10/28 $\boxed{\text{Review}}$

* ML estimation for incomplete data

  Examples $t = 1, 2, \cdots, T$

  Visible nodes $V^{(t)}$

* EM algorithm

  E-step : compute posteriors $P(X_i = x, pa_i = \pi \mid V^{(t)})$ inference

  M-step : update CPTs

$$P(X_i = x \mid pa_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, pa_i = \pi \mid V^{(t)})}{\sum_t P(pa_i = \pi \mid V^{(t)})}$$

  Converges to local maximum of $\mathcal{L} = \sum_t \log P(V^{(t)})$

$\boxed{\text{Example}}$

$(A) \longrightarrow (B) \longrightarrow (C)$ A and C are observed. B is hidden.

* Posterior probability

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

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

* Incomplete data set $\{(a_t, c_t)\}_{t=1}^{T}$ (I.I.D.)

  Log-likelihood

$$\begin{aligned} \mathcal{L} &= \sum_t \log P(a_t, c_t) \\ &= \sum_t \log \sum_b P(a_t, b, c_t) \\ &= \sum_t \log \sum_b \left[P(a_t)\, P(b \mid a_t)\, P(c_t \mid b) \right] \end{aligned}$$

  M-step for this example :

$$P(B=b \mid A=a) \longleftarrow \frac{\sum_t P(A=a, B=b \mid A=a_t, C=c_t)}{\sum_t P(A=a \mid A=a_t, C=c_t)}$$

  Simplify RHS $= \dfrac{\sum_t I(a, a_t)\, P(b \mid a_t, c_t)}{\sum_t I(a, a_t)}$

$$P(C=c \mid B=b) \longleftarrow \frac{\sum_t P(c, b \mid A=a_t, C=c_t)}{\sum_t P(b \mid A=a_t, C=c_t)}$$

  Simplify RHS $= \dfrac{\sum_t I(c, c_t)\, P(b \mid a_t, c_t)}{\sum_t P(b \mid a_t, c_t)}$

Ex : Markov models of language.

* Let w_ℓ denote ℓ^{th} word in text.

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

Model	$P(\vec{w})$	ML estimate	DAG G
unigram	$\prod_\ell P_1(w_\ell)$	$P_1(w) = \frac{\text{count}(w)}{\sum_{w'} \text{count}(w')}$	$W_1 \quad W_2 \quad \cdots \quad W_L$
bigram	$\prod_\ell P_2(w_\ell \mid w_{\ell-1})$	$P_2(w' \mid w) = \frac{\text{count}(w, w')}{\text{count}(w)}$	$W_1 \rightarrow W_2 \rightarrow \cdots \rightarrow$

* Evaluating n-gram models

train on corpus A : $P_1(\vec{w}) \le P_2(\vec{w}) \le P_3(\vec{w}) \le \cdots$

test on corpus B : $P_2(\vec{w}) = 0$ if there are unseen bigrams.

$P_3(\vec{w}) = 0$ " " " " trigrams.

Word clustering

* Alternative to bigram model

$$W \longrightarrow W'$$

Replace by:

$$W \longrightarrow Z \longrightarrow W'$$

(words w, w' observed
(cluster z hidden.

CPTs in BN:

$P(z \mid w)$ = prob that word w is mapped into cluster z.

$P(w' \mid z)$ = prob that word in cluster z is followed by word w'.

In cluster model: $P(w' \mid w) = \sum_{z=1}^{C} P(w' \mid z) P(z \mid w)$.

* compact representation

\# words $= V$ (vocabulary size)

\# clusters $= C$ (clusters)

\# parameters $= 2CV$ (reduce to unigram if $C = 1$

\# bigrams $= V^2$ (recover bigrams if $C = V$.

* learning : how to estimate $P(z \mid w)$ and $P(w' \mid z)$?

E-step :

$$P(z \mid w, w') = \frac{P(w' \mid z) P(z \mid w)}{\sum_{z'=1}^{C} P(w' \mid z') P(z' \mid w)} \quad \text{Bayes rule.}$$

M-step : update CPTs

$$P(z \mid w) \longleftarrow \frac{\sum_\ell I(w, w_\ell) P(z \mid w_\ell, w_{\ell+1})}{\sum_\ell I(w, w_\ell)}$$

$$P(w' \mid z) \longleftarrow \frac{\sum_\ell I(w_{\ell+1}, w') P(z \mid w_\ell, w_{\ell+1})}{\sum_\ell P(z \mid w_\ell, w_{\ell+1})}$$

* Experimental results

$V = 60{,}000$ word vocabulary.

$L = 80$ million word corpus of WSJ articles.

count(w, w') is sparse : 99.8% elements are zero.

$\text{count}(w, w') = \sum_{\ell} I(w, w_\ell)\, I(w', w_{\ell+1})$

$C = 32$ model trained by EM.

$P(z|w)$ and $P(w'|z)$ — approx 4 million parameters.

Converges in ~ 30 iterations.

What clusters are discovered? For each word w, what is $\max_z P(z|w)$?

$\boxed{\text{Linear interpolation of Markov models}}$

$$\underbrace{P_M(w_\ell \mid w_{\ell-1}, w_{\ell-2})}_{\text{mixture model}} = \lambda_1 \underbrace{P_1(w_\ell)}_{\text{unigram}} + \lambda_2 \underbrace{P_2(w_\ell \mid w_{\ell-1})}_{\text{bigram}} + \lambda_3 \underbrace{P_3(w_\ell \mid w_{\ell-1}, w_{\ell-2})}_{\text{trigram}}$$

n-gram models are trained on corpus A.

How to estimate λ_i where $\lambda_i \geq 0$ and $\sum_{i=1}^{3} \lambda_i = 1$?

* Methodology

Train P_1, P_2, P_3 on corpus A.

Fix P_1, P_2, and P_3.

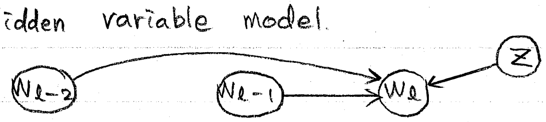
Train $\lambda_1, \lambda_2, \lambda_3$ on corpus C.

Estimate λ to maximize log-likelihood of corpus C.

- Don't estimate λ_i's on corpus A : $\left.\begin{array}{l}\lambda_1 = \lambda_2 = 0 \\ \lambda_3 = 1\end{array}\right\}$ always favor trigram model.
- Test $P_M = \sum_{i=1}^{3} \lambda_i P_i$ on corpus B.

  Don't estimate λ_i's on corpus B : cheating.

* Hidden variable model.



hidden $P(Z=i) = \lambda_i \quad (i = 1, 2, 3)$

<u>CPT for w_ℓ</u>

$$P(w_\ell \mid z, w_{\ell-1}, w_{\ell-2}) = \begin{cases} P_1(w_\ell) & \text{if } z=1 \\ P_2(w_\ell \mid w_{\ell-1}) & \text{if } z=2 \\ P_3(w_\ell \mid w_{\ell-1}, w_{\ell-2}) & \text{if } z=3 \end{cases}$$

In this model:
$$\begin{aligned} P(w_\ell \mid w_{\ell-1}, w_{\ell-2}) &= \sum_{z=1}^{3} P(w_\ell, z \mid w_{\ell-1}, w_{\ell-2}) \quad \text{marginalization} \\ &= \sum_{z=1}^{3} P(z \mid w_{\ell-1}, w_{\ell-2})\, P(w_\ell \mid z, w_{\ell-1}, w_{\ell-2}) \quad \text{product rule} \\ &= \sum_{z=1}^{3} P(z)\, P(w_\ell \mid z, w_{\ell-1}, w_{\ell-2}) \quad \text{conditional independence} \\ &= \lambda_1 P_1(w_\ell) + \lambda_2 P_2(w_\ell \mid w_{\ell-1}) + \lambda_3 P_3(w_\ell \mid w_{\ell-1}, w_{\ell-2}) \end{aligned}$$

* E-step

Compute posterior probability

$$P(Z=i \mid w_\ell, w_{\ell-1}, w_{\ell-2}) = \frac{P(w_\ell \mid Z=i, w_{\ell-1}, w_{\ell-2})\, P(Z=i)}{\sum_j P(w_\ell \mid Z=j, w_{\ell-1}, w_{\ell-2})\, P(Z=j)}$$

$$= \frac{\lambda_i P_i(w_\ell \mid Z=i, w_{\ell-1}, w_{\ell-2})}{\lambda_1 P_1(w_\ell) + \lambda_2 P_2(w_\ell \mid w_{\ell-1}) + \lambda_3 P_3(w_\ell \mid w_{\ell-1}, w_{\ell-2})}$$

* M-step

Update parameters $\lambda_i = P(Z=i)$

General EM : $P(X_i = x \mid pa_i = \pi) \longleftarrow \dfrac{\sum_t P(X_i = x, pa_i = \pi \mid V^{(t)})}{\sum_{x'} \sum_t P(X_i = x', pa_i = \pi \mid V^{(t)})}$

Translation : t^{th} example $\longleftrightarrow$ ℓ^{th} word triplet

$X_i \longleftrightarrow$ node Z

$pa_i \longleftrightarrow \{\ \}$ because Z has no parents.

For mixture model: $P(Z=i) \longleftarrow \dfrac{\sum_\ell P(Z=i \mid w_\ell, w_{\ell-1}, w_{\ell-2})}{\sum_{j=1}^{3} \sum_\ell P(Z=j \mid w_\ell, w_{\ell-1}, w_{\ell-2})}$

$\lambda_i \longleftarrow \dfrac{\sum_\ell P(Z=i \mid w_\ell, w_{\ell-1}, w_{\ell-2})}{L}$ ← # words in corpus.

*Iterate EM: Guaranteed improvement of log-likelihood:

$$\mathcal{L}(\lambda_1, \lambda_2, \lambda_3) = \sum_{\ell=1}^{L} \log \underbrace{P_M(w_\ell \mid w_{\ell-1}, w_{\ell-2})}_{\lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3}$$