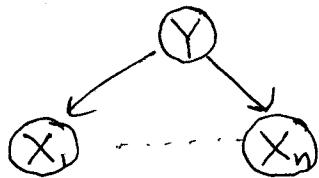


Review

- ML Estimation from complete data

$$P_{ML}(X_i=x | p_{a_i}=\pi) = \frac{\text{count}(X_i=x, p_{a_i}=\pi)}{\text{count}(p_{a_i}=\pi)}$$

- Naive Bayes model for document classification



$Y \in \{1, 2, \dots, m\}$ topic label

$X_i \in \{0, 1\}$ is i th word in document?

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

$P_{ML}(X_i=1 | Y=y)$ = fraction of documents on topic y with i th word

Markov models of language

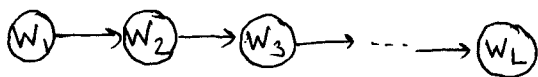
- How to model word sequences w_1, w_2, \dots, w_L in text?

- Simplifying assumptions:

(i) $P(w_L | w_1, w_2, \dots, w_{L-1}) = P(w_L | w_{L-1})$

(ii) $P(w_{L+1}=w' | w_L=w) = P(w_L=w' | w_{L-1}=w)$ } bigram model

- Belief network



same CPTs at all (non-root) nodes in network

- Learning bigram model

- collect large corpus of text $\sim 10^8$ words

- vocabulary size $V \sim 10^5$

- count c_{ij} = # times that word j follows word i

- count c_i = # times that word i appears

Estimate $P_{ML}(w_L=j | w_{L-1}=i) = \frac{c_{ij}}{c_i}$

- Note: no generalization to unseen word pairs
 - n-gram model: condition on previous n-1 words

$$P(w_L | w_1, w_2, \dots, w_{L-1}) = P(w_L | w_{L-n+1}, \dots, w_{L-1})$$
 - n=1 unigram
 - n=2 bigram
 - n=3 trigram
- n-gram counts get increasingly sparse for large n.

MLE from incomplete data

- Given: fixed DAG over discrete nodes $\{X_1, X_2, \dots, X_n\}$
Also: data set of T partial instantiations of $\{X_1, X_2, \dots, X_n\}$
- Goal: estimate CPTs $P(X_i=x | pa_i=\pi)$ that maximize the probability of partially observed data
- Variables in BN
 - X = all $X = H \cup V$
 - H = hidden
 - V = visible
- Log-likelihood
Assume T examples are I.I.D. from joint distribution $P(X_1, X_2, \dots, X_n)$.

$$\begin{aligned}
 \mathcal{L} &= \log P(\text{data}) \\
 &= \log \left[\prod_{t=1}^T P(V=v^{(t)}) \right] \begin{array}{l} \text{visible nodes} \\ \text{only on } t^{\text{th}} \text{ example} \end{array} \\
 &= \sum_t \log P(V=v^{(t)}) \\
 &= \sum_t \log \sum_h P(V=v^{(t)}, H=h) \begin{array}{l} \text{marginalizing over} \\ \text{joint for } X=V \cup H \end{array} \\
 &= \sum_t \log \sum_h \prod_i P(X_i=x | pa_i=\pi) \Big|_{\substack{H=h \\ V=v^{(t)}}}
 \end{aligned}$$

- More complicated to optimize \mathcal{L} for incomplete data.
No "closed form" solution.
Alternative: iterative solution.

- Expectation-Maximization (EM) algorithm
iterative procedure to maximize $\mathcal{L}(\text{data})$
when incomplete in terms of CPTs

- Intuition— by analogy, ML estimates for complete data

$$P_{ML}(X_i=x | p_{a_i}=\pi) = \frac{\text{count}(X_i=x, p_{a_i}=\pi)}{\text{count}(p_{a_i}=\pi)} = \frac{\sum_{\mathcal{I}} \mathbb{I}(X_i^{(\mathcal{I})}, x) \mathbb{I}(p_{a_i}^{(\mathcal{I})}, \pi)}{\sum_{\mathcal{I}} \mathbb{I}(p_{a_i}^{(\mathcal{I})}, \pi)}$$

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

$$P(X_i=x | p_{a_i}=\pi) \leftarrow \frac{\sum_{\mathcal{I}} P(X_i=x, p_{a_i}=\pi | V=v^{(\mathcal{I})})}{\sum_{\mathcal{I}} P(p_{a_i}=\pi | V=v^{(\mathcal{I})})}$$

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

- Iterative algorithm

E-step: compute posterior probabilities

$$P(X_i=x, p_{a_i}=\pi | V=v^{(\mathcal{I})}) \quad [\text{inference}]$$

M-step: Update CPTs using (*).

Iterate until convergence.

Why iterative? RHS of (*) depends on current CPTs.

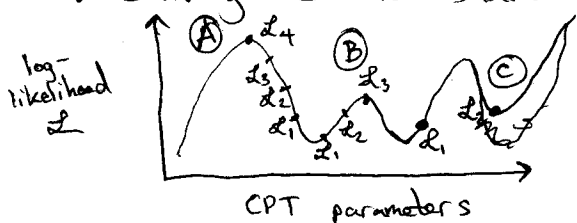
- Key properties

- Monotonic convergence

Each iteration of EM improves the log-likelihood: $\mathcal{L} = \sum_{\mathcal{I}} \log P(V=v^{(\mathcal{I})})$

If $\mathcal{L}_k = \log\text{-likelihood}$ at k^{th} iteration of EM, then $\mathcal{L}_k \geq \mathcal{L}_{k-1}$

- Convergence to stationary point



- (A) global maximum: most desirable, but not guaranteed
- (B) local maximum: usual result
- (C) local minimum: possible in theory, not an issue in practice.

• Key properties (cont'd)

• No tuning parameters

no learning rates, no step sizes,
no line searches, no backtracking...

Example



A and C observed/visible
B is hidden

• Posterior probability

$$P(B=b | A=a, C=c) = \frac{P(C=c | B=b, A=a) P(B=b | A=a)}{\sum_{b'} P(C=c | B=b', A=a) P(B=b' | A=a)} \quad \text{Bayes rule}$$
$$= \frac{P(C=c | B=b) P(B=b | A=a)}{\sum_{b'} P(C=c | B=b') P(B=b' | A=a)} \quad \text{conditional independence.}$$

shorthand:

$$P(b|a, c) = \frac{P(c|b) P(b|a)}{\sum_{b'} P(c|b') P(b'|a)}$$

• Log-likelihood of incomplete data set $\{(a_t, c_t)\}_{t=1}^T$

$$\mathcal{L} = \sum_t \log P(A=a_t, C=c_t)$$

$$= \sum_t \log \sum_b P(A=a_t, B=b, C=c_t)$$

$$= \sum_t \log \sum_b [P(A=a_t) P(B=b | A=a_t) P(C=c_t | B=b)]$$

$$= \sum_t \log \sum_b [P(a_t) P(b|a_t) P(c_t|b)]$$