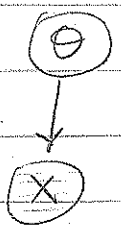
2/6/14

BN: PARAMETER SHARING & STRUCTURE LEARNING

① Meta BN (or Bayesian view of BN)
→ explicitely draw parameters as variables in the BN

e.g #1 $X \sim Cat(\vec{\theta})$
$x \in \{1, \ldots, k\}$

$P(\vec{\theta}) = Dir(\vec{\alpha})$

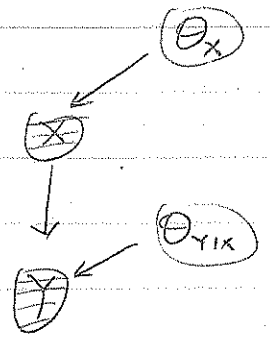


$$P(X=i \mid \vec{\theta}) = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{bmatrix}$$

s.t. $\sum_{i=1}^{k} \theta_i = 1$, $\theta_i \geq 0$ } Simplex

Here $\vec{\alpha}$ is a hyper parameter = fixed.
If we start modeling uncertainty over $\vec{\alpha}$ & start estimating $\vec{\alpha}$ from data, then we can put $\vec{\alpha}$ in the BN.

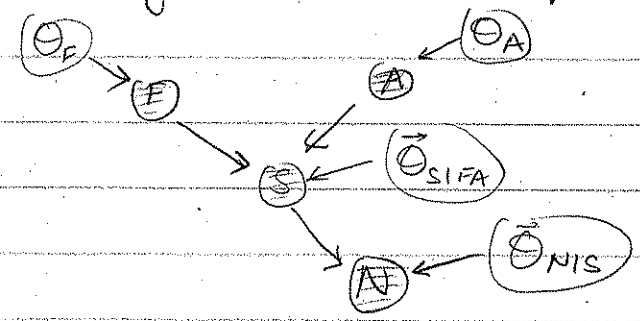
eg #2



① Global Parameter Independence
→ very common assumption
→ all CPT parameters are indep.

$$P(\theta_X, \theta_{Y|X}) = P(\theta_X) P(\theta_{Y|X})$$

Revisiting an old example:



Assumption: $P(\vec{\theta}) = P(\theta_F)\, P(\theta_A)\, P(\theta_{SIFA})\, P(\theta_{NIS})$

$\Rightarrow \log P(\vec{\theta}) = \log P(\theta_F) + \log P(\theta_A) + \log P(\theta_{SIFA}) + \log P(\theta_{NIS})$

fully factorized

$\log P(D|\theta)$

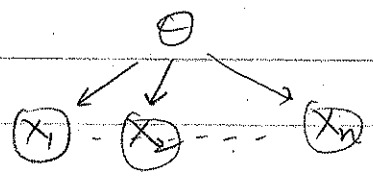
factorizes according to G

These two together lead to $\vec{\theta}_{MAP}$ becoming separate problems at each variable.

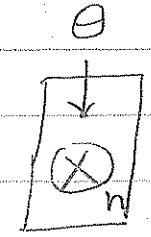
(2) Parameter sharing

→ What if $x_1, \ldots, x_n$ are outcome of n tosses of SAME coin?

Appropriate BN:



(3) Plate Notation



Variable inside plate is repeated n times
Variable outside plate is not (It's shared)

→ There can be nested plates ≡ hierarchical Bayesian models

(4) Structure Learning

Setup:

→ Random Vars: $\vec{X} = \{X_1 \cdots X_n\}$ (all categorical for now)

→ Fully observed dataset of samples from P^*

$$\begin{bmatrix} x_1^{(1)} & \cdots & x_n^{(1)} \\ \vdots & & \\ x_1^{(M)} & & x_n^{(M)} \end{bmatrix}_{M \times n}$$

→ Goal: Estimate G (& associated θ CPTs)

→ Formulation: Score-based approach

$$\hat{G} = \underset{G \in \{\text{DAGs on } n \text{ vars}\}}{\operatorname{argmax}} \; \text{Score}(G : D)$$

problem: # (undirected) graphs on n vars $= 2^{{}^{n}C_2} \approx 2^{O(n^2)}$

" trees (spanning) " " " $= n^{n-2} \approx 2^{O(n \log n)}$

(4.1) What's a good score?

How about the log-likelihood?

Specifically $\text{Score}(G : D) = \log P(D \mid G, \hat{\theta}_{MLE})$

A graph is "good" if it maximizes the chances of observing the dataset I observed (under MLE parameters)

(Very imp; Later we'll change this)

(4.2) e.g. #1

G: (X) (Y)

D:

X	Y
0	0
0	1
0	0
1	1
1	0
0	0

$M \times 2$

$$\text{Score}(G) = \sum_{j=1}^{M} \log P(X = x^{(j)}, Y = y^{(j)} \mid \hat{\theta}^{G}_{MLE}, G)$$

$$= \sum_{j=1}^{M} \left[\log P(X = x^{(j)} \mid \hat{\theta}_{MLE}, G) + \log P(Y = y^{(j)} \mid \hat{\theta}_{MLE}, G) \right]$$

$$= \sum_{x \in \{0,1\}} \text{count}(x) \log P(x \mid \hat{\theta}_{MLE}, G) + (\;)$$

$$= \sum_{x} \text{count}(x) \log \hat{P}(x) + (\;)$$

$\hat{P}$: empirical distribution estimated from D

Why? $\because \hat{\theta}_{MLE}$!

$$= \sum_{x} M \cdot \hat{P}(x) \log \hat{P}(x) + (\cdot)$$

$$\Rightarrow \text{Score}(G) = M \sum_x \hat{P}(x) \log \hat{P}(x) + M \sum_y \hat{P}(y) \log \hat{P}(y)$$

$$= M \cdot \left[\underbrace{-H_{\hat{P}}(X)}_{\text{Entropy of } \hat{P}(X)} - H_{\hat{P}}(Y) \right]$$

(4.3) e.g #2

G: $X \longrightarrow Y$

$$\text{Score}(G) = \sum_{j=1}^{M} \log P(X=x^{(j)}, Y=y^{(j)} \mid \hat{\theta}_{MLE}, G)$$

$$= \sum_{j=1}^{M} \left[\log P(X=x^{(j)} \mid \hat{\theta}_{MLE}, G) + \log P(Y=y^{(j)} \mid X=x^{(j)}, \hat{\theta}_{MLE}, G) \right]$$

$$= -M \cdot H_{\hat{P}}(X) + \sum_x \sum_y \text{count}(x,y) \log P(y|x, \hat{\theta}_{MLE}, G)$$

$$= -M \cdot H_{\hat{P}}(X) + \sum_x \sum_y M \cdot \hat{P}(x,y) \log \hat{P}(y|x)$$

$$= -M H_{\hat{P}}(X) + M \sum_x \sum_y \hat{P}(x,y) \log \frac{\hat{P}(x,y) \cdot \hat{P}(y)}{\hat{P}(x) \hat{P}(y)}$$

$$= -M \cdot H_{\hat{P}}(X) + M \sum_x \sum_y \hat{P}(x,y) \log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)} + M \sum_x \sum_y \hat{P}(x,y) \log \hat{P}(y)$$

$$= M \cdot \left[\underbrace{I_{\hat{P}}(X,Y)}_{\text{Mutual Information}} - H_{\hat{P}}(X) - H_{\hat{P}}(Y) \right]$$

$\equiv$ KL-divergence between $\underbrace{\hat{P}(X,Y)}_{\text{Joint}}$ & $\underbrace{\hat{P}(X)\hat{P}(Y)}_{\text{Factorized}}$ ■

$\equiv H(X) - H(X|Y)$

$\equiv H(Y) - H(Y|X)$

In general

$$\text{Score}(G : D) = M \cdot \left[\sum_i I_{\hat{P}}(X_i, Pa_{X_i}) - \sum_i H_{\hat{P}}(X_i) \right]$$

$\sum_i H_{\hat{P}}(X_i)$: Indep. of G constant!

→ Good news: score seems reasonable; high mutual info between X_i & Pa_{X_i} is good!

→ Unfortunately $I_{\hat{P}}(X, Y) \geq 0$ [see HW0 in ECE4984/5984]

⇒ score is always non-negative

⇒ adding an edge never hurts!

⇒ Problem!