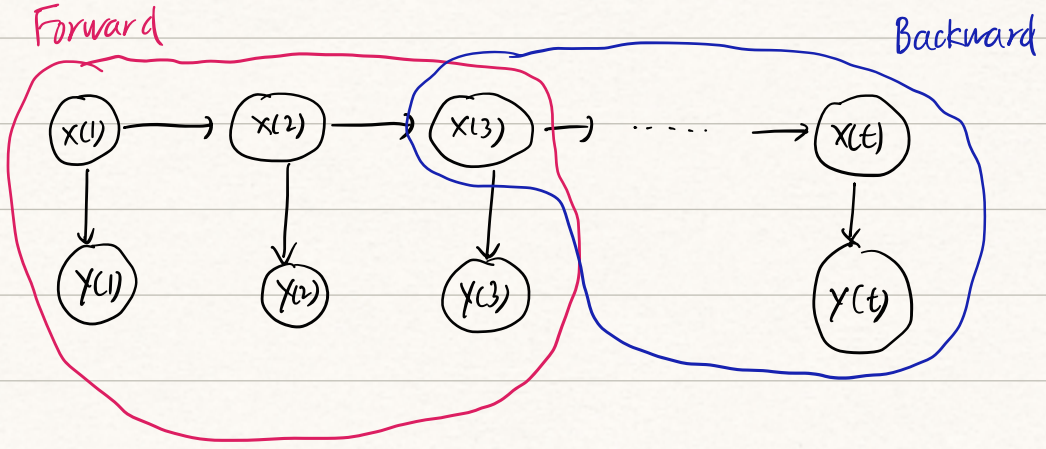
MA4270 Lec20 B-W Algorithm + Bayesian Network

Recall → Forward & Backward Algorithm ⇒ calculate $P(y_1, \dots, y_n)$

$$\begin{cases} \alpha_t(j) = P(y_1, \dots, y_t, X(t) = j), \ j \in S \rightsquigarrow \text{forward message} \\ \beta_t(j) = P(y_{t+1}, \dots, y_n \mid X(t) = j), \ j \in S \rightsquigarrow \text{Backward message} \end{cases}$$



Task 3: (B-W Alg.)

Given observation $(y_1, \dots, y_n)$ from HMM. We want to estimate the paras of the HMM.

paras:
- $\{q(i)\}_{i=1}^{k}$ ⇝ initial
- $\{p_{ij}\}_{i,j=1}^{k}$ ⇝ transition
- $\{P(y \mid i)\}_{i=1}^{k}, \ y \in \mathcal{Y}$ ⇝ emission

⇓

an example of EM

What is GOOD to have (to estimate paras)?

→ It would be good to also have $x_1, \dots, x_n$ in addition to our Hidden Variables

Real Observation.

$(y_1, \dots, y_n)$

如果 Incomplete → $\{y_1, \dots, y_n\}$

likelihood $L(x; \theta) = \prod_{i=1}^{k} P(x_i)$

Complete Dataset = $\{ \underbrace{x_1, \dots, x_n}_{\text{Hidden MC}}, y_1, \dots, y_n \}$ $= \prod_{t=1}^{K} \sum_{s} P(s)\, P(x_t|s)$

Hidden Variable

$$\delta(i|t) = \begin{cases} 1, & \text{if } x_t = i \in S \\ 0, & \text{else} \end{cases} \rightsquigarrow \{ \delta(i|t) : i \in S,\ t = 1, \dots, n \}$$

Hidden Variable

$$\delta(i,j|t) = \begin{cases} 1, & x_t = i,\ x_{t+1} = j \\ 0, & \text{else} \end{cases} \rightsquigarrow \{ \delta(i,j|t),\ i,j \in S,\ t = 1, \dots, n \}$$

$$\log P = \log q(x_1) \cdot \prod_{t=1}^{n-1} P_{x_t, x_{t+1}} \prod_{t=1}^{n} P(y_t | x_t)$$

Given the complete dataset = $\{ x_1, \dots, x_n, y_1, \dots, y_n \}$

we can write down the complete log-likelihood:

$$\ell(\{x_t\}, \{y_t\}; \theta) = \sum_{i=1}^{K} \delta(i|1) \log(q(i)) + \sum_{i=1}^{K} \left(\sum_{t=1}^{n} \delta(i|t) \log P(y_t | i) \right)$$

(initial; emission)

$$+ \sum_{i,j=1}^{K} \left(\sum_{t=1}^{n-1} \delta(i,j|t) \right) \log P_{ij}$$

(transition)

M-step: $\mathbb{E}_{p(\underline{X}|\, y, \theta^{(e)})} [\ell(\{x_t\}, \{y_t\}; \theta)] \rightarrow$ maximize

(X is hidden)

$\Downarrow$

$\theta^{(t+1)}$

E-step: $p^{(e)}(i|t) = P(x(t) = i \mid y_1, \dots, y_n) = \mathbb{E}[\delta(i|t) \mid \{y_t\}, \theta^{(e)}]$

(the quantity we care; $\theta^{(e)}$)

$$p^{(e)}(i,j|t) = P(x(t) = i,\ x(t+1) = j \mid y_1, \dots, y_n)$$

($\theta^{(e)}$)

$$= \mathbb{E}[\delta(i,j|t) \mid \{y_t\}, \theta^{(e)}]$$

Q: How to compute $p(i \mid t, \{y_k\}, \underline{\theta}^{(e)})$, $p(i,j \mid t, \{y_k\}, \underline{\theta}^{(e)})$ $\Rightarrow$ Posterior

① Ans: $p^{(t)}(X(t)=i \mid y_1, \dots y_n, \underline{\theta}^{(e)})$

$$= \frac{P(X(t)=i, y_1, \dots y_n \mid \underline{\theta}^{(e)})}{P(y_1, \dots y_n \mid \underline{\theta}^{(e)})}$$

$$= \frac{P(X(t)=i, y_1, \dots y_n \mid \underline{\theta}^{(e)})}{\sum_{i'} P(X(t)=i', y_1, \dots y_n \mid \underline{\theta}^{(e)})} = \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{i'} \alpha_t(i')\,\beta_t(i')}$$

Numerator $P(X(t)=i, y_1, \dots y_n ; \underline{\theta}^{(e)})$

$= \alpha_t(i)\,\beta_t(i)$ $\rightsquigarrow$ according to paras $\underline{\theta}^{(e)}$

② $p^{(t)}(i,j \mid t) = P(X(t)=i, X(t+1)=j \mid y_1, \dots, y_n ; \underline{\theta}^{(e)})$

$$= \frac{P(y_1, \dots, y_n, X(t)=i, X(t+1)=j ; \underline{\theta}^{(e)})}{\sum_{i',j'} P(y_1, \dots, y_n, X(t)=i', X(t+1)=j' ; \underline{\theta}^{(e)})}$$

Numerator:

$P(y_1, \dots, y_n, X(t)=i, X(t+1)=j)$ (omit $\underline{\theta}^{(e)}$)

$= P(y_{t+2}, \dots y_n \mid X(t+1)=j) \cdot P(X(t+1)=j, X(t)=i, y_1, \dots y_{t+1})$

$$= P(y_{t+2}, \dots, y_n \mid X(t+1)=\tilde{j}) \cdot \underbrace{P(y_{t+1} \mid \tilde{j})\, P(\tilde{i},\tilde{j})}_{P(y_{t+1}, X(t+1)=\tilde{j} \mid X(t)=\tilde{i})} \cdot P(y_1, \dots, y_t, X(t)=\tilde{i})$$

$$= \alpha_t(\tilde{i})\, p_{\tilde{i}\tilde{j}}\, P(y_{t+1} \mid \tilde{j})\, \beta_{t+1}(\tilde{j})$$

$$\text{Thus, } P(\tilde{i},\tilde{j} \mid t) = \frac{\alpha_t(\tilde{i})\, p_{\tilde{i}\tilde{j}}\, P(y_{t+1} \mid \tilde{j})\, \beta_{t+1}(\tilde{j})}{\sum_{\tilde{i}',\tilde{j}'} \alpha_t(\tilde{i}')\, p_{\tilde{i}'\tilde{j}'}\, P(y_{t+1} \mid \tilde{j}')\, \beta_{t+1}(\tilde{j}')}$$

Baum-Welch Alg.

Given $\{y_t\}_{t=1}^{n}$, estimate $\{q(\tilde{i})\}$ $\{p_{\tilde{i}\tilde{j}}\}$ $\{P(y \mid \tilde{i})\}$

origin hidden Variable

→ Would be good to have $\{X_t\}$ induce $\{\delta(\tilde{i},\tilde{j} \mid t)\}$ (sufficient hidden variable) and $\{\delta(\tilde{i} \mid t)\}$

$\mathbb{E}[\delta(\tilde{i},\tilde{j} \mid t) \mid \theta^{(e)}, y]$

→ But $p(\tilde{i} \mid t)$ & $p(\tilde{i},\tilde{j} \mid t)$ can be estimated recursively using α_t & $\beta(t)$

Bayesian Networks

i) A directed acylic (no direct cycle) graph (DAG)

ii) An associated prob. distribution that respects the graph

Example

① → this is OK

②

arc / directed edge

(edge has no direction!)

Directed Cycle!

this is NOT OK!

DAG provides a qualitative description of the (conditional) independence relations among

Prob. dist^n provides a quantitative description.

DAG can be used to ① explain properties of dist^n

② perform efficient computations

HMM: ① $P(y_1, \dots, y_n) \rightarrow$ F-B alg. $\rightarrow O(k^2 n)$

② $(x_1^*, \dots, x_n^*) = \underset{x_1, \dots, x_n}{\arg\max}\ P(x_1, \dots, x_n, y_1, \dots, y_n) \rightarrow$ Viterbi $\rightarrow O(n)$

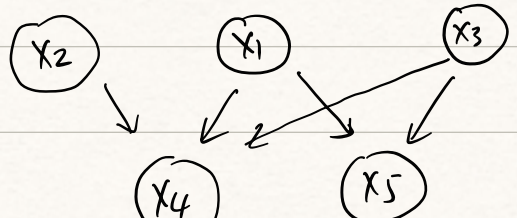
⇓

this can be generated to any DAG

⇓

Exploit the DAG structure to do the above ① & ② efficiently!

[E.g.] Form a joint dist^n.



x_1 & x_2 are parents of x_4

$$P(X_1, \dots, X_5) = P(X_1)\, P(X_2)\, P(X_3)\, P(X_4 \mid X_1, X_2, X_3)\, P(X_5 \mid X_1, X_3)$$

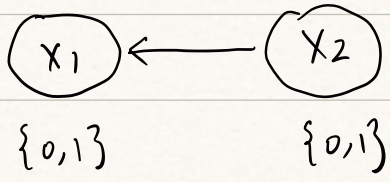
More generally, $\underline{X} = (X_1, \dots, X_d)$

$$P(\underline{X}) = \prod_{k=1}^{d} P(X_k \mid X_{pa(k)})$$

① $pa(k)$ is the set of parents of k

② $X_u = (X_i : i \in u)$

Consider the case where all variables are binary, i.e., $\{0,1\}$

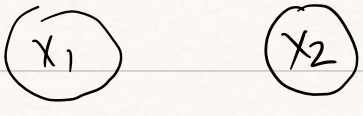
① 

$$\rightsquigarrow P(X) = P(X_1)\, P(X_2 \mid X_1)$$

Q: How many paras ?

Ans: 3 paras.

$$\begin{cases} P(X_1 = 0) \\ P(X_2 = 0 \mid X_1 = 0) \\ P(X_2 = 0 \mid X_1 = 1) \end{cases}$$

② 

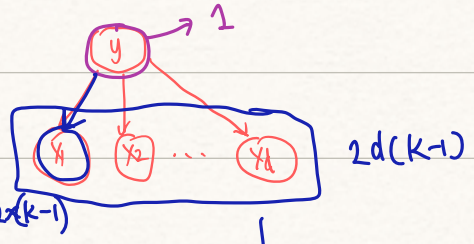
$$P(X) = P(X_1)\, P(X_2)$$

$$\Downarrow$$

# of para : 2.

③ $X_i \in \{1, \dots, K\}$ $\quad y \in \{0,1\}$

$i = 1, 2, \dots, d$



Naive Bayes Model for Binary classification $y \in \{0, 1\}$

Q: How many paras? $\Rightarrow$ # of paras: $2d(K-1)+1$

Ans: $P(x_i | y=0)$ $P(x_i | y=1)$

$K-1$ $K-1$

1 feature $\rightsquigarrow$ $2(K-1)$ paras

d features $\rightsquigarrow$ $2d(K-1)$ paras

$$P(x, y) = P(x|y)\,P(y)$$

$2d(K-1)$ 1

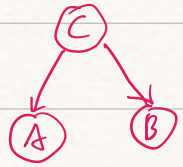
Conditional Independence & Explain away.

Def: A & B are conditional independent given C if

$$P(A=a, B=b \mid C=c) = P(A=a \mid C=c)\, P(B=b \mid C=c)$$

for all a, b, c

E.g:



P(a|

$$P(a, b, c) = P(c)\, P(a|c)\, P(b|c)$$

Qn 1: A & B independent ? → NO

Pf: $p(a,b) = \sum_c p(c)\, p(a|c)\, p(b|c) \neq p(a)\, p(b)$

Qn 2: A & B cond. indepedent given C ? → YES!

Pf: $p(a,b,c) = p(c)\, p(a|c)\, p(b|c)$

$\Leftrightarrow$ $p(a,b|c) = p(a|c)\, p(b|c)$

$\Rightarrow$ $A \perp\!\!\!\perp B \mid C$

E.g. 

$p(a,b,c) = p(a)\, p(c|a)\, p(b|c)$

Qn 1: A & B independent ? → NO

Pf: $p(a,b) = \sum_c p(a)\, p(c|a)\, p(b|c) = p(a) \sum_c p(c|a)\, p(b|c) \neq p(a)\, p(b)$

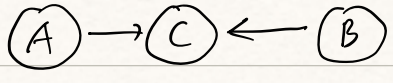
Qn 2: A & B cond. indepedent given C ? → YES!

Pf: $p(a,b|c) = \dfrac{p(a,b,c)}{p(c)}$

$= \dfrac{p(a)\, p(c|a)\, p(b|c)}{p(c)}$

$= p(a|c)\, p(b|c)$

$\Rightarrow A \perp\!\!\!\perp B \mid C$

E.g.: 

$p(a,b,c) = p(a)\, p(b)\, p(c|a,b)$

Qn 1: A & B independent ? ⇒ YES!

Pf: $P(a,b) = \sum_c P(a,b,c)$

$= P(a) \cdot P(b) \cdot \sum_c P(c|a,b)$

$= P(a)\,P(b)$

Qn 2: A & B cond. indep. given C ? ⇒ NO!

Pf: $P(a,b|c) = \dfrac{P(a,b,c)}{P(c)}$

$= \dfrac{P(a)\,P(b)\,P(c|a,b)}{P(c)}$

$\neq P(a|c)\,P(b|c)$ in general

Intuition: A → C ← B

↳ C is a head-to-head node

Conditioned on C, path between A & B become unblocked → A & B are DEPENDENT

Explain Away

[E.g.]

$\begin{cases} R=1 \Rightarrow \text{Rain Last Night} \\ S=1 \Rightarrow \text{Sprinkle On} \\ G=1 \Rightarrow \text{Grass wet} \end{cases}$

$P(R=1) = P(S=1) = 0.9$

S \ R	0	1
0	0.1	0.2
1	0.2	0.8

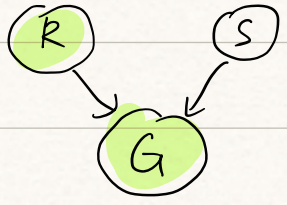
$P(G=1 \mid R=r, S=s)$

① $P(S=0) = 0.1$

② Suppose we observe $G=0$

$P(S=0 \mid G=0) = 0.257\ldots$

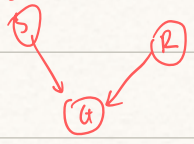
↳ 说明 S和G不独立 (S) → (G)



③ Suppose we observe $G=0$ & $R=0$

$P(S=0 \mid G=R=0) = 0.111\ldots$ ① > 0.1 ② < 0.257

说明 Given G S和R不独立



Explain Away the fact that $S=0$ → make the prob closer to prior 0.1

If $R \perp\!\!\!\perp S \mid G$ (which is not true), then

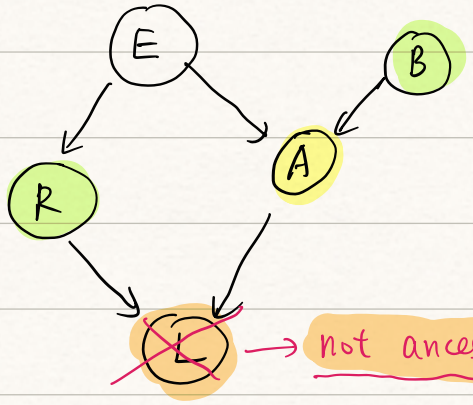
$P(S=0 \mid G=0) = P(S=0 \mid G=0, R=0)$

⇓

But this equality does not hold.

Instead of writing down the joint dist^n & marginalizing to answer Cond. indep. questions, we want a systematic way to answer C. I. question!

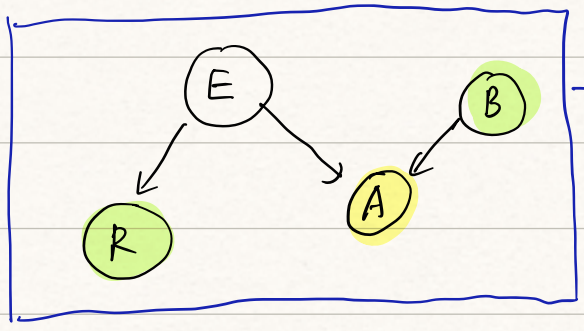
Moralization



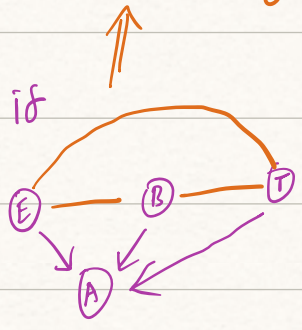
Qn : Are R & B C.I given A ?

Ans: Four steps :

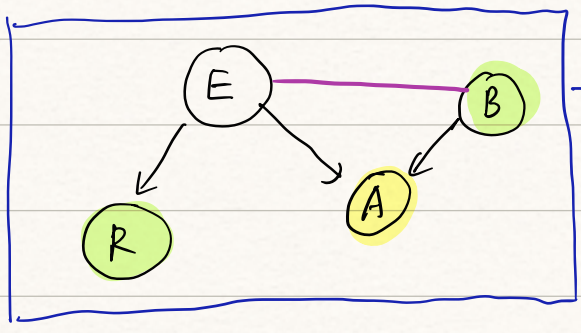
Step 1 : Construct ancestral graph of nodes / variable of interest (R, A, B)



Pair-wise marry !

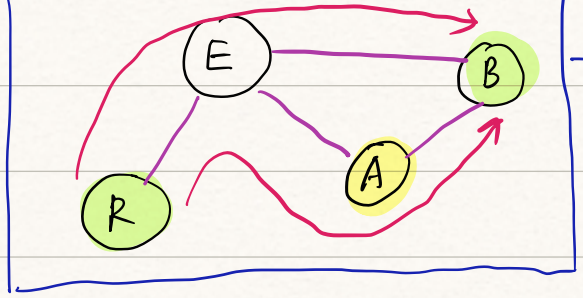


Step 2 : Moralize the ancestral graph by marrying the parents ⇒ only the parents (not all ancestors)



(Takes care of head-to-head explaining away phenomenon)

Step 3: Change arcs to edges



→ ancestral graph

Step 4: Answer !

↓

Is A a seperator of R & B in the undirected graph

Defn: S is a seperator of A and B ⇔ every path from A to B passes through some nodes in S

NO !

⇓

Conclusion: $R \not\perp B \mid A$

⇓

$R \perp B \mid (A, E)$ is TRUE !