Lecture - 13

11/07/2019

REVIEW

- EM updates

  - root nodes (nodes without parents) : $P(X_i = x) \longleftarrow \frac{1}{T} \sum_t P(X_i = x \mid V^{(t)})$

  - nodes with parents : $P(X_i = x \mid Pa_i = \pi) \longleftarrow \dfrac{\sum_t P(X_i = x, Pa_i = \pi \mid V^{(t)})}{\sum_t P(Pa_i = \pi \mid V^{(t)})}$

Updates converge monotonically in log-likelihood $\sum_t \log P(V^{(t)})$

Example # 1

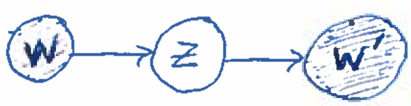
Incomplete data set : $\{(a_t, c_t)\}_{t=1}^{T}$

EM updates : $P(B=b \mid A=a) \longleftarrow \dfrac{\sum_t I(a, a_t)\, P(b \mid a_t, c_t)}{\sum_t I(a, a_t)}$

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

Application # 1

Word Clustering



$w, w' \in \{1, 2, \ldots, V\}$

$z \in \{1, 2, \ldots, k\}$ $(k << V)$

Example # 2

Hidden : H

Observed : A, B, C

- Posterior $P(H \mid A, B, C) = \dfrac{P(C \mid H, A, B)\, P(H \mid A, B)}{P(C \mid A, B)}$ Bayes Rule

$$= \dfrac{P(C \mid H, A, B) \cdot P(H)}{\sum_h P(C \mid H=h, A, B)\, P(H=h)}$$

marginal independence

normalization

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

log (conditional) likelihood : $\mathcal{L} = \sum_t \log P(c_t | a_t, b_t)$

$= \sum_t \log \left(\sum_h P(c_t, h | a_t, b_t) \right)$ marginalization

$= \sum_t \log \sum_h \left[P(h | \cancel{a_t}, \cancel{b_t}) \cdot P(c_t | a_t, b_t, h) \right]$ Product rule & marginal independence

* EM update for CPT at node H : $P(H=h) \leftarrow \frac{1}{T} \sum_t P(H=h | a_t, b_t, c_t)$

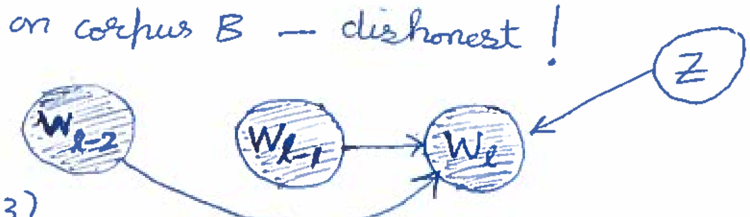
• Aside : $P(V_1 = v, H = h | V_1^{(t)}, V_2^{(t)}) = I(v, V_1^{(t)}) \cdot P(h | V_1^{(t)}, V_2^{(t)})$

Application #2 : Linear Interpolation of Markov Models

• $P_M(w_\ell | w_{\ell-1}, w_{\ell-2}) = \lambda_1 P_1(w_\ell) + \lambda_2 P_2(w_\ell | w_{\ell-1}) + \lambda_3 P(w_\ell | w_{\ell-1}, w_{\ell-2})$

mixture model

* Suppose n-gram models are trained on large corpus A
* How to estimate λ_i ; where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$?
* Methodology : Train P_1, P_2, P_3 on corpus A ; fix these models.
  Train $\lambda_1, \lambda_2, \lambda_3$ on corpus C
  Estimate λ_i to maximize log-likelihood of mixture model on corpus C.
* Don't use corpus A to estimate λ_i.
  - otherwise you find : $\lambda_3 = 1$; $\lambda_1 = \lambda_2 = 0$.
* Test $P_M = \sum_{i=1}^{3} \lambda_i P_i$ on corpus B.
  - Don't estimate λ_i on corpus B — dishonest !
* <u>Hidden Variable Model</u> :



$P(Z=i) = \lambda_i$; $Z \in \{1,2,3\}$

Define CPT at node W_ℓ :

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

In this model,

$\cancel{P(w_\ell | w_{\ell-1}, w_{\ell-2}, z)} = \sum_{z=1}^{3} P(w_\ell, Z=z | w_{\ell-1}, w_{\ell-2})$ marginalization

$= \lambda_1 P_1(w_\ell) + \lambda_2 P_2(w_\ell | w_{\ell-1}) + \lambda_3 P_3(w_\ell | w_{\ell-1}, w_{\ell-2})$

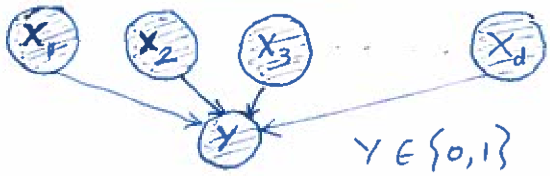
EM Update : $\underset{\lambda_i}{P(Z=i)} \longleftarrow \frac{1}{L_c} \sum_\ell P(Z=i \mid w_\ell, w_{\ell-1}, w_{\ell-2})$

$L_c \hookrightarrow$ Length of corpus c

• If we change the BNws :

: $P(Z=i \mid w_{\ell-1}) = \lambda_i(w_{\ell-1})$

Example / Application # 3 : NOISY-OR (HW 6)

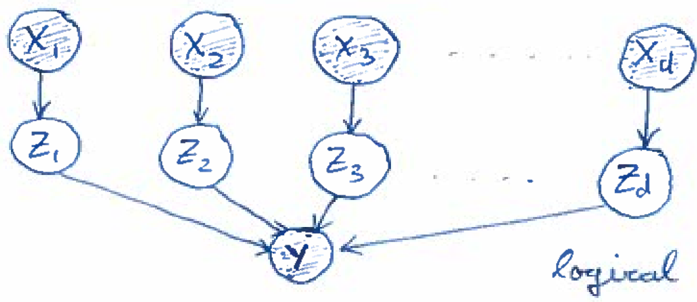


$x_i \in \{0,1\}$

$Y \in \{0,1\}$ $\quad P(Y=1 \mid x_1, x_2, \dots, x_d) = 1 - \prod_i (1-p_i)^{x_i}$

• One approach to ML : gradient ascent / Newton's method for complete data and parameterized CPTs .

Consider Hidden Variable Model :



logical - OR : $Y = \text{OR}(Z_1, Z_2, \dots, Z_d)$

Show $P(Y=1 \mid X_1, \dots, X_d)$ in above model matches noisy - OR .

Hidden Markov Models

* Random Variables : $S_t \in \{1, 2, 3, \dots, n\}$ (hidden) state at time t

$O_t \in \{1, 2, 3, \dots, m\}$ observed at time t

Observations O_t are noisy, partial reflections of true underlying state S_t at time t .

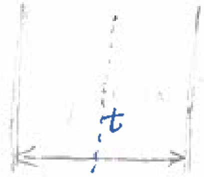
Ex : house - training puppy

$S_t \in \{$ has-to-go , doesn't need to go , went $\}$

$O_t \in \{$ barking , waiting by door , sleeping , hiding , ... $\}$

$P(S_t = i \mid O_1, O_2, \dots, O_t)$ do I take the puppy outside ?

Ex: Speech Recognition



O_t = acoustic measurements of windowed signal around time t.

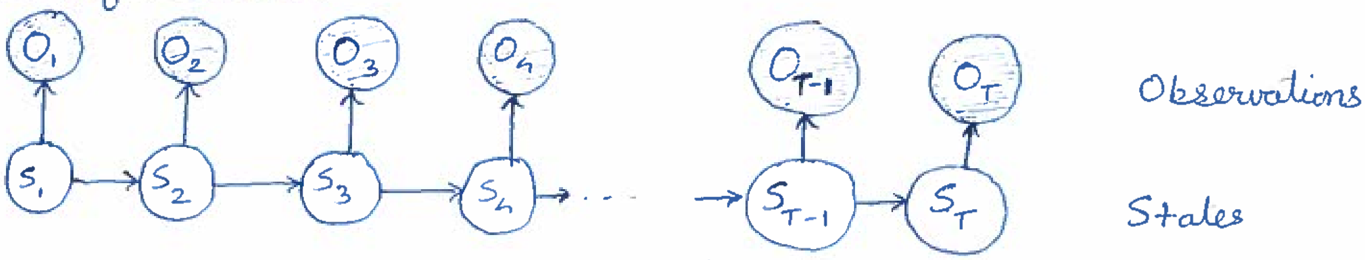
S_t = unit of language (eg. phoneme, letter, syllable) being uttered at time t.

Ex: Robotics

O_t : sensor readings

S_t : location / orientation

* Belief Network



POLYTREE !

* Conditional Independence assumptions

- finite context : $P(S_t \mid S_1, S_2, \ldots, S_{t-1}) = P(S_t \mid S_{t-1})$

$$P(O_t \mid S_1, S_2, \ldots, S_{t-1}, S_t, S_{t+1}, \ldots, S_T) = P(O_t \mid S_t)$$

* CPTs are shared across time

$$P(S_{t+1} = s' \mid S_t = s) = P(S_{t+1+\Delta} = s' \mid S_{t+\Delta} = s)$$

$$P(O_t = o \mid S_t = s) = P(O_{t+\Delta} = o \mid S_{t+\Delta} = s)$$

* Joint Distribution

$$P(\underbrace{S_1, S_2, \ldots, S_{T-1}, S_T}_{\vec{S}}, \underbrace{O_1, O_2, \ldots, O_{T-1}, O_T}_{\vec{O}})$$

$$= \underbrace{P(S_1)}_{\text{initial state}} \left\{ \prod_{t=2}^{T} P(S_t \mid S_{t-1}) \right\} \left\{ \prod_{t=1}^{T} P(O_t \mid S_t) \right\}$$

* Parameters

$a_{ij} = P(S_{t+1} = j \mid S_t = i)$ $n \times n$ transition matrix

$b_i(k) = b_{ik} = P(O_t = k \mid S_t = i)$ $n \times m$ emission matrix

$\pi_i = P(S_1 = i)$ initial state distribution. ($n \times 1$ vector)

* Key Computations / Questions in HMMs

1) how to compute likelihood $P(O_1, O_2, \cdots, O_T)$?

2) how to compute most likely hidden state sequence ?

$$\underset{S_1, S_2, \cdots, S_T}{\arg\max} P(S_1, S_2, \cdots, S_T \mid O_1, O_2, \cdots, O_T)$$

Inference

3) How to estimate parameters $\{\pi_i, a_{ij}, b_{ik}\}$ that maximize $P(O_1, O_2, \cdots, O_T)$ & maybe $\underset{\text{sequence } i}{\prod} P(O_1^{(i)}, O_2^{(i)}, \cdots, O_T^{(i)})$?

Learning