Review

- Inference in BNs

  evidence nodes E

  query node Q

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

- Polytrees
  - singly connected networks
  - polynomial time inference

- Loopy BNs

  Exact inference: node clustering

Learning

↳ as a form of uncertain reasoning from observations

Agents can handle uncertainty using probability, but must learn probabilistic theories of the world from experience

- BN = DAG + CPTs not always available from experts

How to learn from data (examples / observations)?

- Issues
  - structure (DAG): known or unknown?
  - evidence: complete data vs. "incomplete" data
    ↳ partial instantiation of nodes in BN
  - optimization:

    combinatorial (eg: learning DAG) vs continuous (e.g. learning CPTs)
  - algorithms: non-iterative vs. iterative (loop many times over data set)
  - solution: local vs. global optima in model estimation

- Maximum likelihood estimation (ML, MLE)
  - simplest form of learning
  - choose ("estimate") model (DAG + CPTs) to maximize $\underbrace{P(\text{observed data} \mid \text{model})}_{\text{"likelihood"}}$,
  - we'll focus on parameter learning: finding numerical parameters for a probabilistic model whose structure is fixed

Ex: biased coin

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

trivial BN

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

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

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

  Intuition: $p = \dfrac{\#\text{ heads}}{\#\text{ flips}_{\rightarrow T}}$

- 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:

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

$$= P(X = x^{(1)})\, P(X = x^{(2)}) \cdots P(X = x^{(T)}) \quad \text{coin tosses independent}$$

$$= \prod_{t=1}^{T} P(X = x^{(t)})$$

- Log-probability $\mathcal{L}$

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

$\nearrow$ og-likelihood

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

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

- Notation :

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

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

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

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

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

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

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

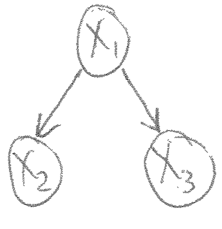
Intuitively, ML estimate is relative empirical frequency of heads

Discrete BNs with "complete data"

- Given: fixed DAG over discrete nodes $\{X_1, X_2, \ldots, X_n\}$
- CPTs enumerate $P(X_i = x \mid \underset{\substack{\nearrow \\ \text{parents of} \\ X_i}}{pa(X_i)} = \underset{\substack{\nwarrow \\ \text{configuration} \\ \text{of parents}}}{\pi})$ as lookup tables
- Data is T complete instantiations of all nodes in BN

$$\{(x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)})\}_{t=1}^{T}$$

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

- Probability of data

joint prob. of t^{th} example

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

- Work out t^{th} term:

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

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

prod. rule

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

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

conditional dependence in BN

- Log-likelihood

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

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

$$= \log \prod_{t=1}^{T} \prod_{i=1}^{n} P\left(x_i^{(t)} \mid pa_i^{(t)}\right)$$

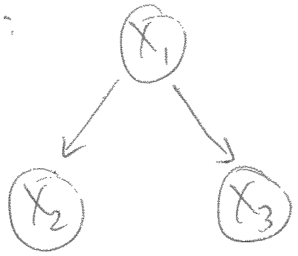
$$= \sum_{t=1}^{T} \sum_{i=1}^{n} \log P\left(x_i^{(t)} \mid pa_i^{(t)}\right)$$

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

swap order of sums

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

Ex:



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	1	1

$\text{count}(X_1 = 1) = 3$
$\text{count}(X_3 = 1, X_1 = 0) = 2$
$\text{count}(X_2 = 1, X_1 = 1) = 1$

- Log-likelihood:

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

- ML estimation:

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

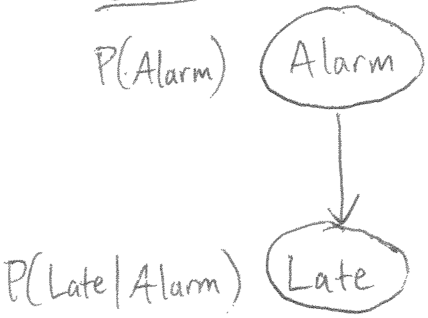
- ML solution (w/out proof):

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

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

Ex: 30-day experiment, set alarm to wake up for class

CPTs



$\text{count}(\text{Alarm}=0) = 10$
$\text{count}(\text{Alarm}=1) = 20$

$\text{count}(\text{Late}=0, \text{Alarm}=0) = 2$
$\text{count}(\text{Late}=1, \text{Alarm}=0) = 8$
$\text{count}(\text{Late}=0, \text{Alarm}=1) = 17$
$\text{count}(\text{Late}=1, \text{Alarm}=1) = 3$

$P_{ML}(\text{Alarm}=0) = 10/30$
$P_{ML}(\text{Alarm}=1) = 20/30$

$P_{ML}(\text{Late}=0 \mid \text{Alarm}=0) = 2/10$
$P_{ML}(\text{Late}=1 \mid \text{Alarm}=0) = 8/10$
$P_{ML}(\text{Late}=0 \mid \text{Alarm}=1) = 17/20$
$P_{ML}(\text{Late}=1 \mid \text{Alarm}=1) = 3/20$

- 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 $\text{count}(pa_i = \pi) = 0$

- Other useful notation:

Indicator function:

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

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

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