<u>Review</u>

* Inference in BNs

evidence nodes E

query node Q

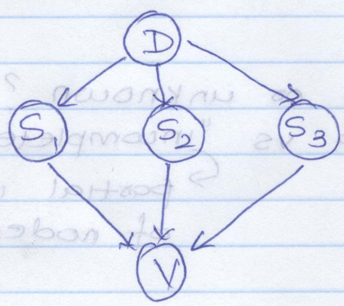
How to compute $P(Q|E)$?

* Polytrees

- singly connected networks
- polynomial time inference.

* Loopy BNs

One approach: node clustering



disease $D \in \{0,1\}$

symptoms $S_i \in \{0,1\}$

visit $V \in \{0,1\}$

Merge nodes to form polytree

Merge S_1, S_2, S_3 into "mega-node" S — take on 2^3 values

Merge CPTs into "mega-CPTs".

S_1	S_2	S_3	S
0	0	0	0
0	0	1	1
⋮	⋮	⋮	⋮
1	1	1	7

S	$P(S|D)$
0	$P(S_1=0|D)\,P(S_2=0|D)\,P(S_3=0|D)$
1	$P(S_1=0|D)\,P(S_2=0|D)\,P(S_3=1|D)$
⋮	
7	$P(S_1=1|D)\,P(S_2=1|D)\,P(S_3=1|D)$

* ~~Polynomial~~ Polytree algorithm linear in size of CPTs, but CPTs growing exponentially with clustering.

* How to choose optimal clustering? computationally hard problem.

$\boxed{\text{Learning}}$

* BN = DAG + CPTs not always available from experts

How to learn from examples?

* Issues
  - structure (DAG) - known or unknown?
  - evidence: complete data vs "incomplete" data
    ↳ partial instantiation of nodes in BN
  - optimization:
    combinatorial v/s ~~iter~~ continuous
    (eg: learning DAG) (eg: learning CPTs)
  - algorithms - non-iterative v/s iterative
    (loop many times over data set)
  - solution: local v/s global optima in model estimation.

* Maximum likelihood estimation (ML)
  - simplest form of learning
  - choose ("estimate") model (DAG + CPTs) to maximize $\underbrace{P(\text{observed data} \mid \text{model})}_{\text{"likelihood"}}$

$\boxed{\text{Ex : biased coin}}$

$X \in \{heads, tails\}$

$P(X = heads) = p$ (X) trivial BN
$P(X = tails) = 1-p$

* How to estimate p from T examples (i.e, results of T coin tosses)?

* I.I.D. assumption
Samples are independently, identically distributed according to $P(X)$.
$\{X^{(1)}, X^{(2)}, \ldots, X^{(T)}\}$ T samples

* Probability of I.I.D. data:

$$P(data) = P(X = x^{(1)}, X = x^{(2)}, X = x^{(3)}, \ldots, X = x^{(T)})$$
$$= P(X = x^{(1)})\, P(X = x^{(2)}) \ldots P(X = x^{(t)})$$ b/c coin tosses independant
$$= \prod_{t=1}^{T} P(X = x^{(t)})$$

* Log-probabilities $\mathcal{L}$

$$\mathcal{L} = \log P(data)$$
$$= \log \prod_{t=1}^{T} P(X = x^{(t)})$$
$$= \sum_{t=1}^{T} \log P(X = x^{(t)})$$

log-likelihood

* Notation :

Let N_H = count (X = heads)

N_T = count (X = tails)

Clearly : $N_H + N_T = T$ (total # samples)

In terms of counts :

$$\mathcal{L} = N_H \log p + N_T \log(1-p)$$

* Maximum likelihood estimation :

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{N_H}{p} + \frac{N_T}{1-p}(-1) = 0 \quad \text{at maximum}$$

$$N_H(1-p) - N_T p = 0$$

$$N_H - p(N_H + N_T) = 0$$

$$p = \frac{N_H}{N_H + N_T} = \frac{N_H}{T}$$

- intuitively, ML estimate is relative empirical frequency of heads.

$\boxed{\text{Discrete BNs with "complete data"}}$

* Given : fixed DAG over discrete nodes $\{X_1, X_2, \ldots, X_n\}$

* CPTs enumerate $P(X_i = x \mid pa(X_i) = \pi)$ as lookup tables

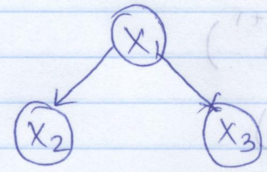
parents of X_i

configuration of parents

* Data is T complete instantiations of all nodes in BN.

$$\left\{ \left(X_1^{(t)}, X_2^{(t)}, X_3^{(t)}, \ldots, X_n^{(t)} \right) \right\}_{t=1}^{T}$$

Ex:



$X_i \in \{0,1\}$

$n = 3$

Data

t	X_1	X_2	X_3
1	1	0	1
2	1	1	1
3	0	1	1
4	0	1	0
⋮			
T	0	1	0

* Each n-tuple of values is called an "example"
Goal: learn from examples
estimate CPTs $P(X_i = x \mid pa_i = \pi)$
that maximize $\underbrace{\text{probability}}_{\text{likelihood}}$ of data

* I.I.D. assumption
Examples are independantly, identically distributed according to $P(X_1, X_2, \ldots, X_n)$

* Probability of data ← Joint prob of t^{th} example

$$P(\text{data}) = \prod_{t=1}^{T} P(X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \ldots, X_n = x_n^{(t)})$$

* Work out t^{th} term: Product Rule

$$P(X_1 = x_1^{(t)}, \ldots, X_n = x_n^{(t)}) = P(X_1 = x_1^{(t)})\, P(X_2 = x_2^{(t)} \mid X_1 = x_1^{(t)}) \ldots\ P(X_n = x_n^{(t)} \mid X_1 = x_1^{(t)} \ldots X_{n-1}^{(t)} = x_{n-1}^{(t)})$$

$$= \prod_{i=1}^{n} P(X_i = x_i^{(t)} \mid X_1 = x_1^{(t)}, \ldots, X_{i-1} = x_{i-1}^{(t)})$$

conditional dependence in BN

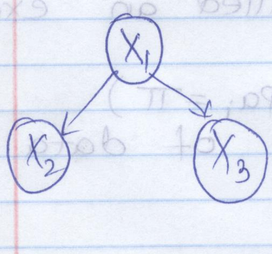
$$= \prod_{i=1}^{n} P(X_i = x_i^{(t)} \mid pa(X_i) = pa_i^{(t)})$$

* Log-Likelihood

$$\mathcal{L} = \log P(\text{data})$$
$$= \log \prod_{t=1}^{T} P(X_1^{(t)}, X_2^{(t)}, \ldots, X_n^{(t)})$$
$$= \log \prod_{t=1}^{T} \prod_{i=1}^{n} P(x_i^{(t)} \mid pa_i^{(t)})$$
$$= \sum_{t=1}^{T} \sum_{i=1}^{n} \log P(x_i^{(t)} \mid pa_i^{(t)})$$

$$\boxed{\mathcal{L} = \sum_{i=1}^{n} \sum_{t=1}^{T} \log P(X_i = x_i^{(t)} \mid pa(X_i) = pa_i^{(t)})}$$

swapped order of sums



* Let count$(X_i = x, pa_i = \pi)$ denote #examples for which $X_i = x$ and $pa_i = \pi$

t	X_1	X_2	X_3
1	0	0	1
2	1	1	0
3	1	0	1
4	1	0	0
5	0		

count$(X_3 = 1, X_1 = 0) = 2$

count$(X_1 = 1) = 3$

count$(X_2 = 1, X_1 = 1) = 1$

* Log-likelihood

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{x} \sum_{\pi} \underbrace{\text{count}(X_i = x, pa_i = \pi)}_{\text{completely determined by data}} \log \overbrace{P(X_i = x \mid pa = \pi)}^{\text{CPTs that we can choose}}$$

x: values of X_i

π: values of $pa(X_i)$

* ML estimation

How to choose $P(X_i = x \mid pa_i = \pi)$ to maximize $\mathcal{L}(\text{data})$?

* ML solution (w/o proof)

$$P_{ML}(X_i = x \mid pa_i = \pi) = \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x'_i, pa_i = \pi)}$$

$$= \frac{\text{count}(X_i = x, pa_i = \pi)}{\text{count}(pa_i = \pi)}$$

* Properties of MLE

- Asymptotically correct: $P_{ML}(X_1, X_2, \ldots, X_n) \to P(X_1, X_2, \ldots, X_n)$ as $T \to \infty$

  with enough data, you recover the true model

- Problematic for sparse data:

  $P_{ML}(X_i = x \mid pa_i = \pi) = 0$ if $\text{count}(X_i = x, pa_i = \pi) = 0$

  $P_{ML}(X_i = x \mid pa_i = \pi)$ undefined if $\text{count}(pa_i = \pi) = 0$