E} \frac{\theta_{ij}(x_i, x_j)}{\theta_i(x_i)\, \theta_j(x_j)}$$

$$= \underset{p \in T_d}{\operatorname{argmax}} \sum_{t=1}^{n} \log (x_t \mid \theta, G)$$

In other words, we want to find the/a tree-structured MRF that maximizes likelihood $\log P(\mathcal{D} \mid \theta, G)$

Rmk: This is an optimization over 2 objects

① Undirected edge set E

② Given the edges learnt in E, find the parameters

$$\{ \theta_{ij}(x_i, x_j) \; ; \; x_i, x_j \in [r] \}_{\{i,j\} \in E}$$

Problem: ① There are d^{d-2} trees (undirected) on d nodes

$\Downarrow$

Intractable even for moderate d! $(d \approx 10)$

MLE for Learning
$P \triangleq (\theta, G)$

Def: Given a joint dist^{n} $P_{ij}(x_i, x_j)$ on x_i, x_j, the mutual information is

$$I(x_i; x_j) = \sum_{x_i, x_j \in [r] \times [r]} P_{ij}(x_i, x_j) \log \frac{P_{ij}(x_i, x_j)}{P_i(x_i)\, P_j(x_j)}$$

Fun fact: a) $I(x_i; x_j)$ is a measure of the dependence between x_i & x_j!

$\Downarrow$

b) Note: If $x_i \perp\!\!\!\perp x_j$, then $I(x_i; x_j) = 0$!

c) $I(X_i ; X_j) \geq 0$

d) If $I(X_i ; X_j) = 0$, then $X_i \perp\!\!\!\perp X_j$!

e) $I(X_i ; X_j) \leq \log r$ where $r = |X_i| = |X_j|$

Thm: [Chow-Liu (1968)]

The MLE sol^{n} is a tree-structured MRF with

edge set $$E^* = \underset{E:\, G=(V,E) \text{ is a tree}}{\arg\max} \sum_{\{i,j\} \in E} \hat{I}(X_i ; X_j)$$

where $\hat{I}$ is the mutual information of X_i & X_j

computed w.r.t

$r \times r$ matrix for fixed i, j

$$\hat{P}_{ij}(x_i, x_j) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}\{X_{ti} = x_i ; X_{tj} = x_j\}$$

$x_i, x_j \in [r] \times [r]$

$\downarrow$

计算 d 两个 node 间的 $\hat{I}(i;j)$

$$\boxed{\hat{I}(i ; j)} = \sum_{x_i, x_j} \underbrace{\hat{P}_{ij}(x_i, x_j)}_{r \times r \text{ matrix}} \log \frac{\hat{P}_{ij}(x_i, x_j)}{\hat{P}_i(x_i)\, \hat{P}_j(x_j)}$$

$\binom{d}{2}$

Rmk: The optimization for E^* is known as a maximum weight spanning tree (MWST) problem and can be implemented in time $O(d^2 \log d)$ operations.

Prim's Alg:

initialize all vertices

1. Let $E = \phi$, $U = \{1\}$, $V = [d]$
2. While: $U \neq V$

3. let $\{i,j\}$ be the highest weight, s.t $i \in U$, $j \in V \setminus U$

4. $E \leftarrow E \cup \{i,j\}$

5. $U \leftarrow U \cup \{j\}$

6. end

[GREEDY Algorithm]

Procedure for learning ML tree (Chow-Liu Tree)

1. Calculate the empirical (pair wise) mutual information

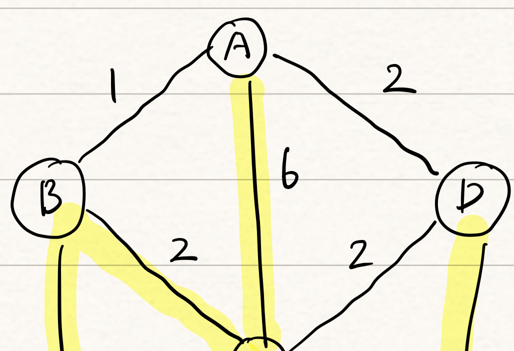
$$\forall\, i,j \in \binom{V}{2} \qquad \hat{I}(X_i; X_j) = f(\hat{P}_{ij})$$

2. Initialize a tree with a single node (say 1)

3. Grow the tree by 1 edge of the edges that connect the tree to vertices not in the tree, find the maximum weight edge and transfer to tree

4. Repeat until we have d-1 edges.

Eg.



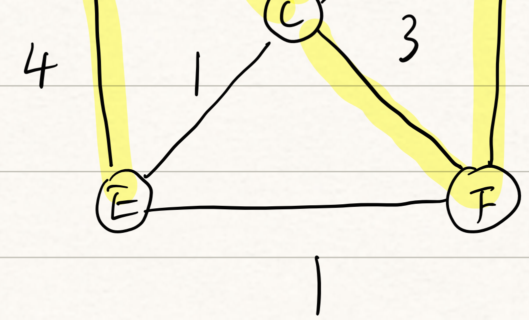
① $U = \{A\}$

② $U = \{A, C\}$

③ $U = \{A, C, F\}$

④ $U = \{A, C, F, D\}$

⑤ $U = \{A, C, F, D, B\}$



⑥ $\mathcal{U} = \{A, C, F, D, B, E\}$

Qa: Why Chow-Liu Thm is RIGHT?

Pf of Chow-Liu Thm:

Idea: Learning of E^* and the parameters $\theta^* = \{\theta_{ij}^*(x_i, x_j) : x_i, x_j \in [r]\}_{\{i,j\} \in E}$ are neatly decoupled!

Pf: $\max\limits_{P} \sum\limits_{t=1}^{n} \log P(x_t \mid \theta, E)$ → x_t is generated from the graph G

$$\equiv \boxed{\max_{E} \max_{\theta}} \sum_{t=1}^{n} \log \left[\prod_{i \in V} \theta_i(x_{ti}) \prod_{\{i,j\} \in E} \frac{\theta_{ij}(x_{ti}, x_{tj})}{\theta_i(x_{ti})\, \theta_j(x_{tj})} \right]$$

↓ critical

Fix a particular set of edges E

Consider inner opt. problem:

Step 1

$$\max_{\theta} \sum_{t=1}^{n} \log \left[\prod_{i \in V} \theta_i(x_{ti}) \prod_{\{i,j\} \in E} \frac{\theta_{ij}(x_{ti}, x_{tj})}{\theta_i(x_{ti})\, \theta_j(x_{tj})} \right]$$

$$= \max_{\theta} \sum_{t=1}^{n} \left[\sum_{i \in V} \log \theta_i(x_{ti}) + \sum_{\{i,j\} \in E} \log \frac{\theta_{ij}(x_{ti}, x_{tj})}{\theta_i(x_{ti})\, \theta_j(x_{tj})} \right]$$

$$= \max_{\theta} n \left[\sum_{i \in V} \sum_{x_i \in [r]} \hat{P}_i(x_i) \log \theta_i(x_i) + \sum_{\{i,j\} \in E} \sum_{x_i, x_j} \hat{P}_{ij}(x_i, x_j) \log \frac{\theta_{ij}(x_i, x_j)}{\theta_i(x_i)\, \theta_j(x_j)} \right]$$

Solve θ for

fix E!

$$= n \left[\sum_{i \in V} \sum_{x_i} \hat{P}_i(x_i) \log \hat{P}_i(x_i) + \sum_{\{i,j\} \in E} \sum_{x_i, x_j} \hat{P}_{ij}(x_i, x_j) \log \frac{\hat{P}_{ij}(x_i, x_j)}{\hat{P}_i(x_i) \hat{P}_j(x_j)} \right]$$

挑选哪两个 node 的参数

Seep2 :

$$\max_E n \left[\sum_{i \in V} \sum_{x_i} \hat{P}_i(x_i) \log \hat{P}_i(x_i) + \sum_{\{i,j\} \in E} \sum_{x_i, x_j} \hat{P}_{ij}(x_i, x_j) \log \frac{\hat{P}_{ij}(x_i, x_j)}{\hat{P}_i(x_i) \hat{P}_j(x_j)} \right]$$

$$= \max_E \sum_{\{i,j\} \in E} \sum_{x_i, x_j} \hat{P}_{ij}(x_i, x_j) \log \frac{\hat{P}_{ij}(x_i, x_j)}{\hat{P}_i(x_i) \hat{P}_j(x_j)}$$

$$= \boxed{\max_E \sum_{\{i,j\} \in E} \hat{I}(X_i ; X_j)}$$

$\Downarrow$

discrete optimization