$\boxed{\text{Review}}$ Learning in BNs

1/29/15
CSE 150

* Maximum likelihood estimation (ML)
  Estimate CPTs that maximize probability of observed data (evidence)

* Complete data (a.k.a. fully observed)
  Data set $\{(X_1^{(t)}, X_2^{(t)}, \ldots, X_n^{(t)})\}_{t=1}^{T}$ is T complete instantiations of nodes $X_1, X_2, \ldots, X_n$

t	X_1	X_2	...	X_n
0	0	1		0
1	1	1		0
2	0	1		1
3			...	
⋮	⋮	⋮		⋮
T	1	1		0

data set

* ML estimates for CPTs
  Nodes with parents:

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

Root nodes:

$$P_{ML}(X_i = x) = \frac{\text{count}(X_i = x)}{T}$$ where T is # of examples

* Other notation:

Indicator function: $I(x,x') = \begin{cases} 1 & \text{if } x=x' \\ 0 & \text{otherwise} \end{cases}$

- $\text{count}(X_i = x) = \sum_{t=1}^{T} I(x, x_i^{(t)})$ ($x_i^{(t)}$ = i^{th} column in t^{th} example)

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

* Properties of ML estimation

- Asymptotically correct

$$P_{ML}(x_1, x_2, \ldots, x_n) \longrightarrow P(X_1, X_2, \ldots, X_n) \text{ as } T \to \infty$$

- Problematic for small #'s examples (T)
("opposite limit of sparse data")

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

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

Ex: Model for document classification

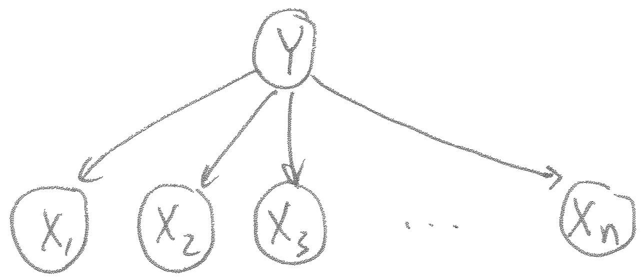
* Variables

$Y \in \{1, 2, \ldots, m\}$ possible document topics

$X_i \in \{0, 1\}$ does the i^{th} word in vocabulary appear in the document?

(This corrects any document to a fixed length bit vector)

* BN = DAG + CPTs



$P(Y=y)$

$P(X_i=1|Y=y)$

* How would you classify documents into topics with this model?

$$\underbrace{P(Y=y|\vec{X}=\vec{x})}_{\text{document's bit vector}} = \frac{P(\vec{X}=\vec{x}|Y=y)\cdot \overbrace{P(Y=y)}^{\text{get from CPT}}}{P(\vec{X}=\vec{x})}$$ Bayes rule

other term in numerator:

$$P(\vec{X}=\vec{x}|Y=y) = \prod_{i=1}^{n} P(X_i=x_i|Y=y)$$ cond. indep with d-sep case II

Denominator:

$$P(\vec{X}=\vec{x}) = \sum_{y'=1}^{m} \left\{ \prod_{i=1}^{n} P(X_i=x_i|Y=y') \right\} P(Y=y')$$ marginalization

* How to learn this model?

Estimate CPTs from a large corpus of $\overset{T}{\uparrow}$ $\underbrace{\text{labeled}}_{\text{of known topic}}$ documents

$$P_{ML}(Y=y) \quad \frac{\text{count}(Y=y)}{T}$$ fraction of documents corresponding to topic y

$$P_{ML}(X_i=1|Y=y) = \frac{\text{count}(X_i=1, Y=y)}{\text{count}(Y=y)}$$ fraction of documents with topic y that contain i^{th} word.

* Weaknesses of model

(i) "bag-of-words" representation (ignores word ordering)

(ii) strong (overly strong) assumption that words are independent given topic (naive Bayes)

Ex: Markov models of language

* Let w_ℓ denotes word at ℓ^{th} position in sentence
  How to model $\underbrace{P(w_1, w_2, \ldots, w_{L-1}, w_L)}_{\text{prob. of sentence with } L \text{ words}}$?

* Simplifying assumptions
  (i) finite context/memory

$$P(w_\ell | w_1, w_2, \ldots, w_{\ell-1}) = P(w_\ell | \underbrace{w_{\ell-(k-1)}, \ldots, w_{\ell-2}, w_{\ell-1}}_{\text{only condition on } (k-1) \text{ preceding words}})$$

"k-gram" model

$$P(w_\ell | w_1, w_2, \ldots, w_{\ell-1}) = P(w_\ell | w_{\ell-1}) \quad \text{is "bigram" model } (k=2)$$

  (ii) position invariance

$$P(w_{\ell+1} = w' | w_\ell = w) = P(w_\ell = w' | w_{\ell-1} = w)$$

* Belief network for bigram model of language

$$w_1 \to w_2 \to w_3 \to \cdots \to w_{L-2} \to w_{L-1} \to w_L$$

same CPT at all non-root nodes in BN

* Learning a bigram model
  - collect large corpus of text ~ 10^8 words
  - commit to vocabulary size $V \sim 10^5$ dictionary entries

* ML estimates

parent i co-occurs with word j

$$P_{ML}(w_{\ell+1} = j | w_\ell = i) = \frac{\text{count(word } i \text{ is followed by word } j)}{\text{count(word } i \text{ occurs)}} = \frac{C_{ij}}{C_i}$$

* Note: no generalization to unseen word combinations

* k-gram model : conditioning on (k-1) previous words

k=1 unigram model

k=2 bigram "

k=3 trigram "

tension: more powerful model but also more prob. where count = 0. as k increases

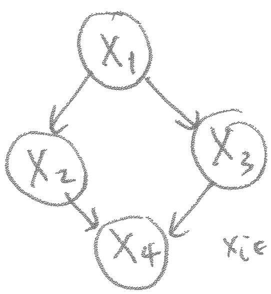
(ngrams.google.com)

Material Up To Quiz 1

Learning from incomplete data

* Given some fixed DAG over discrete nodes $\{X_1, X_2, \ldots, X_n\}$
Also given data set of T partial instantiations of $\{X_1, X_2, \ldots, X_n\}$

Ex :



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

t	X_1	X_2	X_3	X_4
1	1	0	?	1
2	0	?	?	1
3	1	0	?	1
⋮	⋮	?	?	⋮
T	0	1	?	1

* Goal : estimate CPTs $P(X_i = x \mid pa_i = \pi)$ that maximize <u>marginal</u> prob. of <u>partially</u> observed data
(vs. before when we maximize joint prob of complete data)

* Variables in BN

X = all nodes

H = hidden nodes

V = visible nodes

$X = H \cup V$

* Log - likelihood

- Assume that T examples are iid from joint distribution $P(X_1, X_2, \dots, X_n)$ of BN

$$
\begin{aligned}
\mathcal{L} &= \log P(\text{DATA}) \\
&= \log \prod_{t=1}^{T} P(V = v^{(t)}) \qquad \text{marginal prob. of visible nodes in } t^{th} \text{ example} \\
&= \sum_{t=1}^{T} \log P(V = v^{(t)}) \\
&= \sum_{t=1}^{T} \log \Big[\underbrace{\sum_h P(V = v^{(t)}, H = h)}_{\text{joint}} \Big] \qquad \text{marginalization}
\end{aligned}
$$

... much more complicated to optimize b/c sum inside log.