Review

- Learing in BNs
- Maximum likelihood (ML) estimation

Estimate CPTs that maximize $\overbrace{\text{probability of observed data}}^{\text{"likelihood"}}$

- Complete data

$\{(x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)})\}_{t=1}^{T}$ T complete instantiations of nodes $X_1, \ldots, X_n$

- Notation

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

- ML estimates

$$P_{ML}(X_i = x \mid pa_i = \pi) = \begin{cases} \dfrac{\text{count}(X_i = x,\ pa_i = \pi)}{\text{count}(pa_i = \pi)} & \text{for nodes w/ parents} \\[2ex] \dfrac{\text{count}(X_i = x)}{T} & \text{for root nodes} \end{cases}$$

$$= \begin{cases} \dfrac{\sum_{t=1}^{T} I(X_i^{(t)}, x)\, I(pa_i^{(t)}, \pi)}{\sum_{t=1}^{T} I(pa_i^{(t)}, \pi)} \\[3ex] \dfrac{\sum_{t=1}^{T} I(X_i^{(t)}, x)}{T} \end{cases}$$

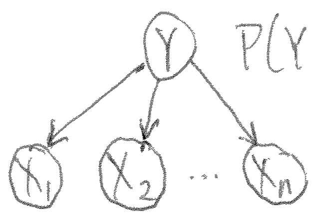
Ex: Naive Bayes model for document classification

- Variables

  $Y \in \{1, 2, \ldots, m\}$ possible document topics (e.g. $Y \in \{\text{spam}, \text{not spam}\}$)

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

- BN = DAG + CPTs



$P(Y=y)$

$P(X_i=1 \mid Y=y)$

- ML estimation of CPTs

  Collect and label a large corpus of N documents

$$P_{ML}(Y=y) = \frac{\text{count}(Y=y)}{N}$$

= fraction of documents w/ topic y

$$P_{ML}(X_i=1 \mid Y=y) = \frac{\text{count}(X_i=1, Y=y)}{\text{count}(Y=y)}$$

= fraction of documents of topic y that contain i^{th} word in vocabulary

- Document classification

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

$$= \frac{\prod_{i=1}^{n} P(X_i = x_i \mid Y=y)\, P(Y=y)}{\sum_{y'} P(\vec{X}=\vec{x}, Y=y')} \qquad \begin{array}{l}\text{conditional independence} \\ \\ \text{marginalization}\end{array}$$

$$= \frac{\prod_{i=1}^{n} P(X_i = x_i \mid Y=y)\, P(Y=y)}{\sum_{y'} \prod_{i=1}^{n} P(X_i = x_i \mid Y=y')\, P(Y=y')} \qquad \begin{array}{c}\text{product rule} \\ + \\ \text{cond. ind.}\end{array}$$

- Strengths of model
  (1) easy to estimate from a large corpus of documents
  (2) simplest baseline

- Weaknesses of model
  (1) Assumption that words appear independently given the topic (naive!)
  (2) "Bag of words" representation ignores order
  (3) Documents have only one topic

Ex: Markov models of language

Why do we need language models?

 $\rightarrow$ speech recognizer / acoustic analysis $\rightarrow$

- eyelid in joyful
- Alvin lap hula
- I live in La Jolla

- Let w_ℓ denote word at ℓ^{th} position in sentence

  How to model $P(w_1, w_2, \ldots, w_L)$?

  Probability of sentence with L words $w_1, w_2, \ldots, w_L$.

- Simplifying assumptions

  (1) finite context/memory

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

"n-gram" model

special case: "bigram" model

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

  (2) position invariance

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

- BN for bigram model of language

$$W_1 \rightarrow W_2 \rightarrow \cdots \rightarrow W_{L-1} \rightarrow W_L$$

same CPTs at all non-root nodes

- Learning bigram model
  - collect large corpus of text, $\sim 10^8$ words
  - vocabulary size $V \sim 10^5$ dictionary entries
  - count c_i = # times word i appears

    c_{ij} = # times word j follows word i

  estimate $P_{ML}(w_\ell = j \mid w_{\ell-1} = i) = \dfrac{c_{ij}}{c_i}$

- Note: no generalization to unseen word combinations (will have 0 probability)

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

  $n=1$ unigram
  $n=2$ bigram
  $n=3$ trigram

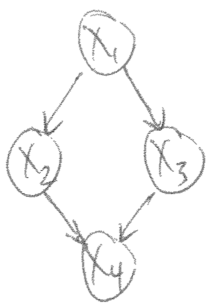
  $$P(w_\ell \mid w_1, \ldots, w_{\ell-1}) = P(w_\ell \mid w_{\ell-(n-1)}, \ldots, w_{\ell-1})$$

  n-gram model counts get more sparse as n increases

ML estimation from incomplete data

- Given fixed graph (DAG) over discrete nodes $\{X_1, X_2, \ldots, X_n\}$
  Also data set of T partial instantiations of $\{X_1, X_2, \ldots, X_n\}$

Ex:



t	X_1	X_2	X_3	X_4
1	0	?	1	1
2	1	?	?	1
3	0	?	1	1
4	?	?	?	0

- Goal: estimate CPTs $P(X_i = x \mid pa_i = \pi)$ that maximize the marginal (not joint) probability of partially observed data.
  (not complete)

- Variables in BN
  X = all nodes $\qquad\qquad X = H \cup V$
  H = hidden nodes
  V = visible nodes

- Log-likelihood: assume T examples are i.i.d. from joint distribution $P(X_1, X_2, \ldots, X_n)$

visible nodes on t^{th} example

$$\begin{aligned}
\mathcal{L} &= \log P(\text{data}) \\
&= \log \prod_{t=1}^{T} P(V^{(t)} = v^{(t)}) \\
&= \sum_{t=1}^{T} \log P(V^{(t)} = v^{(t)})
\end{aligned}$$

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

marginalizing over joint for $X = H \cup V$

hidden nodes

$$= \sum_{t=1}^{T} \log \sum_{h} \prod_{i=1} P(X_i = x \mid pa_i = \pi) \Big|_{\substack{V^{(t)}=v^{(t)} \\ H^{(t)}=h}}$$

- more complicated to optimize $\mathcal{L}$ from incomplete data
  - no "closed-form" solution
  - <u>iterative</u> solution

<u>Expectation - Maximization (EM) algorithm</u>

Iterative procedure to maximize $\mathcal{L}$(data) for incomplete data in terms of CPTs.

By analogy, ML estimates for complete data

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

For incomplete data, we must "fill in" hidden values:

$$P_{ML}(X_i = x \mid pa_i = \pi) \leftarrow \frac{\sum_{t=1}^{T} P(X_i = x, pa_i = \pi \mid V = v^{(t)})}{\sum_{t=1}^{T} P(pa_i = \pi \mid V = v^{(t)})}$$

Intuition: expected statistics ("counts") under $P(H|V)$ substitute for observed counts in complete data case