Review 7/7

* d-separation

For sets of nodes X, Y, E

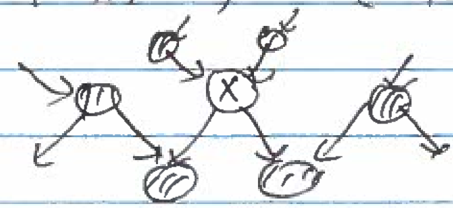
when is
$$\begin{cases} P(Y \mid X, E) = P(Y \mid E) \\ P(X \mid Y, E) = P(X \mid E) \\ P(X, Y \mid E) = P(X \mid E)\, P(Y \mid E) \end{cases}$$

* True if all paths from nodes in X to nodes in Y are "blocked".
A path is blocked if it has a node z that satisfies :

1) $z \in E$ $\quad \rightarrow z \rightarrow$ intervening cause
2) $z \in E$ $\quad \leftarrow z \rightarrow$ common cause
3) $z \notin E$ $\quad \rightarrow z \leftarrow$ no observed common effect
$desc(z) \notin E$

* Markov blanket $B_x$ of node X consists of parents, children and spouses of X.
other parents of children

* Thm : $P(X \mid B_x, Y) = P(X \mid B_x)$ if $Y \notin \{X, B_x\}$



* Inference in BNs
  Query node Q
  Evidence nodes E
  How to compute $P(Q|E)$ ?

* Polytrees
  - singly connected networks
  - polynomial time inference

* loopy BNs
  - exact inference : node clustering, ....
  - approximate inference : stochastic simulation, ...

---

## 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 vs. continuous
(eg. learning DAGs) (eg. learning CPTs)

- algorithms.

non-iterative vs iterative
(loop over data many times)

- solution : local vs global optima

* Maximum likelihood estimation :

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

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

$\boxed{P(X=heads)}$

X

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

$P(X=heads) = p$

$P(X=tails) = 1-p$

* How to estimate $p$ from observed saples (eg. T coin tosses) ?

* IID assumption

Samples are independently, identically distributed according to $P(X)$

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

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

log likelihood

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

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

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

$\Rightarrow N_H + N_T = T$

* In terms of counts:

$$\mathcal{L}(p) = \underbrace{N_H \log p}_{\text{heads}} + \underbrace{N_T \log(1-p)}_{\text{tails}}$$

* Maximum likelihood (ML) estimation:

$$0 = \frac{dL}{dp} = \frac{N_H}{p} + \frac{N_T(-1)}{1-p} \Rightarrow N_H(1-p) + N_T p = 0$$

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

ML estimate of $p = P(\text{heads})$ is just empirical frequency...

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

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

* CPTs enumerate $P(X_i = x \mid \underbrace{pa(X_i)}_{\text{parents of } X_i} = \pi)$ as lookup tables

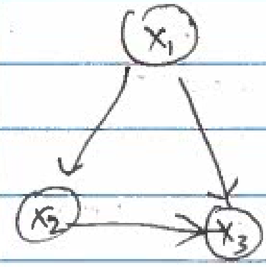
some configuration of parents.

* Data is T complete

instantiations of all nodes in BN

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

Ex:



$n=3$

| $t$ | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| 1 | 0 | 1 | 3 |
| 2 | 1 | 2 | 4 |
| 3 | 0 | 7 | 2 |
| ⋮ | 0 | ⋮ | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| T | 0 | 4 | 5 |

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

Goal: learn from examples
estimate CPTs $\xi$ $P(X_i = x \mid pa_i = \pi)$ that maximize $\underbrace{\text{probability}}_{\text{likelihood}}$ of data set.

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

* Probability of (IID) data set

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

joint probability of $t^{th}$ example.

* Work out $t^{th}$ term in product :

$$P(X_1 = x_1^{(t)}, \dots X_n = x_n^{(t)})$$

product rule

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

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

$$= \prod_{i=1}^{n} P(X_i = x_i^{(t)} | pa(X_i) = pa_i^{(t)})$$ cond. ind

* log-likelihood

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

$$= \log \prod_{t=1}^{T} P(x_1^{(t)}, x_2^{(t)}, \dots x_n^{(t)})$$

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

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

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} \log P(X_i^{(t)} | pa(X_i) = pa_i^{(t)})$$ (swap order of sums)

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

Ex count $(X_2 = 2, X_1 = 0) = 1$

count $(X_2 = 1, X_1 = 0) = 2$

$\vdots$

* log-likelihood

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

values that $X_i$ can assume

configuration of parents of $X_i$

* ML estimation

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

* ML solution (without proof):

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

(empirical frequency of $X_i = x$ and $pa_i = \pi$ in your data.)

$$= \begin{cases} \dfrac{\text{count}(X_i = x,\ pa_i = \pi)}{\text{count}(pa_i = \pi)} & \text{when } X_i \text{ has parents} \\ \\ \dfrac{\text{count}(X_i = x)}{\underbrace{T}_{\#\text{ examples}}} & \text{when } X_i \text{ is root node} \end{cases}$$

* Properties of ML estimation

- Asymptotically correct:

$$P_{ML}(X_1, X_2 \ldots X_n) \longrightarrow P(X_1 X_2 \ldots X_n) \quad \text{as } T \to \infty$$

- Problematic for sparse data ($T$ small):

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

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

* Other useful notation

Indicator function:

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

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

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

---

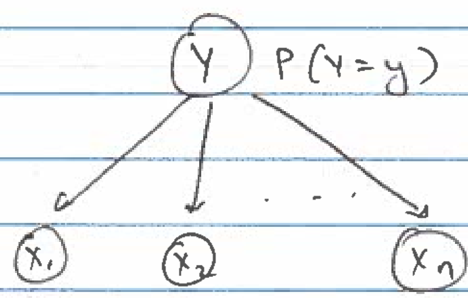
Ex: Naive Bayes model for document classification.

* Variables

$Y \in \{1, 2, \ldots n\}$ — e.g. 1 = sports, 2 = politics, ⋮

$(i = 1..n)$ $X_i \in \{0, 1\}$ — does the $i^{th}$ word in the dictionary appear in document?

* BN = DAG + CPTs




$P(Y=y)$

$P(X_i = 1 \mid Y = y)$ for $i = 1 \ldots n$

* How to use model for document classification.

$$P(Y = y \mid \vec{X} = \vec{x}) = \frac{P(\vec{X} = \vec{x} \mid Y = y)\, P(Y = y)}{P(\vec{X} = \vec{x})} \qquad \text{Bayes rule}$$

$$= \frac{\left\{\prod_{i=1}^{n} P(X_i = x_i \mid Y = y)\right\} P(Y = y)}{\sum_{y'} \left\{\prod_{i=1}^{n} P(X_i = x_i \mid Y = y')\right\} P(y = y')} \qquad \begin{array}{l}\text{prod rule} \\ + \text{CI} \\ \text{normalization}\end{array}$$

* How to learn a model from a large corpus of documents?

$P_{ML}(Y=y)$ fraction of documents with topic $y$

$P_{ML}(X_i=1 \mid Y=y)$ fraction of documents with topic $y$ that contain $i^{th}$ word in dict

* Weaknesses of model

- "naive Bayes" assumption that words appear independently given topic

- "bag-of-words" representation ignores word ordering

Ex: $\boxed{\text{Markov models of language}}$

* Let $w_\ell$ denote the word at $\ell^{th}$ position in sentence.

How to model $P(w_1, w_2, \ldots, w_L)$? probability of sentence with $L$ words.

* Simplifying assumptions:

(1) finite context / memory. "k-gram" model

$$P(w_\ell \mid w_1, w_2, \ldots w_{\ell-1}) = P(w_\ell \mid \underbrace{w_{\ell-1}, w_{\ell-2}, \ldots w_{\ell-(k-1)}}_{k-1 \text{ previous words}})$$

eg. $P(w_\ell \mid w_1 \ldots w_{\ell-1}) = P(w_\ell \mid w_{\ell-1})$ "bi-gram" model.

(2) position invariance (for bigram model)

$$P(w_{l+1} = w' \mid w_l = w) = P(w_{l+b+1} = w' \mid w_{l+b} = w)$$

b is any positive or negative shift.

* BN for bigram-model of language

$$w_1 \to w_2 \to w_3 \to \cdots \to w_{l-1} \to w_L$$

Also : same CPT is used at all non-root nodes in BN.

* learning bigram model.

- collect large corpus of text $\sim 10^8$ words (at least)
- vocabulary size $V \sim 10^5$ dictionary entries

* Count $C_{ij}$ = # times that word $j$ follows word $i$

count $C_i$ = # times that word $i$ appears in corpus

estimate $P_{ML}(w_{l+1} = j \mid w_l = i) = \dfrac{C_{ij}}{C_i}$

* Note : no generalization to unseen word combinations

* n-gram model: condition on previous n-1 words

$$P(w_\ell \mid w_1 \ldots w_{\ell-1}) = P(w_\ell \mid w_{\ell-1}, \ldots w_{\ell-(n-1)})$$

n=1 unigram
n=2 bigram
n=3 trigram
⋮

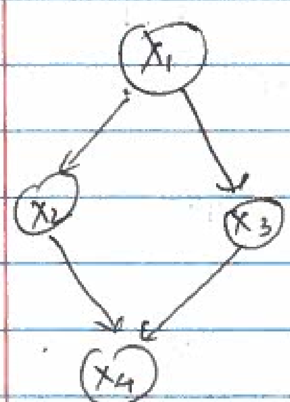
n-gram counts get increasingly sparse for larger n.

## Learning (ML estimation) from incomplete data

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

Also: data set of T examples, but each example is a partial instantiation of nodes in BN.

Ex:




| t | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| 1 | 1 | ? | 4 | ? |
| 2 | 0 | ? | ? | 1 |
| 3 | 1 | | 5 | 3 |
| ⋮ | | | | |
| T-1 | 0 | | 3 | ? |
| T | 1 | ? | ? | |

* Goal : estimate CPTs $P(X_i = x \mid pa_i = \pi)$ that maximize (marginal) probability of partially observed data.

* Variables in BN

$X$ = all nodes in BN

$H$ = subset of nodes that are unobserved ("hidden")

$V$ = subset of nodes observed ("visible").

* Log-likelihood

Assume that $T$ examples are i.i.d from joint distribution $P(X_1, X_2, \ldots, X_n)$

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

visible nodes on $t^{th}$ example

$$= \log \prod_{t=1}^{T} P(V = v^{(t)})$$

marginal probability (not joint)

$$= \sum_{t=1}^{T} \log P(V = v^{(t)})$$