CSE 250A

→ Learning in BNs

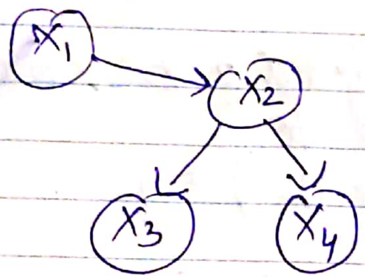
<u>Case I</u> Fixed DAG, complete data, lookup CPTs

Nodes $\{X_1, X_2, \ldots, X_n\}$

CPTs enumerate $P(X_i = x \mid pa_i = \pi)$

<u>IID Data</u> $\left\{\left(x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)}\right)\right\}_{t=1}^{T}$

T complete instantiations of BN.



t	X_1	X_2	X_3	X_4
1	.	.	.	:
2	.	.	.	.
⋮	.	.	.	.
T	.	.	.	.

filled table

- Log-likelihood of IID data

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

$$= \sum_{i=1}^{n} \sum_{x} \sum_{\pi} \underbrace{\text{count}(X_i = x, pa_i = \pi)}_{\text{properties of data}} \cdot \underbrace{\log P(X_i = x \mid pa_i = \pi)}_{\text{unknown CPTs to be estimated from data}}$$

x: Possible values of X_i

π: values of parents of X_i

- Write $\mathcal{L} = \sum_{i\pi} \mathcal{L}_{i\pi}$

  where $\mathcal{L}_{i\pi} = \sum_x \text{count}(X_i = x, pa_i = \pi) \log P(X_i = x | pa_i = \pi)$

  We can independently optimize each row of each CPT in BN! [Only true for complete data]

- ML Estimation

  For each node X_i, and for each row π maximize $\mathcal{L}_{i\pi}$ subject to:

  1.) $\sum_x P(X_i = x | pa_i = \pi) = 1$

  2.) $P(X_i = x | pa_i = \pi) \geq 0$

- Shorthand:

  Let $C_\alpha = \text{count}(X_i = \alpha, pa_i = \pi)$

  Let $p_\alpha = P(X_i = \alpha | pa_i = \pi)$

  How to maximize $\sum_\alpha C_\alpha \log p_\alpha$ such that

  $p_\alpha \geq 0$,

  $\sum p_\alpha = 1$?

  How to minimize $\sum_\alpha C_\alpha \log \frac{1}{p_\alpha}$

  equivalent to minimizing $\sum_\alpha C_\alpha \log \frac{C_\alpha}{p_\alpha}$

  $\because C_\alpha$ are constants

Also same as minimizing $\sum_{\alpha} \left(\frac{C_\alpha}{\sum_\beta C_\beta}\right) \log \frac{(C_\alpha / \sum_\beta C_\beta)}{p_\alpha}$

KL distance

Solution: $\boxed{p_\alpha = \frac{C_\alpha}{\sum_\beta C_\beta}}$

ML solution:

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

- Properties:

- Asymptotically correct

$P_{ML}(X_1, X_2, \ldots X_n) \rightarrow P(X_1, X_2, \ldots, X_n)$ as $T \rightarrow \infty$

- Problematic in non-asymptotic regime (sparse data):

$$P_{ML}(X_i = x \mid pa_i = \pi) = \begin{cases} 0 & \text{if count}(X_i = x, pa_i = \pi) \text{ but count}(pa_i = \pi) \neq 0 \\ \text{undefined} & \text{if count}(pa_i = \pi) = 0 \end{cases}$$

Ex: Markov Models of Language.

- Let w_ℓ denote ℓ^{th} word in sentence (or text)
- How to model $P(w_1, w_2, \ldots, w_L)$?
- Simplifying assumptions:

1) Finite context / history / memory:

$$P(w_\ell | w_1, w_2, \ldots, w_{\ell-1}) = P(w_\ell | w_{\ell-(n-1)}, \ldots, w_{\ell-2}, w_{\ell-1})$$

2) Position invariance:

$$P(w_\ell = w' | w_{\ell-(n-1)}, \ldots, w_{\ell-2}, w_{\ell-1}) =$$
$$P(w_{\ell+s} = w' | w_{\ell-(n-1)+s}, \ldots, w_{\ell-2+s}, w_{\ell-1+s})$$

- Markov Model

$$P(w_1, w_2, \ldots w_L) = \prod_\ell P(w_\ell | w_1, w_2, \ldots w_{\ell-1})$$

Product Rule

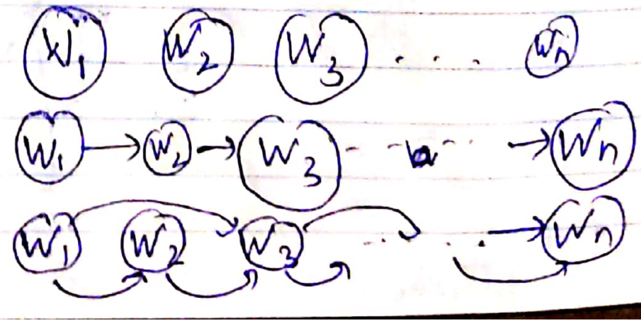
$$= \prod_\ell P(w_\ell | w_{\ell-(n-1)}, \ldots, w_{\ell-1}) \quad CI$$

- Models of different orders:

n=1 unigram

n=2 bigram

n=3 trigram



- Focus on bigram ($n = 2$) :

Same CPT $P(w_\ell = w' \mid w_{\ell-1} = w)$ used at each node ($\ell \geq 1$)

- How to learn P

Collect large corpus of text $\sim 10^{10}$ words
Commit to vocabulary size $\sim 10^{4-6}$

Count C_{ij} = # times that i^{th} word is followed by j^{th} word in vocab.

C_i = # times that i^{th} word is followed by anything. (i.e. $C_i = \sum_j C_{ij}$)

Estimate $P_{ML}(w_\ell = j \mid w_{\ell-1} = i) = C_{ij}/C_i$

- <u>Problems</u> with ML estimates for n-gram models:

- no generalization to unseen n-grams.

- n-gram counts become increasingly sparse as n increases.

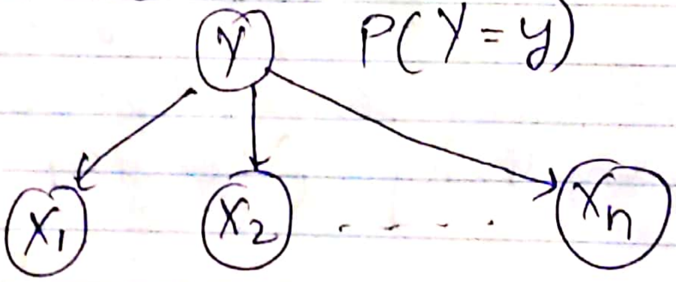
Ex: Naive Bayes model (for document classification)

- variables $Y \in \{1, 2, \dots m\}$ topic label (eg. sports, politics)

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

  Used X_i to represent each document as a fixed length bit vector.

- BN = DAG + CPT



$P(Y = y)$

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

CPTs are unknown. How to estimate from data?

- How to learn P?

- Collect corpus of documents and labels for each document.
- Estimate: $P_{ML}(Y = y)$ = fraction of documents with label y in corpus.

$P_{ML}(X_i = 1 | Y = y)$ = fraction of documents with label y that contain i^{th} word in dictionary

- How to classify?

$$P(Y=y \mid X_1, X_2, \ldots, X_n) = \frac{P(X_1, \ldots X_n \mid Y=y) P_{ML}(Y=y)}{P(X_1, X_2, \ldots X_n)}$$

$$= \frac{P_{ML}(Y=y) \prod_{i=1}^{n} P_{ML}(X_i \mid Y=y)}{\sum_{y'} P_{ML}(Y=y') \prod_{i=1}^{n} P_{ML}(X_i \mid Y=y')}$$

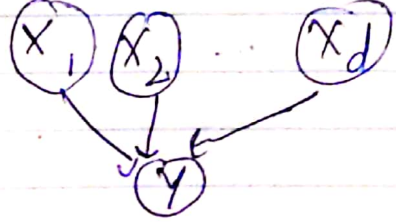
- Weaknesses:

1) "Naive" Bayes assumption that words appear independently given the topic.

2) "Bag of Words" representation ignores word ordering.

→ $\boxed{\text{Case II}}$ Fixed DAG, complete data, parametrized CPTs.

(Preview) II A. linear regression



How to predict real valued $Y \in \mathbb{R}$ from parents $\vec{X} \in \mathbb{R}^d$?

- Gaussian CPT

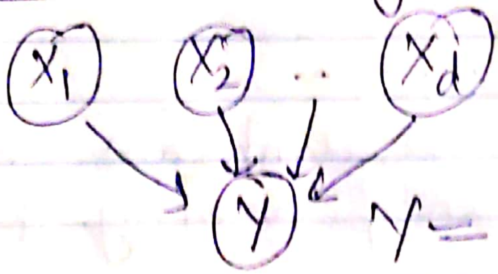
$$P(Y=y \mid \vec{X}=x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y - \sum_{i=1}^{d} w_i x_i\right)^2\right\}$$

σ^2 ↓ variance

How to estimate $\vec{w}$ and σ^2?

$\vec{w}$ ↓ $(w_1, w_2, \ldots, w_d)$

- Case II B Logistic Regression



$\vec{X} \in \mathbb{R}^d$

$Y = \{0, 1\}$ binary

How to predict binary Y from a real valued $\vec{X}$

Sigmoid CPT

$$P(Y=1 \mid \vec{X}) = \sigma(\vec{w} \cdot \vec{x})$$

How to estimate $\vec{w} = (w_1, \ldots w_d)$ from data?