ML for Graphical Models (fully observed) $G = (V, E)$

- ML - easy to get table
- for directed graphical models $\rightarrow$ ML decouples
- for undirected $\rightarrow$ decomposable (normalizer $Z$ is hard to compute)

- Directed $p(x) = \prod_{i=1}^{n} p(x_i | x_{pa_i})$ graph, no tables yet
  $p(x|\theta) = \prod_{i=1}^{n} p(x_i | x_{pa_i}, \theta)$ $\theta$ specifies tables

- CPT's can be varied indep. so $\theta$ breaks down

- IID have $N$ observations of $x$ (which is a set of MVs)
  - i.e. have $N$ patients, some in & observe
    - fly, headache, sinus, fever, etc... for each
  - get $N$ graphs: $G^{(N)} = (U^{(N)}, E^{(N)})$

- Replicates graph...
to index nodes in super graph: $X_{\text{m,n}}$

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

$1 \leq i \leq N$

$C$ is a set of vars, i.e., $X_{c}$ for $c \in \mathcal{C}$

$X_{\text{m,n}}$ is the $i^{th}$ copy of the set of vars $X_{c}$

$X_{\mathcal{C}} = \{X_{1,\ldots,N}\}$

our dataset $D = \{X_{1,\ldots,N}\}$

or early $X$

Recall $\underbrace{\theta^* = \arg \max}_{\theta} \log p(D|\theta)}$

$= \log \frac{1}{n} \prod_{i=1}^{n} p(x_{i}|\theta)$

$= \log \frac{1}{n} \sum_{i=1}^{n} p(x_{i}|X_{\text{m,n}}, \theta_{u})$

$= \frac{1}{n} \sum_{i=1}^{n} \log p(x_{i}|X_{\text{m,n}}, \theta_{u}) \equiv l(\theta|D)$

(1) $\theta$ decouples, i.e., $\frac{2}{n} l(\theta|D) = \frac{2}{n} \sum_{i=1}^{n} \log p(x_{i}|X_{\text{m,n}}, \theta_{u})$

i.e. only consider single $\theta_{u}$ at a time!

i.e. estimate each cpdf at a time.

i.e. only look at a few vars at a time. $p(x_{i}|x_{\text{m,n}})$

graphical models $\Rightarrow$ memory efficiency

$\Rightarrow$ estimation efficiency

Estimating CPDFs... First some notation...

$\delta(x_{i}, x_{j}) = 1$ if $x_{i} = x_{j}$

$\delta(x_{i}, x_{j}) = 0$ otherwise

$m(x_{i}) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_{i}, x_{j})$

$m(x_{i}) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_{i}, x_{j})$

$m(x_{i}) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_{i}, x_{j})$

$m(x_{i}) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_{i}, x_{j})$
\[
m(X_c) = \sum_{x \in C} m(x) \quad \text{at times on a subset of vars}
\]
\[
i.e. m(x,i) = \sum_{x_i} m(x,x_i,\epsilon)
\]
\[
m(x) = \sum_{x} m(x,\epsilon,\epsilon)
\]
\[
\sum_{x_i} m(x_i) = N \quad \text{(total n of data)}
\]

Define: \(X_{\Phi} = \{x_i, x_{\Phi}\} \quad \text{node i & its parents (cpdt)}\)

\[
m(X_{\Phi}) = \sum_{x \cup \Phi} m(x)
\]

\[
\text{forming a table of counts over } X_{\Phi} \text{ & its parents}
\]

\[
\Theta_{\Phi} \text{ is a parameter vector. If } \Theta \text{ is over } X_{\Phi} \text{ by itself, it is a (d) table}
\]
\[
\begin{array}{c|c|c|c|c|c}
X_{\Phi} & X_{\Phi} & X_{\Phi} & X_{\Phi} & X_{\Phi} & X_{\Phi} \\
\hline
x_{\Phi=0} & x_{\Phi=1} & x_{\Phi=0} & x_{\Phi=1} & x_{\Phi=0} & x_{\Phi=1} \\
\hline
\end{array}
\]

\[
\Theta_{\Phi} \text{ is a table over } X_{\Phi} \quad x_{\Phi=0}, x_{\Phi=1}
\]

\[
\text{i.e. } \Theta_{\Phi}(X_{\Phi}) \quad \text{like we had for coin flips}
\]

\[
\Theta(x_{\Phi=0}) = \Theta(x_{\Phi=1}) = \frac{1}{2} \quad (\text{we call it } \Theta_{1} \text{ & } \Theta_{2} \ldots)
\]

\[
\sum_{x_{\Phi}} \Theta_{\Phi}(X_{\Phi}) = \sum_{x_{\Phi}} \Theta_{\Phi}(x_{\Phi}, x_{\Phi}) = 1
\]

\[
\text{i.e. } p(x_{\Phi}|x_{\Phi}, \Theta_{\Phi}) = \Theta_{\Phi}(X_{\Phi})
\]

\[
\Sigma p(x_{\Phi}|\Theta) = \int p(x_{\Phi}|x_{\Phi}, \Theta_{\Phi}) = \int \Theta_{\Phi}(X_{\Phi})
\]

\[
p(x_{\Phi}, x_{\phi}|\Theta) = \int p(x_{\Phi}, x_{\phi}|\Theta_{\Phi})
\]

\[\text{log-likelihood } \log p(D|\Theta) = \log \int p(x_{\Phi}|\Theta) \Sigma(x_{\Phi}, x_{\phi})
\]

\[= \sum_{x_{\Phi}} \log p(x_{\Phi}|\Theta) \Sigma(x_{\Phi}, x_{\phi})
\]

\[= \frac{1}{2} \sum_{x_{\Phi}} \log \int p(x_{\Phi}|\Theta) \Sigma(x_{\Phi}, x_{\phi})
\]
\[
\begin{align*}
\mathcal{D} &= \{x_{w1}, x_{w2}, x_{w3}\} \\
\rho(x_{w1}|\theta) &= \frac{1}{2} \prod_x \psi(x) \\
\rho(x_{w1}, x_{w2}|\theta) &= \frac{1}{2} \prod_x \psi(x) \\
\rho(x_{w1}, x_{w2}, x_{w3}|\theta) &= \prod_x \rho(x|\theta) \\
\end{align*}
\]
\[ L(\theta|D) = \log p(D|\theta) = \log \prod_{x,y} p(x,y|\theta) \]
\[ = \log \prod_{x} \prod_{y \neq x} p(x,y|\theta) \]
\[ = \sum_{x} \sum_{y \neq x} S(x,y, x \cup y) \log p(x,y|\theta) \]
\[ = \sum_{x} m(x) \log \psi(x) - \sum_{x} m(x) \log 2 \]
\[ = \sum_{x} m(x) \log \psi(x) - N \log 2 \]

Use max likelihood to set \( \theta \) or the tables \( \psi(x) \).

Two cases: decomposable \& non-decomposable.

Decomposable:
- Divide graph into disjoint subsets \( A, B, S \)
- Recursively, must have below for graph \& subgraphs
- \( S \) separates \( A \& B \), \( S \) must be complete
- \( \psi(x) \) complete
- \( \psi(x) \) not complete

If decomposable:
\[ \psi(x) \leftarrow \rho(x) = \frac{1}{N} m(x) \]
\[ \rho(x) \leftarrow \frac{\psi(x)}{\rho(x)} \] for all non-empty intersections

More general:
- Non-decomposable (or otherwise) \( \Rightarrow \) IPF algorithm
- if decomposable, converges in finite steps
Iterative Prop Fitting: Solution properties proof

1. If

\[ \text{IPF:} \]

2. Guess values for \( \Psi_c(x_c) \)

3. Lock all \( \Psi_c(x_c) \)

4. For \( c = 1, \ldots, C \)

\[ \Psi_c(x_c) \rightarrow \Psi_c(x_c) \]

where \( \delta \Psi_c(x_c) = \frac{1}{2} \nabla \Psi_c(x_c) \)

\[ \rho(x_c) \]

is computed from current

\[ \rho(x) = \frac{1}{2} \sum \Psi_c(x_c) \]

\[ \rho(x_c) = \sum_{x_c} \rho(x) \]

5. For decomposable graphs, IPF converges in finite steps.

Note: \( z = \frac{1}{x_c} \sum \Psi_c(x_c) \)

so \( z \) can change with \( t_i \)

\[ \rho^{(t)}(x_c) = \frac{1}{x_c} \rho^{(t)}(x) \]

\[ \rho^{(t+1)}(x_c) = \frac{1}{x_c} \rho^{(t+1)}(x) \]

\[ \rho^{(t+1)}(x_c) = \frac{1}{x_c} \rho^{(t+1)}(x) \]

\[ \sum \text{both sides over } x_c \]

\[ \sum_{x_c} \rho^{(t+1)}(x_c) = \sum_{x_c} \rho^{(t+1)}(x_c) \]

\[ \sum \sum = \frac{1}{2} \]

\[ \rho^{(t+1)} = \frac{1}{2} = \text{constant with IPF steps.} \]