CSE 150 – 1/27/10

Review

* Inference in BNs
  evidence node E
  query node Q
  How to compute $P(Q|E)$?

* Polytrees
  - singly connected networks
  - polynomial time inference

* Loopy BNs
  Exact inference: node clustering
  Approximate inference: $\underbrace{\text{stochastic simulation}}_{\text{[covered later]}}$

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 (e.g. learning DAG) vs. continuous (e.g. learning CPTs)
  - algorithms: non-iterative vs. iterative (loop over data many times)
  - solution: local vs. global optimum.

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

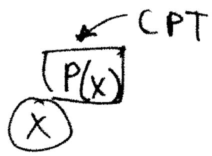
$\boxed{\text{Ex: biased coin}}$

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

$P(X = \text{heads}) = p$

$P(X = \text{tails}) = 1 - p$

Trivial BN



* How to estimate p from observed samples (results of T coin tosses)?

* IID assumptions

  Samples are independently, identically distribute to $P(X)$.

  $\longrightarrow \{x^{(1)}, x^{(2)}, \ldots, x^{(T)}\}$ T samples

* Probability of IID data:

$$P(\text{data}) = P(X = x^{(1)})\, P(X = x^{(2)}) \ldots P(X = x^{(T)})$$
$$= \prod_{t=1}^{T} P(X = x^{(t)})$$

* Log-probability $\mathcal{L}$

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

"log-likelihood"

Let $N_H = \text{count}(X = \text{heads})$
Let $N_T = \text{count}(X = \text{tails})$

Clearly: $N_H + N_T = T$ ← total samples

In terms of counts:

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

* Maximum Likelihood estimation

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{N_H}{p} + \frac{N_T}{1-p} \cdot (-1) = 0$$

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

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

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

intuitively, maximum likelihood estimate of $p = P(X = \text{heads})$ is relative frequency in observed coin ~~tosses~~ tosses.

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

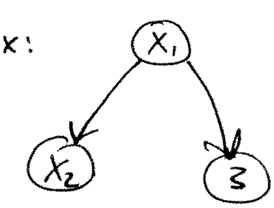
* Given: fixed DAG over discrete nodes $\{X_1, X_2, \dots, X_n\}$

* CPTs enumerate $P(X_i = x_i \mid pa(x_i) = \pi)$ as look up tables

$pa(x_i)$ ← parents of X_i; π ← parent configuration

* Data is T complete instantiations of nodes in BN
$\{x_1^{(t)}, x_2^{(t)}, \dots, x_n^{(t)}\}_{t=1}^{T}$

Ex:



$X_i \in \{0,1\}$
$n = 3$

Data

t^{th} sample	X_1	X_2	X_3
1	0	0	0
2	1	1	0
3	1	1	0
⋮			
T	0	1	3

* Each n-tuple of values is called an "example".

Goal: ~~lean~~ from examples;

estimate CPTs $P(X_i = x \mid pa_i = \pi)$

that maximize $\underbrace{\text{probability}}_{\text{likelihood}}$ of data set

* I.I.D. Assumption

Samples are independently, identically distributed according to $P(X_1, X_2, \ldots X_n)$.

* Probability of I.I.D. set:

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

probability of t^{th} exampe

* Work out t^{th} term:

$$P(X_1 = x_1^{(t)}, \ldots, X_n = x_n) = P(X_1 = x_1^{(t)})\, P(X_2 = x_2^{(t)} \mid P(X_1 = x_1^{(t)}) \times \ldots$$

product rule

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

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

conditional dependence

* Log-likelihood $\mathcal{L}$

$$\mathcal{L} = \log P(\text{data})$$

$$= \log \prod_{t=1}^{T} P(X_1^{(t)}, X_2^{(t)}, \ldots, X_n^{(t)})$$

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

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

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

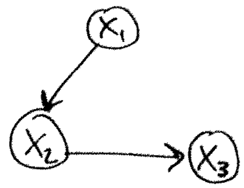
swap order of sums

Data set

t	X_1	X_2	X_3	$\cdots X_n$
1	0	0	1	1
2	1	1	0	0
3	0	1	0	

* Let $\text{count}(X_i = x, pa_i = \pi)$ denote examples for which $X_i = x$ and $pa(X_i) = \pi$.

Ex:



$n = 3$
$T = 3$

$\text{count}(X_3 = 0 \mid X_2 = 1) = 2$
$\text{count}(X_3 = 1 \mid X_2 = 0) = 1$

* Log-likelihood:

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{x} \sum_{\pi} \text{count}(X_i = x, pa_i = \pi) \log P(X_i = x_i \mid pa_i = \pi)$$

values of X parent configuration

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

* ML Solution (without proof):

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

Equivalently:

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

* Properties of MLE
  - Asymptotically correct: $P_{ML}(X_1, X_2, \ldots, X_n) \to P(X_1, X_2, \ldots, X_n)$ as $T \to \infty$
  - Problematic for sparse data:
    $P_{ML}(X_i = x \mid pa_i = \pi) = 0$ if $\text{count}(X_i = x \mid pa_i = \pi) = 0$
    $P_{ML}(X_i = x \mid pa_i = \pi)$ undefined if $\text{count}(pa_i = \pi) = 0$

- Other useful notation:

  Indicator function:

$$I(x, x') = \begin{cases} 0 & \text{if } x \neq x' \\ 1 & \text{if } x = x' \end{cases}$$

$$\text{count}(pa_i = \pi) = \sum_{t=1}^{T} I(pa_i^{(t)}, \pi)$$

$$\text{count}(X_i = x,\ pa_i = \pi) = \sum_{t=1}^{T} I(pa_i^{(t)}, \pi)\, I(X_i^{(t)}, x)$$