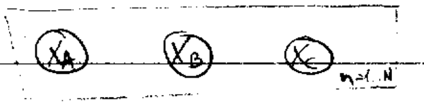
- Assignment 2 due tomorrow
- Assignment 3 on web tomorrow
- Project Info 1-2 people, April 1, May 6, on web tomorrow
- Elimination Algo's, compute marg, conds, but only 1 query node & 1 at a time
- Learn tables, Junction Tree: compute marg/conds but over several nodes & global

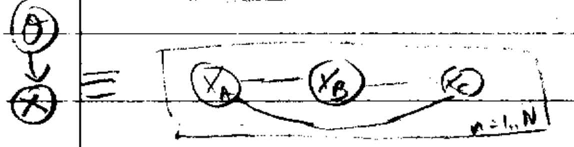
ML for Graphical Models (fully observed)

$G = (U, E)$

nodes vertices — edges



→ easy to get table



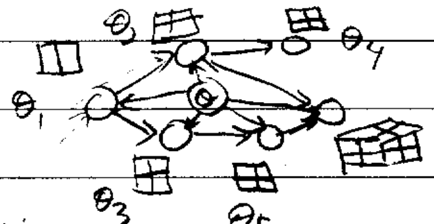
→ easy to get table

for directed graphical models → ML decouples

for undirected → decomposable (normalizer Z is hard to compute)

Directed $\quad p(X) = \prod_{u=1}^{M} p(x_u \mid x_{\pi_u})$ $\quad$ graph, no tables yet

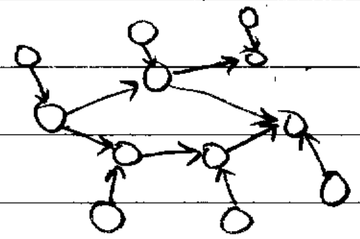
$p(X \mid \theta) = \prod_{u=1}^{M} p(x_u \mid x_{\pi_u}, \theta)$ $\quad$ θ specifies tables



c.p.t's

can be varied indep.

So θ breaks down



$$p(x \mid \theta) = \prod_{u=1}^{M} p(x_u \mid x_{\pi_u}, \theta_u)$$

IID $\quad$ have N observations of X (which is a set of M vars)

i.e. have N patients come in & observe

flu, headache, sinus, fever, etc... for each

get N graphs: $\quad G^{(N)} = (U^{(N)}, E^{(N)})$

Replicates graph

to index nodes in super graph: $X_{u,n}$

$u \in \mathcal{U}$, the set of nodes

$n \in \{1, \dots, N\}$

C is a set of vars, i.e. X_C for $C \subseteq \mathcal{U}$

$X_{C,n}$ is the n'th copy of the set of vars X_C.

$$X_{\mathcal{U}} = \{X_1, \dots, X_M\}$$

our dataset $\mathcal{D} = \{X_{\mathcal{U},1}, \dots, X_{\mathcal{U},N}\}$

← or curly $\mathcal{X}$

Recall <u>ML</u>
$$\theta^* = \underset{\theta}{\operatorname{argmax}} \log p(\mathcal{D}|\theta)$$
$$= \log \prod_n p(X_{\mathcal{U},n} | \theta)$$
$$= \log \prod_n \prod_u p(x_{u,n} | x_{\pi_u,n}, \theta_u) \quad \text{log likelihood}$$
$$= \sum_n \sum_u \log p(x_{u,n} | x_{\pi_u,n}, \theta_u) \equiv \ell(\theta|D)$$

← $\ell(\theta)$ decouples, i.e. $\frac{\partial}{\partial \theta_k}\left(\ell(\theta|D)\right) = \sum_n \frac{\partial}{\partial \theta_k} \log p(x_{k,n} | x_{\pi_k,n}, \theta_k)$

← only consider single θ_k at a time!

i.e. estimate each cpdf at a time.

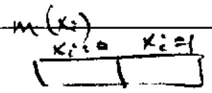
i.e. only look at a few vars at a time $p(x_i|x_{\pi_i})$

graphical models ⟹ memory efficiency
⟹ <u>estimation efficiency</u>

<u>Estimating CPDFs</u> ... first some notation...

$$\delta(x_i, x_j) = \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{otherwise} \end{cases}$$

$m(x_i)$



$m(x_i) = \sum_n \delta(x_i, x_{i,n})$ ← # of data points where x_i assumes given val

$m(x_{\mathcal{U}}) = \sum_n \delta(x_{\mathcal{U}}, x_{\mathcal{U},n})$ ← # of data points where X assumes value.

$m(x_{\mathcal{U}})$

Sum over all X in $\mho$ except X in C

$$m(X_C) = \sum_{X_{\mho \setminus C}} m(X_\mho) \quad \leftarrow$$ # of times on a subset of vars

i.e. $m(x_1, x_2) = \sum_{x_3} m(x_1, x_2, x_3)$

$m(x_1) = \sum_{x_2} \sum_{x_3} m(x_1, x_2, x_3)$

$\sum_{x_1} m(x_1) = N$ (total # of data)

define: $X_{\phi_i} := \{X_i, X_{\pi_i}\}$ node i & its parents (cpdt)

$\therefore \quad m(X_{\phi_u}) := \sum_{X_{\mho \setminus \phi_u}} m(X_u)$

forming a table of counts over X_u & its parents.

our cond. prob. table

θ_u is a parameter vector. If θ_u is over X_u by itself, it is a 1d table

$X_u=0$	$X_u=1$
0.3	0.7

if $X_u | X_{\pi_u}$

	$X_{\pi_u}=0$	$X_{\pi_u}=1$
$X_u=0$		
$X_u=1$		

$\therefore$ θ_u is a table over X_{ϕ_u} $\qquad \phi_u = u, \pi_u$

i.e. $\theta_u(X_{\phi_u})$ like we had for coin flips

$\theta(x=0) = 0.3 \quad \theta(x=1) = 0.7$ (we called it θ_1 & θ_2 ...)

naturally $\sum_{x_v} \theta_v(X_{\phi_v}) = \sum_{x_v} \theta_v(x_v, X_{\pi_v}) = 1$

i.e. $p(x_v | X_{\pi_v}, \theta_v) = \theta_v(X_{\phi_v})$

$\therefore \quad p(x_\mho | \theta) = \prod_v p(x_v | X_{\pi_v}, \theta_v) = \prod_v \theta_v(X_{\phi_v})$

$p(x_{\mho,n} | \theta) = \prod_{x_\mho} p(x_\mho | \theta)^{\delta(x_\mho, x_{\mho,n})}$

over all possible $\mho$
all settings of the nodes $X_1, \dots, X_M$

<u>log-likelihood</u> $\log p(D|\theta) = \log \prod_n p(x_{\mho,n} | \theta)$

$= \sum_n \log p(x_{\mho,n} | \theta)$

$= \sum_n \log \prod_{x_\mho} p(x_\mho | \theta)^{\delta(x_\mho, x_{\mho,n})}$

$$= \sum_n \sum_{x_\mathcal{V}} \delta(x_\mathcal{V}, x_{\mathcal{V},n}) \log p(x_\mathcal{V}|\theta)$$

$$= \sum_{x_\mathcal{V}} m(x_\mathcal{V}) \log p(x_\mathcal{V}|\theta)$$

$$= \sum_{x_1} \cdots \sum_{x_M} m(x_\mathcal{V}) \log \prod_v p(x_v|x_{\pi_v},\theta)$$

$$= \sum_{x_\mathcal{V}} m(x_\mathcal{V}) \log \prod_v \theta_v(x_{\phi_v})$$

$$= \sum_{x_\mathcal{V}} m(x_\mathcal{V}) \sum_v \log \theta_v(x_{\phi_v})$$

$$= \sum_v \sum_{x_\mathcal{V}} m(x_\mathcal{V}) \log \theta_v(x_{\phi_v})$$

sufficient stats e-family property

$$= \sum_v \sum_{x_{\phi_v}} \underbrace{\sum_{x_{\mathcal{V}\setminus\phi_v}} m(x_\mathcal{V})}_{m(x_{\phi_v})} \log \theta_v(x_{\phi_v})$$

$$= \sum_v \sum_{x_v} \sum_{x_{\pi_v}} m(x_v, x_{\pi_v}) \log \theta_v(x_v, x_{\pi_v})$$

(for each θ_v, it decouples, once again)

add Lagrangians $\ell(\theta|D) - \sum_v \sum_{x_{\pi_v}} \lambda_{\pi_v}\left(\sum_{x_v} \theta_v(x_v, x_{\pi_v}) - 1\right)$

since $\sum_v \theta_v(x_v, x_{\pi_v}) = 1$

$$\frac{\partial}{\partial \theta_{\bar v}(x_{\bar v}, x_{\pi_{\bar v}})} \quad \frac{m(x_{\bar v}, x_{\pi_{\bar v}})}{\theta_{\bar v}(x_{\bar v}, x_{\pi_{\bar v}})} - \lambda_{\pi_{\bar v}} = 0$$

$$\theta_{\bar v}(x_{\bar v}, x_{\pi_{\bar v}}) = \frac{m(x_{\bar v}, x_{\pi_{\bar v}})}{\lambda_{\pi_{\bar v}}}$$

$$\sum_v \theta_v(x_v, x_{\pi_{\bar v}}) = 1$$

implicitly $\Longrightarrow$ $\lambda_{\pi_{\bar v}} = \sum_v m(x_{\bar v}, x_{\pi_{\bar v}}) = m(x_{\pi_{\bar v}})$

$$\hat\theta_v(x_v, x_{\pi_v}) = \frac{m(x_v, x_{\pi_v})}{m(x_{\pi_v})}$$

Simple algorithmically

example

$p(\text{fever}|\text{flu})$ $X_2|X_1$

$m(x_v, x_{\pi_v}) =$

$m(x_{\pi_v}) =$

	patient signs
1	0111
2	00111
3	11..01
4	11100
5	00111
6	
7	
8	
9	

Undirected Models factorize without being conditionals, but $1/Z$ couples them. not decompos.

$$D = \{x_{\mathcal{V},1}, \dots, x_{\mathcal{V},N}\}$$

$$p(x_\mathcal{V}|\theta) = \frac{1}{Z} \prod_C \psi_C(x_C) \qquad Z = \sum_{x_\mathcal{V}} \prod_C \psi_C(x_C)$$

(recall previous $m()$ notation...)

$$p(x_{\mathcal{V},n}|\theta) = \prod_{x_\mathcal{V}} p(x_\mathcal{V}|\theta)^{\delta(x_\mathcal{V}, x_{\mathcal{V},n})}$$

$$\ell(\Theta|D) = \log p(D|\Theta) = \log \prod_n p(x_{\mho,n}|\theta)$$
$$= \log \prod_n \prod_{x_\mho} p(x_\mho|\theta)^{\delta(x_\mho, x_{\mho,n})}$$
$$= \sum_n \sum_{x_\mho} \delta(x_\mho, x_{\mho,n}) \log p(x_\mho|\theta)$$
$$= \sum_{x_\mho} m(x_\mho) \log p(x_\mho|\theta)$$
$$= \sum_{x_\mho} m(x_\mho) \sum_C \log \psi_C(x_C) - \sum_{x_\mho} m(x_\mho) \log Z$$
$$= \sum_C \sum_{x_C} m(x_C) \log \psi_C(x_C) - N \log Z$$

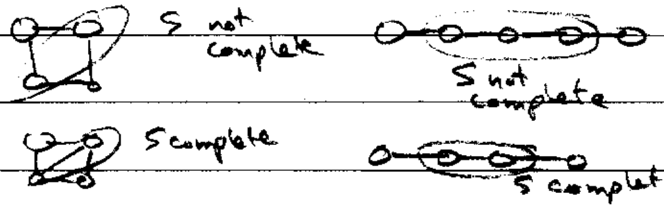
use max likelihood to set Θ or the tables $\psi_C(x_C)$...

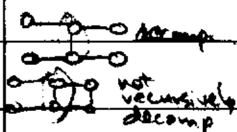
two cases: decomposable & non-decomposable.

<u>Decomposable</u> - divide graph into disjoint subsets A, B, S
- recursively, must have below for graph & sub-graphs

If S separates A & B, S must be complete
$\parallel$
all nodes fully intercon.
$\parallel$
clique

decomp. must recursive on subgraphs





need to divide by middle for 3 nodes

If decomposable: by inspection no algo!

1) $\psi_C(x_C) \leftarrow \tilde{p}(x_C) = \frac{1}{N} m(x_C)$

2) $\psi_C(x_C) \leftarrow \dfrac{\psi_C(x_C)}{p(X_C \cap X_D)}$

$\psi_C(x_D) \leftarrow \psi_D(x_D)$

for all non-empty intersections between pairs divide 1 of the intersecting cliques combo by the empirical marginal of intersection

do example

x_1 — x_2 — x_3

count
1 1 0
0 0 1
1 1 1
1 0 1
0 0 0
0 1 0
0 0 1

$p(x_1, x_2) =$

$p(x_2, x_3) =$

$p(x_2) =$

$\frac{p(x_1,x_2)}{p(x_2)} p(x_2,x_3)$

$\frac{p(x_2,x_3)}{p(x_2)} p(x_1,x_2)$

more general
Non-decomposable (or otherwise) $\Rightarrow$ I.P.F. algorithm

if decomposable, converges in finite steps.

Iterative Prop. Fitting — solution, properties, proof

I.P.F.:
- guess values for $\psi_c(x_c)$
- lock all $\psi_c(x_c)$
- for $c = 1 \ldots C$ $\quad \psi_c(x_c)^{(t+1)} \leftarrow \psi_c(x_c)^{(t)} \dfrac{\hat{p}(x_c)}{p^{(t)}(x_c)}$
  - where $\hat{p}(x_c) = \frac{1}{N} m(x_c)$
  - $p(x_c)^{(t)}$ is computed from current
    $$p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)^{(t)}$$
    $$p(x_c) = \sum_{x \setminus x_c} p(x)$$

for decomposable graphs, IPF converges in finite iters

Note: $Z = \sum_x \prod_c \psi_c(x_c)^{(t)}$ $\quad \therefore\ Z$ can change with (t)

NOPE

$$p^{(t+1)}(x_c) = \sum_{x \supset x_c} p^{(t+1)}(x)$$

only ψ_c changed from t to $t+1$

$$= \sum_{x \supset x_c} \frac{1}{Z^{(t+1)}} \prod_D \psi_D^{(t+1)}(x_D)$$

$$= \frac{1}{Z^{(t+1)}} \sum_{x \supset x_c} \psi_c^{(t+1)}(x_c) \prod_{D \neq c} \psi_D^{(t+1)}(x_D)$$

$$= \frac{1}{Z^{(t+1)}} \sum_{x \supset x_c} \psi_c^{(t)}(x_c) \frac{\hat{p}(x_c)}{p^{(t)}(x_c)} \prod_{D \neq c} \psi_D^{(t)}(x_D)$$

$$= \frac{Z^{t}}{Z^{t+1}} \frac{\hat{p}(x_c)}{p^{(t)}(x_c)} \sum_{x \supset x_c} \frac{1}{Z^{t}} \prod_D \psi_D^{(t)}(x_D)$$

$$= \frac{Z^{t}}{Z^{t+1}} \frac{\hat{p}(x_c)}{p^{(t)}(x_c)} \sum_{x \supset x_c} p^{(t)}(x)$$

$$= \frac{Z^{(t)}}{Z^{(t+1)}} \frac{\hat{p}(x_c)}{p^{(t)}(x_c)} p^{(t)}(x_c)$$

$$p^{(t+1)}(x_c) = \frac{Z^{(t)}}{Z^{(t+1)}} \hat{p}(x_c)$$

sum both sides over x_c

$$\sum_{x_c} p^{(t+1)}(x_c) = \sum_{x_c} \frac{Z^{(t)}}{Z^{(t+1)}} \hat{p}(x_c)$$

$$1 = \frac{Z^{(t)}}{Z^{(t+1)}}$$

$\therefore\ Z^{(t+1)} = Z^{(t)} =$ constant with IPF iters.