* Learning in BNs.
* Maximum likelihood (ML) estimation.

  Estimate CPTs that removes $\overbrace{\text{probability of observed data}}^{\text{"likelihood"}}$
* Complete data.

  $x_1^{(t)}, x_2^{(t)} \dots x_n^{(t)}$ T complete instantiations of nodes $X_1 \dots X_n$.

* ML Estimates.

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

Equivalently.

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

* Other notation.

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

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

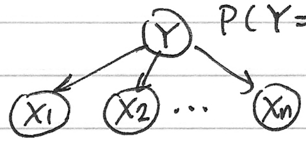
Ex: "Naive Bayes" model for document classification.

* Variables

  $y \in \{1, 2 \dots m\}$ possible document ~~types~~ topics

  $X_i \in \{0, 1\}$. = 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 documents

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

$P_{ML}(X_i = 1 \mid y)$ = fraction of documents of topic y that contains i^{th} word in vocabulary.

* Document Classification.

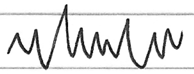
$$P(Y=y \mid \vec{X}=\vec{x}) = \frac{P(\vec{X}=\vec{x} \mid Y=y) \cdot P(Y=y)}{P(\vec{X}=\vec{x})} \quad \text{Bayes rule.}$$

$$= \frac{\prod_{i=1}^{n} P(X_i = x_i \mid Y=y)\ P(Y=y)}{\sum_{y'} \prod_{i=1}^{n} P(X_i = x_i \mid Y=y') \cdot P(Y=y')} \quad \text{Conditional independence.}$$

sum of nominators

* Weakness of this "Naive Bayes" model.
  (1). Assumption that words appear independently given the topic
  (2). Documents have only one topic
  (3). "Bag of words" representation. : ignores order.

Ex: Markov. model of language
why do we need language models.

 $\rightarrow$ $\dfrac{\text{speech recognizer}}{\text{acoustic analysis.}}$ $\rightarrow$
- 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 $P(W_1, W_2 \cdots W_L)$?
  Probability of sentence with L words $W_1, W_2 \cdots W_L$.

* Simplifying assumptions
  (1). finite context/memory.

$$P(W_\ell \mid \underbrace{W_1, W_2 \cdots W_{\ell-1}}_{\text{all previous words}}) = P(W_\ell \mid \overbrace{W_{\ell-(k-1)}, W_{\ell-(k-2)} \cdots W_{\ell-1}}^{k-1 \text{ previous words.}}) \quad \text{"k-gram"}$$

– Special case : "bigram" model.

$$P(W_\ell \mid W_1 \cdots W_{\ell-1}) = P(W_\ell \mid W_{\ell-1}).$$

  (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 language.

$W_1 \to W_2 \to \cdots \to W_{L-1} \to W_L$ $\qquad$ same CPTs at all $\underset{\text{non-root}}{\text{nodes}}$

* Learning bigram model.
  - collect large corpus of text $\sim 10^8$ words
  - 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.

    estimate $P_{ML}(W_\ell = j \mid W_{\ell-1} = i) = \frac{C_{ij}}{C_i}$

* Note: no generalization to unseen word combinations.
* "n-gram" model: condition on $n-1$ previous words

  $P(W_\ell \mid W_1 \cdots W_{\ell-1}) = P(W_\ell \mid W_{\ell-(n-1)}, \cdots W_{\ell-1})$

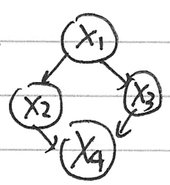
  n-gram model counts get more sparse as n increases.

ML estimation from incomplete data

* Given fixed graph (DAG) over discrete nodes $\{X_1, X_2 \cdots X_n\}$.

  Also data set of T partial instantiations of $\{X_1, X_2 \cdots X_n\}$.

Ex:



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

* Goal is to estimate CPTs $P(X_i = x \mid Pa_i = \pi)$ that maximizes the marginal (not joint) probability of partially observed data.

  (not complete)

* Variables in BN.

  X = all nodes $\qquad X = H \cup V$.

  H = hidden nodes

  V = visible nodes.

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

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

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

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.
  - Iterative solution.