CSE 250A. Principles of AI

Probabilistic Reasoning and Decision-Making

Lecture 13 – More latent variable models

Lawrence Saul
Department of Computer Science and Engineering
University of California, San Diego

Fall 2021

Outline

EM algorithm

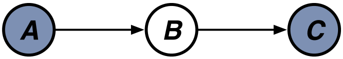
- **Updates**

root nodes	$P(X_i = x)$	$\longleftarrow$	$\frac{1}{T}\sum_t P(X_i = x \mid V_t = v_t)$
nodes with parents	$P(X_i = x \mid \mathrm{pa}_i = \pi)$	$\longleftarrow$	$\frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi \mid V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi \mid V_t = v_t)}$

- **Convergence**

  Each iteration of these updates is guaranteed to increase the log-likelihood $\sum_t \log P(V_t)$ (except at stationary points).

Example 1



Incomplete data $\{(a_t, c_t)\}_{t=1}^{T}$
A and C are observed.
B is hidden.

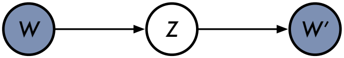
- **E-step** (Inference)

$$P(b|a_t, c_t) \;=\; \frac{P(c_t|b)\,P(b|a_t)}{\sum_{b'} P(c_t|b')\,P(b'|a_t)}$$

- **M-step** (Learning)

$$\begin{aligned} P(a) \;&=\; \frac{1}{T}\,\mathrm{count}(A=a) \\ P(b|a) \;&\longleftarrow\; \frac{\sum_t I(a, a_t)\,P(b|a_t, c_t)}{\sum_t I(a, a_t)} \\ P(c|b) \;&\longleftarrow\; \frac{\sum_t I(c, c_t)\,P(b|a_t, c_t)}{\sum_t P(b|a_t, c_t)} \end{aligned}$$

Application 1: word clustering



$$w, w' \in \{1, 2, \ldots, V\}$$
$$z \in \{1, 2, \ldots, k\} \text{ where } k \ll V$$

1	as cents made make take
2	ago day earlier Friday Monday month quarter reported said Thursday trading Tuesday Wednesday ⟨...⟩
3	even get to
4	based days down home months up work years ⟨%⟩
5	those ⟨,⟩ ⟨—⟩
6	⟨.⟩ ⟨?⟩
7	eighty fifty forty ninety seventy sixty thirty twenty ⟨(⟩ ⟨·⟩
8	can could may should to will would
9	about at just only or than ⟨&⟩ ⟨;⟩
10	economic high interest much no such tax united well
11	president
12	because do how if most say so then think very what when where
13	according back expected going him plan used way
15	don't I people they we you
16	Bush company court department more officials police retort spokesman
17	former the
18	American big city federal general house military national party political state union York

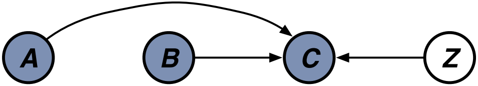
19	billion hundred million nineteen
20	did ⟨"⟩ ⟨'⟩
21	but called San ⟨:⟩ ⟨start-of-sentence⟩
22	bank board chairman end group members number office out part percent price prices rate sales shares use
23	a an another any dollar each first good her his its my old our their this
24	long Mr. year
25	business California case companies corporation dollars incorporated industry law money thousand time today war week ⟨)⟩ ⟨unknown⟩
26	also government he it market she that there which who
27	A. B. C. D. E. F. G. I. L. M. N. P. R. S. T. U.
28	both foreign international major many new oil other some Soviet stock these west world
29	after all among and before between by during for from in including into like of off on over since through told under until while with
30	eight fifteen five four half last next nine oh one second seven several six ten third three twelve two zero ⟨-⟩
31	are be been being had has have is it's not still was were
32	chief exchange news public service trade

$k = 32$

Outline

1. Review
2. $\boxed{\text{Mixture models}}$
3. Noisy-OR models
4. Hidden Markov models

Example 2 — Inference

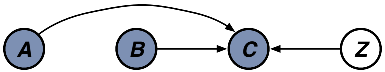


A, B, C are observed.
Z is hidden.

Posterior probability

$$
\begin{aligned}
P(Z|A,B,C) &= \frac{P(C|Z,A,B)\,P(Z|A,B)}{P(C|A,B)} && \boxed{\textbf{Bayes rule}} \\
&= \frac{P(C|Z,A,B)\,P(Z)}{P(C|A,B)} && \boxed{\textbf{marginal independence}} \\
&= \frac{P(C|Z,A,B)\,P(Z)}{\sum_z P(C|Z\!=\!z,A,B)\,P(Z\!=\!z)} && \boxed{\textbf{normalization}}
\end{aligned}
$$

Example 2 — Learning



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

- **Log (conditional) likelihood**

$$
\begin{aligned}
\mathcal{L} &= \sum_t \log P(c_t|a_t, b_t) \\
&= \sum_t \log \sum_z P(z, c_t|a_t, b_t) \quad \boxed{\textbf{marginalization}} \\
&= \sum_t \log \sum_z P(z|a_t, b_t)\, P(c_t|z, a_t, b_t) \quad \boxed{\textbf{product rule}} \\
&= \sum_t \log \sum_z P(z)\, P(c_t|z, a_t, b_t) \quad \boxed{\textbf{marginal independence}}
\end{aligned}
$$

- **EM update**

$$P(z) \leftarrow \frac{1}{T} \sum_t P(z|a_t, b_t, c_t) \quad \boxed{\textbf{root node}}$$

Application

- **Markov models**

  Let $P_1(w)$ be a unigram model.
  Let $P_2(w'|w)$ be a bigram model.
  Let $P_3(w''|w,w')$ be a trigram model.

- **Linear interpolation of Markov models**

$$\underbrace{P_{\text{mix}}(w_\ell|w_{\ell-1},w_{\ell-2})}_{\text{mixture model}} = \lambda_1 P_1(w_\ell) + \lambda_2 P_2(w_\ell|w_{\ell-1}) + \lambda_3 P_3(w_\ell|w_{\ell-1},w_{\ell-2})$$

  We require $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$.
  This ensures a properly normalized distribution.
  But how to estimate $\lambda_1, \lambda_2, \lambda_3$?

Methodology

- **What to do**

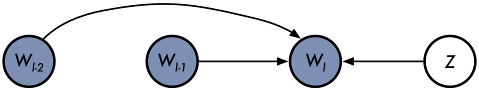
  Use corpus A to estimate $P_1(w)$, $P_2(w'|w)$, $P_3(w''|w, w')$.
  Use corpus B to estimate λ_1, λ_2, λ_3 (only).
  Use corpus C to evaluate the mixture model $P_{\text{mix}}(w''|w, w')$.

- **What not to do**

  Do not use corpus A to estimate $\lambda_1, \lambda_2, \lambda_3$.
  Otherwise you will find $\lambda_3 = 1$ and $\lambda_1 = \lambda_2 = 0$.

  Do not use corpus C to estimate any parameters.
  That would bias the evaluation.

Latent variable model (con't)



Predicting the next word

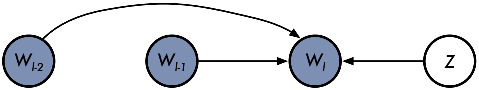
$$P(w_\ell | w_{\ell-1}, w_{\ell-2})$$

$$= \sum_z P(z, w_\ell | w_{\ell-1}, w_{\ell-2}) \quad \boxed{\textbf{marginalization}}$$

$$= \sum_z P(z | w_{\ell-1}, w_{\ell-2})\, P(w_\ell | w_{\ell-1}, w_{\ell-2}, z) \quad \boxed{\textbf{product rule}}$$

$$= \sum_z P(z)\, P(w_\ell | w_{\ell-1}, w_{\ell-2}, z) \quad \boxed{\textbf{marginal independence}}$$

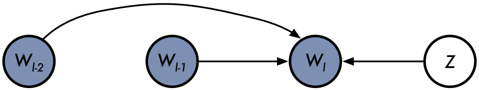
Latent variable model (con't)



Predicting the next word

$$
\begin{aligned}
& P(w_\ell | w_{\ell-1}, w_{\ell-2}) \\
&= \sum_z P(z, w_\ell | w_{\ell-1}, w_{\ell-2}) \quad \boxed{\textbf{marginalization}} \\
&= \sum_z P(z | w_{\ell-1}, w_{\ell-2})\, P(w_\ell | w_{\ell-1}, w_{\ell-2}, z) \quad \boxed{\textbf{product rule}} \\
&= \sum_z P(z)\, P(w_\ell | w_{\ell-1}, w_{\ell-2}, z) \quad \boxed{\textbf{marginal independence}} \\
&= \lambda_1 P_1(w_\ell) \;+\; \lambda_2 P_2(w_\ell | w_{\ell-1}) \;+\; \lambda_3 P_3(w_\ell | w_{\ell-1}, w_{\ell-2}) \quad \boxed{\textbf{!!}}
\end{aligned}
$$

Learning the mixing coefficients



- **Mixing the *n*-gram models**

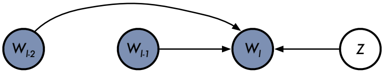
  We learn $P_1(w)$, $P_2(w'|w)$, and $P_3(w''|w,w')$ from corpus A.
  We learn $\lambda_1, \lambda_2, \lambda_3$ from corpus B.

- **EM update for mixing coefficients**

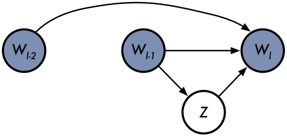
$$\underbrace{P(Z=i)}_{\lambda_i} \longleftarrow \frac{1}{L_B}\sum_{\ell=1}^{L_B} P(Z=i|w_\ell, w_{\ell-1}, w_{\ell-2})$$

Here, L_B is the length in words of corpus B.

Extensions of this model



EM may seem like overkill to learn just 3 numbers $\lambda_1, \lambda_2, \lambda_3$.
But this model can be extended in interesting ways ...



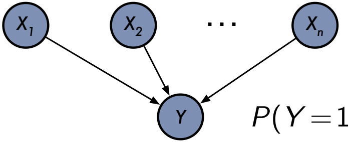
Now the coefficients depend on the previous word:

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

This model has $3V$ coefficients where V is the vocabulary size.
But the EM algorithm hardly changes.

Outline

Example 3: Noisy-OR



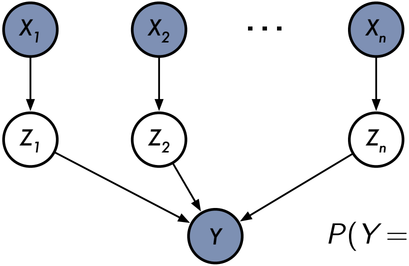
$x_i \in \{0, 1\}$

$P(Y = 1 | x_1, x_2, \ldots, x_n) = 1 - \prod_{i=1}^{n} (1 - p_i)^{x_i}$

The log (conditional) likelihood is $\sum_t \log P(y_t | x_t)$.
How to estimate parameters $p_i \in [0, 1]$ that maximize this?

1. Gradient ascent
2. Newton's method
3. **EM — but how? Isn't the data complete?**

EM for noisy-OR



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

$z_i \in \{0, 1\}$
$P(Z_i\!=\!1|x_i) = p_i x_i$

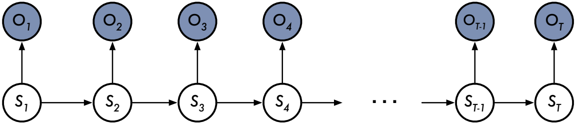
$$P(Y\!=\!1|Z_1, \ldots, Z_n) = \mathbf{OR}(Z_1, \ldots, Z_n)$$

$\boxed{\textbf{HW 6}}$

First you will show that this model is equivalent to noisy-OR.
Then you will derive the EM updates for $p_i \in [0, 1]$.

Outline

Hidden Markov models (HMMs)



- **Random variables**

$$\begin{array}{ll} S_t \in \{1, 2, \ldots, n\} & \text{hidden state at time } t \\ O_t \in \{1, 2, \ldots, m\} & \text{observation at time } t \end{array}$$

- **States versus observations**

  Each observation O_t is a noisy, partial reflection of the true underlying (but hidden) state S_t of the world at time t.

  What makes this model so useful?

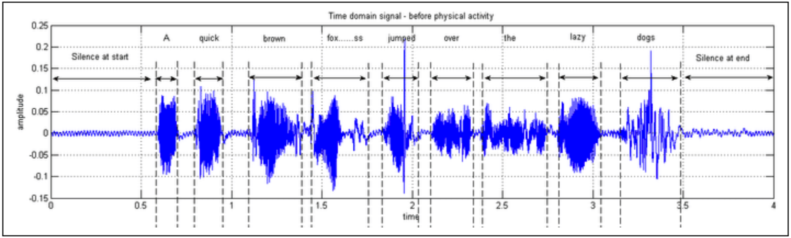
Housetraining a puppy

This is Lilo.
She's a chihuahua-terrier.

$O_t \in \{\texttt{sleeping}, \texttt{eating}, \texttt{barking}, \texttt{waiting by door}, \texttt{etc.}\}$
$S_t \in \{\texttt{playful}, \texttt{hungry}, \texttt{tired}, \texttt{ready to burst}\}$

Does Lilo need to go outside?
What is $P(s_t|o_1, o_2, \ldots, o_t)$?

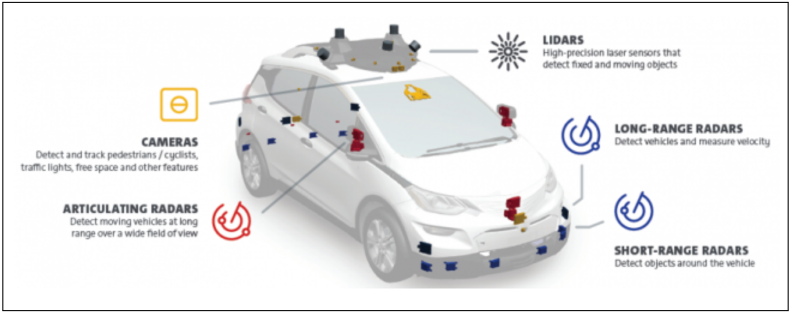
Speech recognition



O_t is the acoustic feature vector for windowed speech at time t.
S_t is the unit of language (e.g., phoneme) being uttered at time t.

What did I just hear?
What is $\text{argmax}_{s_1,s_2,\ldots,s_T} P(s_1, s_2, \ldots, s_T | o_1, o_2, \ldots, o_T)$**?**

Autonomous navigation

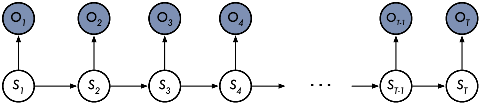


https://www.extremetech.com/computing/305691-the-future-of-sensors-for-self-driving-cars-all-roads-all-conditions

O_t encodes the sensor readings at time t.
S_t encodes the nearby vehicles and pedestrians at time t.

Monitoring the road: what is $P(s_t|o_1, o_2, \ldots, o_t)$?

HMMs as belief networks



- **Conditional independence assumptions**

$$\begin{aligned} P(S_t|S_1, S_2, \ldots, S_{t-1}) &= P(S_t|S_{t-1}) \\ P(O_t|S_1, S_2, \ldots, S_T) &= P(O_t|S_t) \end{aligned}$$

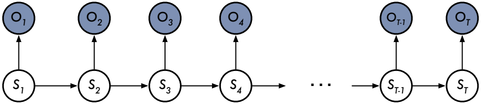
- **CPTs are shared across time**

$$\begin{aligned} P(S_t = s'|S_{t-1} = s) &= P(S_{t+1} = s'|S_t = s) \\ P(O_t = o|S_t = s) &= P(O_{t+1} = o|S_{t+1} = s) \end{aligned}$$

- **Joint distribution**

$P(\, S_1, \ldots, S_T$

HMMs as belief networks



- **Conditional independence assumptions**

$$\begin{aligned} P(S_t|S_1, S_2, \ldots, S_{t-1}) &= P(S_t|S_{t-1}) \\ P(O_t|S_1, S_2, \ldots, S_T) &= P(O_t|S_t) \end{aligned}$$

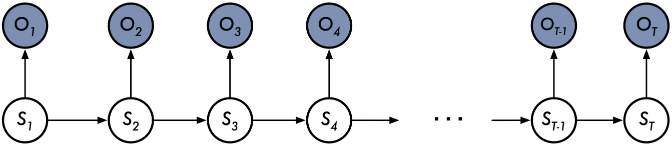
- **CPTs are shared across time**

$$\begin{aligned} P(S_t = s'|S_{t-1} = s) &= P(S_{t+1} = s'|S_t = s) \\ P(O_t = o|S_t = s) &= P(O_{t+1} = o|S_{t+1} = s) \end{aligned}$$

- **Joint distribution**

$$P(\underbrace{S_1, \ldots, S_T}_{\vec{S}}, \underbrace{O_1, \ldots, O_T}_{\vec{O}}) = P(S_1)\, P(O_1|S_1) \prod_{t=2}^{T} \Big[P(S_t|S_{t-1})\, P(O_t|S_t) \Big]$$

Parameters of HMMs

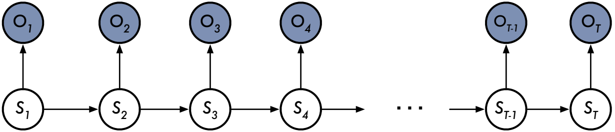


$$a_{ij} = P(S_{t+1}=j|S_t=i)$$ $\boxed{n\times n \textbf{ transition matrix}}$

$$b_{ik} = P(O_t=k|S_t=i)$$ $\boxed{n\times m \textbf{ emission matrix}}$

$$\pi_i = P(S_1=i)$$ $\boxed{n\times 1 \textbf{ initial state distribution}}$

Next lecture: key computations in HMMs



POLYTREE!

Inference

1. How to compute the likelihood $P(o_1, o_2, \ldots, o_T)$?
2. How to compute the most likely state sequence $\text{argmax}_{\vec{s}}\, P(\vec{s}|\vec{o})$?
3. How to update beliefs by computing $P(s_t|o_1, o_2, \ldots, o_t)$?

Learning

How to estimate parameters $\{\pi_i, a_{ij}, b_{ik}\}$ that maximize the log-likelihood of observed sequences?