EM for HMMs a.k.a. The Baum-Welch Algorithm

Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

April 11th, 2007

The general learning problem with missing data

- Marginal likelihood – **x** is observed, **z** is missing:

$$\ell(\theta : \mathcal{D}) = \log \prod_{j=1}^{m} \underbrace{P(\mathbf{x}_j \mid \theta)}_{\text{observed parts}}$$

$$= \sum_{j=1}^{m} \log P(\mathbf{x}_j \mid \theta)$$

$$= \sum_{j=1}^{m} \log \sum_{\mathbf{z}} P(\mathbf{x}_j, \mathbf{z} \mid \theta)$$

sum over (marginalize out) hidden vars

EM is coordinate ascent

$$\overbrace{\ell(\theta : \mathcal{D})}^{\text{marginal likelihood}} \geq \underbrace{F(\theta, Q)}_{\uparrow \text{max.}} = \sum_{j=1}^{m} \sum_{\mathbf{z}} Q(\mathbf{z} \mid \mathbf{x}_j) \log \frac{P(\mathbf{z}, \mathbf{x}_j \mid \theta)}{Q(\mathbf{z} \mid \mathbf{x}_j)}$$

- **M-step**: Fix Q, maximize F over θ (a lower bound on $\ell(\theta : \mathcal{D})$):

$$\ell(\theta : \mathcal{D}) \geq F(\theta, Q^{(t)}) = \sum_{j=1}^{m} \sum_{\mathbf{z}} Q^{(t)}(\mathbf{z} \mid \mathbf{x}_j) \log P(\mathbf{z}, \mathbf{x}_j \mid \theta) + m.H(Q^{(t)})$$

Expected counts.

- **E-step**: Fix θ, maximize F over Q:

$$\ell(\theta^{(t)} : \mathcal{D}) \geq F(\theta^{(t)}, Q) = \ell(\theta^{(t)} : \mathcal{D}) - m \sum_{j=1}^{m} KL\left(Q(\mathbf{z} \mid \mathbf{x}_j) || P(\mathbf{z} \mid \mathbf{x}_j, \theta^{(t)})\right)$$

  - "Realigns" F with likelihood:

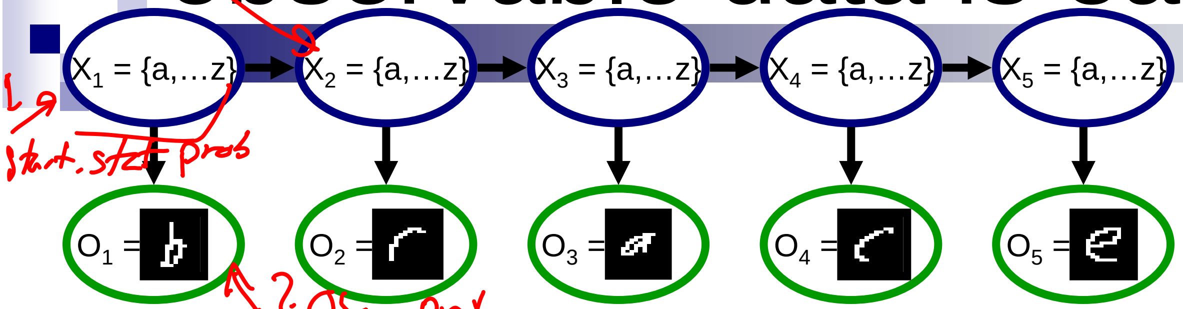
$$F(\theta^{(t)}, Q^{(t+1)}) = \ell(\theta^{(t)} : \mathcal{D})$$

What you should know about EM

- K-means for clustering:
  - algorithm
  - converges because it's coordinate ascent
- EM for mixture of Gaussians:
  - How to "learn" maximum likelihood parameters (locally max. like.) in the case of unlabeled data
- Be happy with this kind of probabilistic analysis
- Remember, E.M. can get stuck in local minima, and empirically it <u>DOES</u>
- EM is coordinate ascent
- General case for EM

Learning HMMs from fully observable data is easy

3 transition model



1 start state prob

2 obs. prob.

Learn 3 distributions:

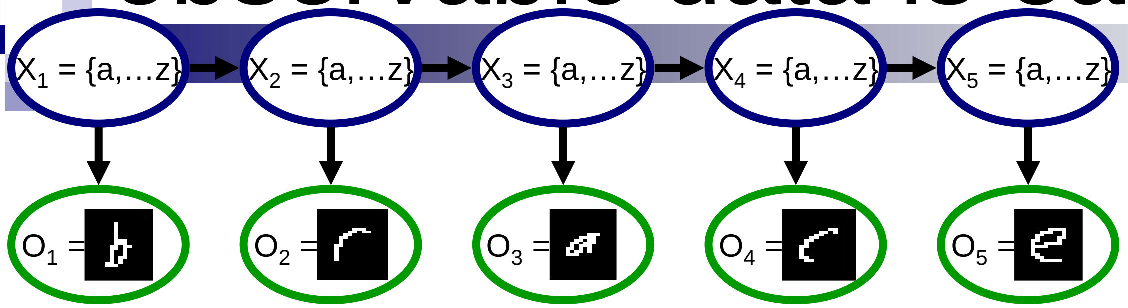
$X_1 =$ anything

1. $$P(X_1 = a) = \frac{\text{Count}(X_1 = a)}{m = \text{Count}(X_1 = ?)}$$

2. $$P(O_i = 17 \mid X_i = b) = \frac{\text{Count}(X_i = b, O_i = 17)}{\text{Count}(X_i = b)}$$

3. $$P(X_i = c \mid X_{i-1} = g) = \frac{\text{Count}(X_{i-1} = g, X_i = c)}{\text{Count}(X_{i-1} = g, X_i = ?)}$$

Learning HMMs from fully observable data is easy



Learn 3 distributions:

$$P(X_1 = a) = \frac{\text{Count}(\#\text{ first letter was } a)}{N = \text{data set size}}$$

select training data where letter was a

$$P(O_i = \text{pixel 17 is white} \mid X_i = a) = \frac{\text{Count}(\text{pixel 17 was white}, X_i = a)}{\ldots}$$

any ... in

$$P(X_i = a \mid X_{i-1} = b)$$

What if **O** is observed, but **X** is hidden

Log likelihood for HMMs when X is hidden

$O = (o_1, \ldots, o_n)$

$X = (x_1, \ldots, x_n)$

for m sequences

$\sum_{j=1}^{m} \log P(o^{(j)}/\theta)$

- Marginal likelihood – **O** is observed, **X** is missing
  - For simplicity of notation, training data consists of only one sequence:

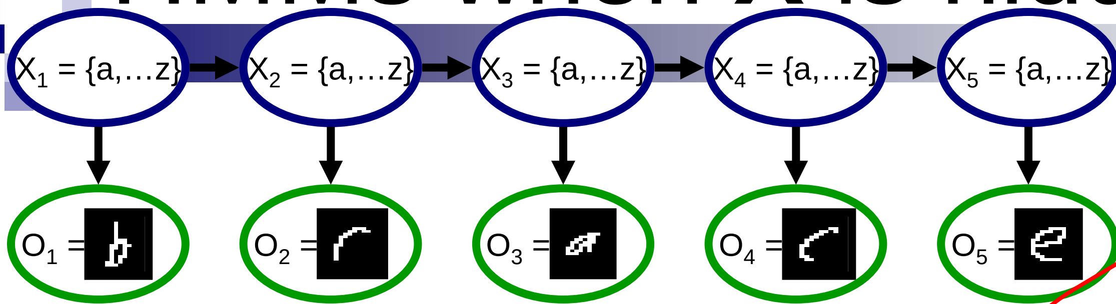
observed

$$\ell(\theta : \mathcal{D}) = \log P(\mathbf{o} \mid \theta)$$
$$= \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o} \mid \theta)$$

  - If there were m sequences:

$$\ell(\theta : \mathcal{D}) = \sum_{j=1}^{m} \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o}^{(j)} \mid \theta)$$

Computing Log likelihood for HMMs when **X** is hidden



if X_i takes on K values

Sum over K^n assignments

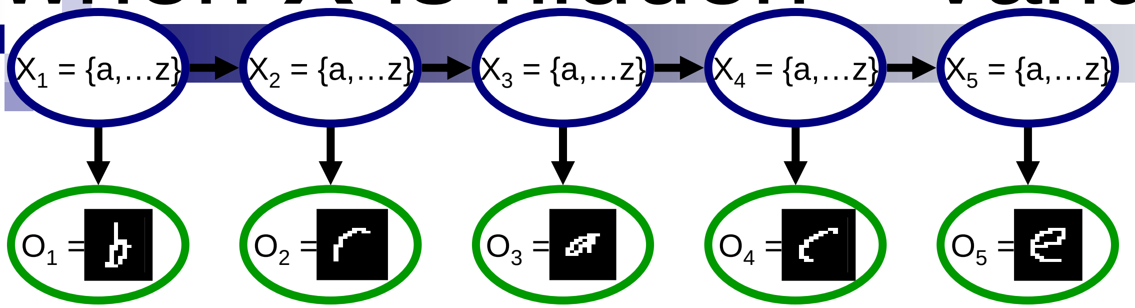
$$\ell(\theta : \mathcal{D}) = \log P(\mathbf{o} \mid \theta)$$
$$= \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o} \mid \theta)$$

$$= \log \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} P(x_1) \cdot P(o_1|x_1) \cdot \prod_{i=2}^{n} P(x_i|x_{i-1}) \cdot P(o_i|x_i)$$

$$= \log \sum_{x_1} \cdots \sum_{x_{n-1}} P(x_1) P(o_1|x_1) \prod_{i=2}^{n-1} P(x_i|x_{i-1}) P(o_i|x_i) \underbrace{\sum_{x_n} P(x_n|x_{n-1}) \cdot P(o_n|x_n)}_{\beta_{n-1}(x_{n-1})}$$

use V.E. to compute $\ell(\theta:D)$ in $O(n)$ time

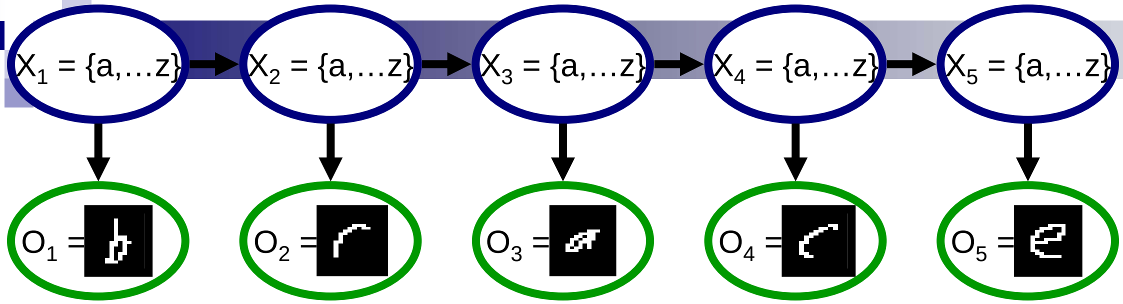
Computing Log likelihood for HMMs when **X** is hidden – variable elimination



- Can compute efficiently with variable elimination:

$$\begin{aligned}\ell(\theta : \mathcal{D}) &= \log P(\mathbf{o} \mid \theta) \\ &= \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o} \mid \theta)\end{aligned}$$

EM for HMMs when **X** is hidden



- E-step: Use inference (forwards-backwards algorithm)

need to compute: $P(X_i, X_{i+1} \mid O_1 \ldots O_n)$, use forwards-backwards

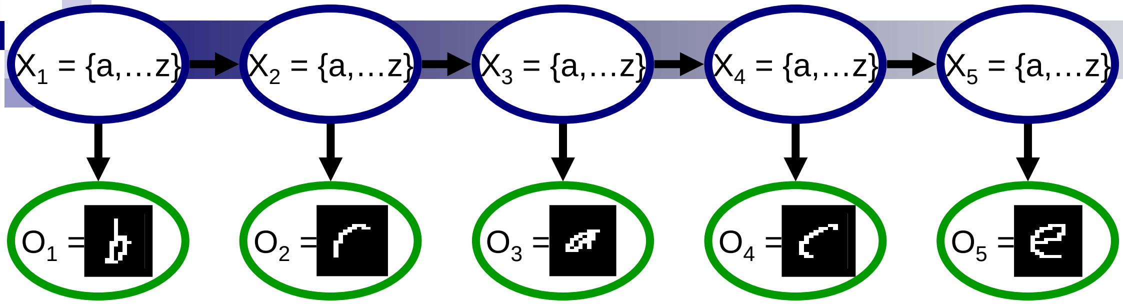
$= Q(X_i, X_{i+1} \mid O_1 \ldots O_n)$ (faster version of Var. elim.)

- M-step: Recompute parameters with weighted data

if fully observable: $\hat{P}(X_1 = a) = \frac{\text{Count}(X_1 = a)}{m}$

if hidden vars: $\hat{P}(X_1 = a) = \frac{\sum_{j=1}^{m} Q(X_1 = a \mid O^{(j)})}{m}$

E-step



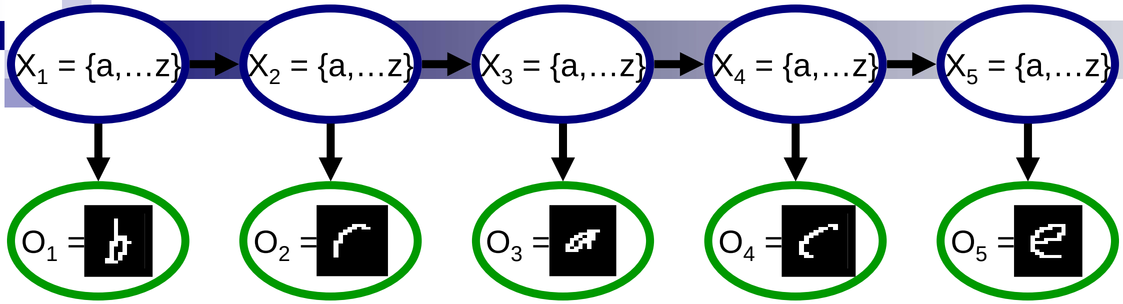
- E-step computes probability of hidden vars **x** given $\mathbf{o}^{(j)}$

j'th training example

$$Q^{(t+1)}(\mathbf{x} \mid \mathbf{o}^{(j)}) = P(\mathbf{x} \mid \mathbf{o}^{(j)}, \theta^{(t)})$$

- Will correspond to inference
  - use forward-backward algorithm!

The M-step



- Maximization step:

$$\theta^{(t+1)} \leftarrow \arg\max_{\theta} \underbrace{\sum_{\mathbf{x}} Q^{(t+1)}(\mathbf{x} \mid \mathbf{o}) \log P(\mathbf{x}, \mathbf{o} \mid \theta)}_{\text{weighted log likelihood}}$$

- Use expected counts instead of counts:
  - If learning requires Count(**x**,**o**)
  - Use $E_{Q(t+1)}$[Count(**x**,**o**)]

$$E_{Q^{(t+1)}}[\text{Count}(X = \{a,b,c\}, O[a][b],[d])] = \frac{\sum_{j=1}^{m} Q^{(t+1)}(X=\{a,b,c\} \mid o^{(j)} = [a][b][d])}{m}$$

Decomposition of likelihood revisited

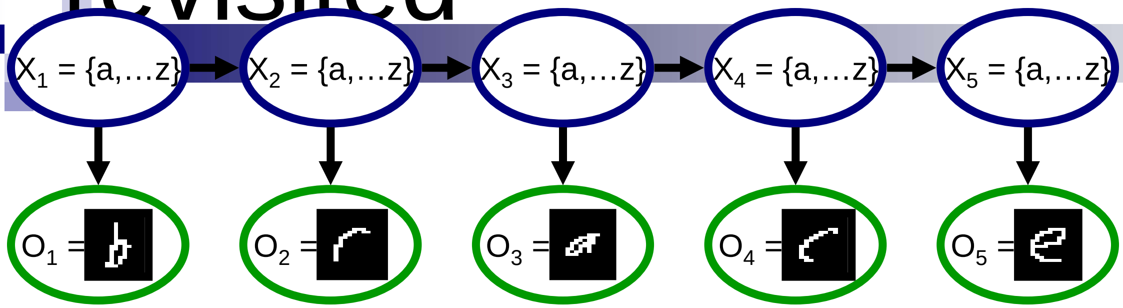
$P(X_1) \leftarrow \theta_{X_1}$

$P(O_i \mid X_i) \leftarrow \theta_{O|X}$

$P(X_i \mid X_{i-1}) \leftarrow \theta_{X_i|X_{i-1}}$

$\log a.b = \log a + \log b$

$\theta = (\theta_{X_1}, \theta_{X_i|X_{i-1}}, \theta_{O|X})$



- Likelihood optimization decomposes:

$$\max_\theta \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}) \log P(\mathbf{x}, \mathbf{o} \mid \theta) =$$

$$\max_\theta \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}) \log \overbrace{P(x_1 \mid \theta_{X_1}) P(o_1 \mid x_1, \theta_{O|X}) \prod_{t=2}^{n} P(x_t \mid x_{t-1}, \theta_{X_t|X_{t-1}}) P(o_t \mid x_t, \theta_{O|X})}^{\text{HMM structure}}$$

$$= \underbrace{\max_{\theta}}_{\max_{\theta_{X_1}} \max_{\theta_{O|X}} \max_{\theta_{X_t|X_{t-1}}}} \sum_x Q(x|o) \left[\log P(x_1|\theta_{X_1}) + \sum_{t=1}^{n} \log P(o_t|x_t, \theta_{O|X}) + \sum_{t=2}^{n} \log P(x_t|x_{t-1}, \theta_{X_t|X_{t-1}}) \right]$$

$$= \underbrace{\left[\max_{\theta_{X_1}} \sum_x Q(x|o) \log P(x_1|\theta_{X_1}) \right]}_{\theta_{X_1} \text{ start state prob.}} + \underbrace{\left[\max_{\theta_{O|X}} \sum_x Q(x|o) \sum_{t=1}^{n} \log P(o_t|x_t, \theta_{O|X}) \right]}_{\theta_{O|X}} + \underbrace{\left[\max_{\theta_{X_t|X_{t-1}}} \sum_x Q(x|o) \sum_{t=2}^{n} \log P(x_t|x_{t-1}, \theta_{X_t|X_{t-1}}) \right]}_{\theta_{X_t|X_{t-1}}}$$

∀ sequences:

Starting state probability $P(X_1)$

- Using expected counts
  - $P(X_1=a) = \theta_{X1=a}$

$Q(\cancel{X}|o)$ explicitly, need K^n-1 parameters

derivation for one sequence

$$\max_{\theta_{X_1}} \sum_{j=1}^{m} \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}^{(j)}) \log P(x_1 \mid \theta_{X_1}) = \max_{\theta_{X_1}} \underbrace{\sum_{x}}_{\sum_{x_1} \cdots \sum_{x_n}} \overbrace{Q(x_1|o) \cdot Q(x_2..x_n|o,x_1)}^{\text{chain rule}} \cdot \log P(x_1|\theta_{X_1})$$

$$= \max_{\theta_{X_1}} \sum_{j=1}^{m} \sum_{x_1} Q(x_1|o^{(j)}) \log P(x_1|\theta_{X_1}) \cdot \underbrace{\sum_{x_2} \cdots \sum_{x_n} Q(x_2..x_n|o^{(j)},x_1)}_{1 \text{ (it's a prob.)}}$$

$$= \max_{\theta_{X_1}} \sum_{j=1}^{m} \sum_{x_1} Q(x_1|o^{(j)}) \log P(x_1|\theta_{X_1})$$

only need $Q(\cancel{X}_1|o)$ ← $o \in$ image in dataset

→ K-1 parameters

→ argmax →

learning from weighted data

argmax ∀ all sequences:

$$\boxed{\theta_{X_1=a} = \frac{\sum_{j=1}^{m} Q(X_1 = a \mid \mathbf{o}^{(j)})}{m}}$$

Transition probability $P(X_t|X_{t-1})$

$\log \prod_t a_t = \sum_t \log a_t$

- Using expected counts
  - $P(X_t=a|X_{t-1}=b) = \theta_{Xt=a|Xt-1=b}$

$$\max_{\theta_{X_t|X_{t-1}}} \sum_{j=1}^{m} \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}^{(j)}) \log \prod_{t=2}^{n} P(x_t \mid x_{t-1}, \theta_{X_t|X_{t-1}}) = \max_{\theta_{X_t|X_{t-1}}} \sum_{x} Q(x|o) \sum_{t=2}^{n} \log P(x_t|x_{t-1}, \theta_{X_t|X_{t-1}})$$

$$= \max_{\theta_{X_t|X_{t-1}}} \sum_{j=1}^{m} \sum_{t=2}^{n} \sum_{x} Q(x|o^{(j)}) \log P(x_t|x_{t-1}, \theta_{X_t|X_{t-1}})$$

$$= \max_{\theta_{X_t|X_{t-1}}} \sum_{j=1}^{m} \sum_{t=2}^{n} \sum_{x_t} \sum_{x_{t-1}} Q(x_{t-1}, x_t|o^{(j)}) \log P(x_t|x_{t-1}, \theta_{X_t|X_{t-1}})$$

$$\sum_{x_1 \ldots x_{t-2}} \sum_{x_{t+1} \ldots x_n} Q(x_1 \ldots x_{t-2}, x_{t+1}, \ldots x_n | o^{(j)})$$

only need $\forall t \in 2\text{-}n$, $Q(x_{t-1}, x_t | o)$

$\equiv (n-1)(k^2-1)$ parameters

argmax $\rightarrow \theta_{X_t|X_{t-1}}$

$$\boxed{\theta_{X_t=a|X_{t-1}=b} = \frac{\sum_{j=1}^{m}\sum_{t=2}^{n} Q(X_t=a, X_{t-1}=b \mid \mathbf{o}^{(j)})}{\sum_{j=1}^{m}\sum_{t=2}^{n}\sum_{i=1}^{k} Q(X_t=i, X_{t-1}=b \mid \mathbf{o}^{(j)})}}$$

Observation probability $P(O_t|X_t)$

- Using expected counts
  - $P(O_t=a|X_t=b) = \theta_{Ot=a|Xt=b}$

$$\max_{\theta_{O|X}} \sum_{j=1}^{m} \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}^{(j)}) \log \prod_{t=1}^{n} P(o_t^{(j)} \mid x_t, \theta_{O|X}) = \max_{\theta_{O|X}} \sum_{j=1}^{m} \sum_{t=1}^{n} \sum_{x} Q(x \mid o^{(j)}) \log P(o_t^{(j)} \mid x_t, \theta_{O|X})$$

$$= \max_{\theta_{O|X}} \sum_{j=1}^{m} \sum_{t=1}^{n} \sum_{x_t} Q(x_t \mid o^{(j)}) \log P(o_t^{(j)} \mid x_t, \theta_{O|X})$$

only need marginals $Q(x_t \mid o)$

for a total of $n \cdot (K-1)$ params

argmax

only count positions where $O_t = a$

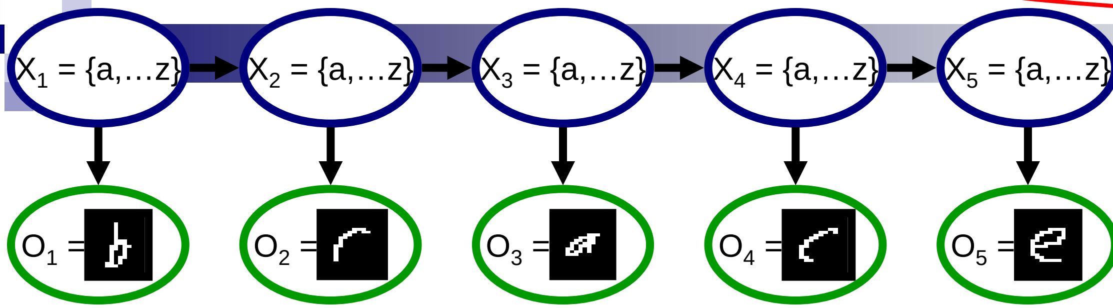
$$\theta_{O_t=a|X_t=b} = \frac{\sum_{j=1}^{m}\sum_{t=1}^{n} \delta(\mathbf{o}_t^{(j)} = a) Q(X_t = b \mid \mathbf{o}^{(j)})}{\sum_{j=1}^{m}\sum_{t=1}^{n} Q(X_t = b \mid \mathbf{o}^{(j)})}$$

j	$O^{(j)}$		
1	B	R	A
2	A	R	G
⋮			

$t=2,\ j=1,\ O_t^{(j)} = \boxed{R}$

E-step revisited

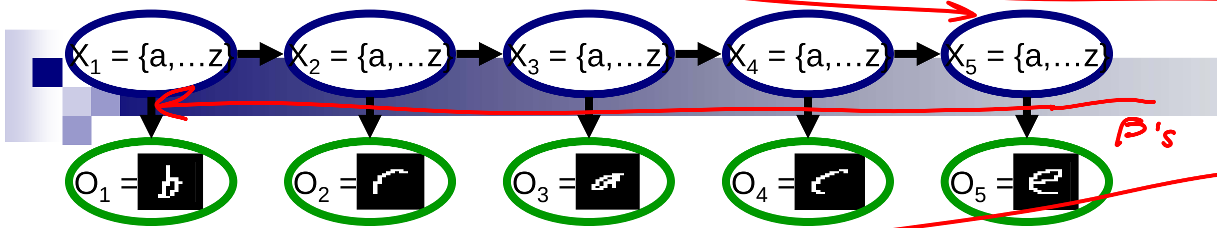
$$Q^{(t+1)}(\mathbf{x} \mid \mathbf{o}) \;=\; P(\mathbf{x} \mid \mathbf{o}, \theta^{(t)})$$



- E-step computes probability of hidden vars **x** given **o**
- Must compute:
  - $Q(x_t=a|\mathbf{o})$ – marginal probability of each position
  - $Q(x_{t+1}=a,x_t=b|\mathbf{o})$ – joint distribution between pairs of positions

use forwards-backwards

α's The forwards-backwards algorithm



$$P(X_i \mid o_{1..n})$$

drop O_1 in notation

- Initialization: $\alpha_1(X_1) = P(X_1)P(o_1 \mid X_1)$
- For i = 2 to n
  - Generate a forwards factor by eliminating X_{i-1}

sum out previous var prob obs transition prob

$$\alpha_i(X_i) = \sum_{x_{i-1}} P(o_i \mid X_i)P(X_i \mid X_{i-1} = x_{i-1})\alpha_{i-1}(x_{i-1})$$

$\alpha_5(a)$
$\alpha_5(b)$
$\vdots$
$\alpha_5(z)$

$\alpha_n(X_n)$ normalized $= P(X_n \mid O_{1:n})$

$\beta_1(X_1)\alpha_1(X_1)$ normalized $= P(X_1 \mid O_{1:n})$

- Initialization: $\beta_n(X_n) = 1$
- For i = n-1 to 1
  - Generate a backwards factor by eliminating X_{i+1}

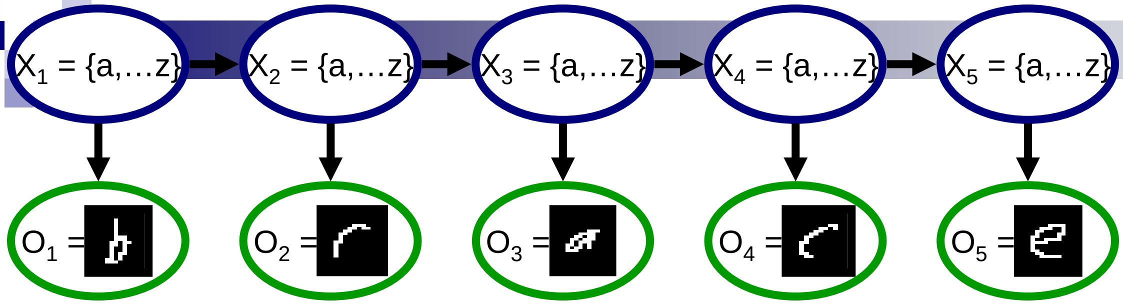
$\forall x_i$

$$\beta_i(\underset{=x_i}{X_i}) = \sum_{x_{i+1}} P(o_{i+1} \mid x_{i+1})P(x_{i+1} \mid \underset{=x_i}{X_i})\beta_{i+1}(x_{i+1})$$

- ∀ i, probability is: $\boxed{P(X_i \mid o_{1..n}) \propto \alpha_i(X_i)\beta_i(X_i)}$

E-step revisited

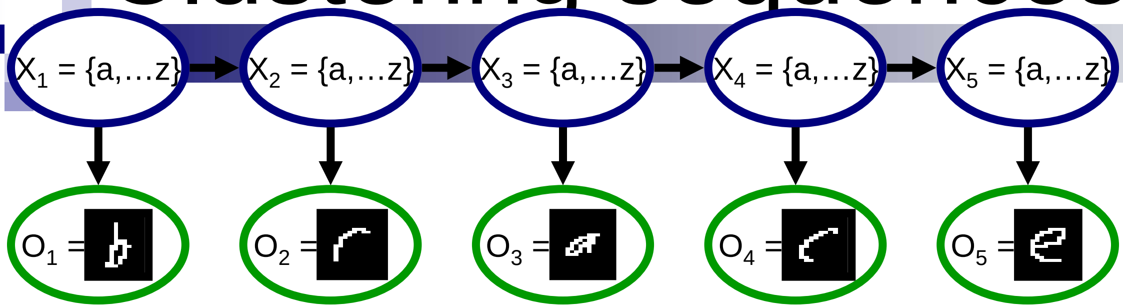
$$Q^{(t+1)}(\mathbf{x} \mid \mathbf{o}) = P(\mathbf{x} \mid \mathbf{o}, \theta^{(t)})$$



- E-step computes probability of hidden vars **x** given **o**
- Must compute:
  - $Q(x_t=a|\mathbf{o})$ – marginal probability of each position
    - Just forwards-backwards!
  - $Q(x_{t+1}=a,x_t=b|\mathbf{o})$ – joint distribution between pairs of positions

see reading [simple eqn.] [maybe homework]

What can you do with EM for HMMs? 1 – Clustering sequences

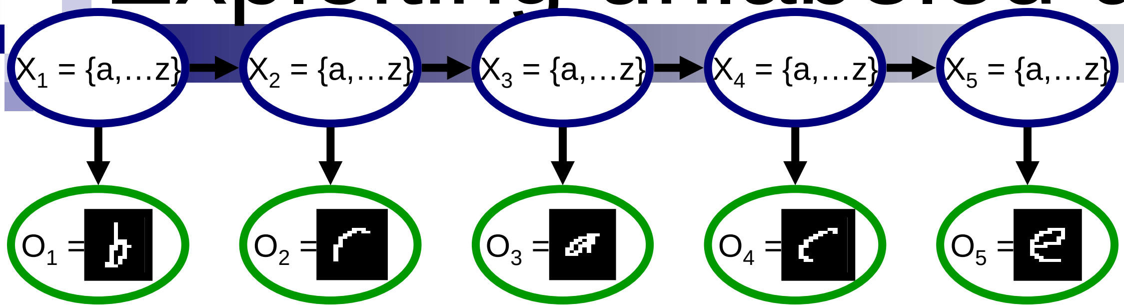


Independent clustering:

Sequence clustering:

speech uses this

What can you do with EM for HMMs? 2 – Exploiting unlabeled data



- Labeling data is hard work → save (graduate student) time by using both labeled and unlabeled data
  - Labeled data:
    - <X=“brace”,O=
  - Unlabeled data:
    - <X=?????,O=

Exploiting unlabeled data in clustering

- A few data points are labeled
  - <x,o>
- Most points are unlabeled
  - <?,o>
- In the E-step of EM:
  - If i'th point is unlabeled:
    - compute $Q(X|o_i)$ as usual
  - If i'th point is labeled:
    - set $Q(X=x|o_i)=1$ and $Q(X\neq x|o_i)=0$

    correct label not label
- M-step as usual

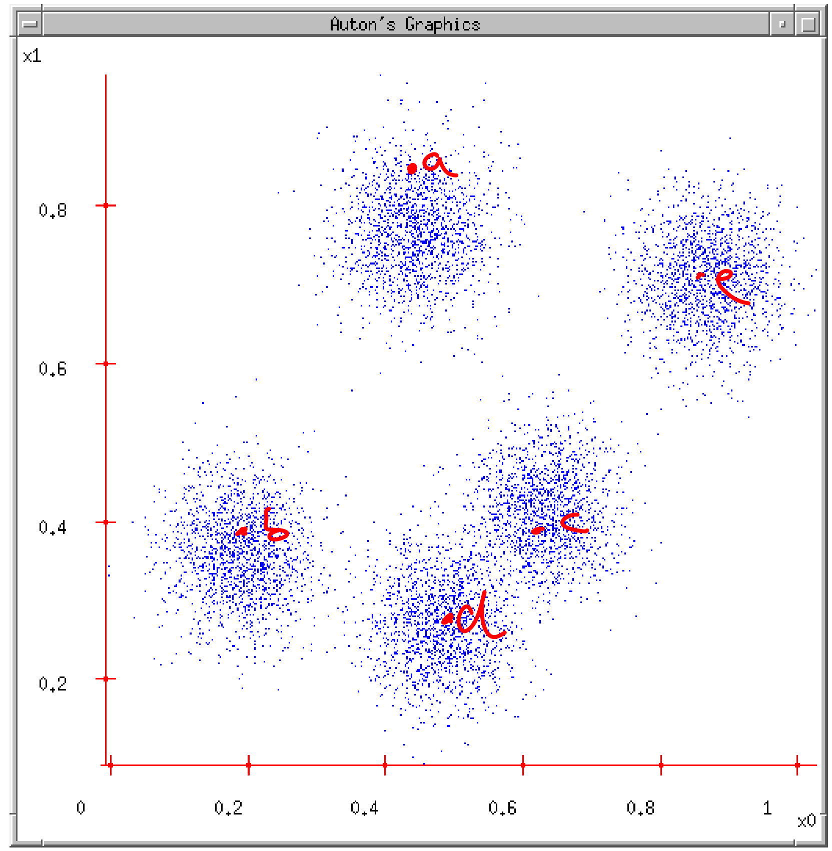
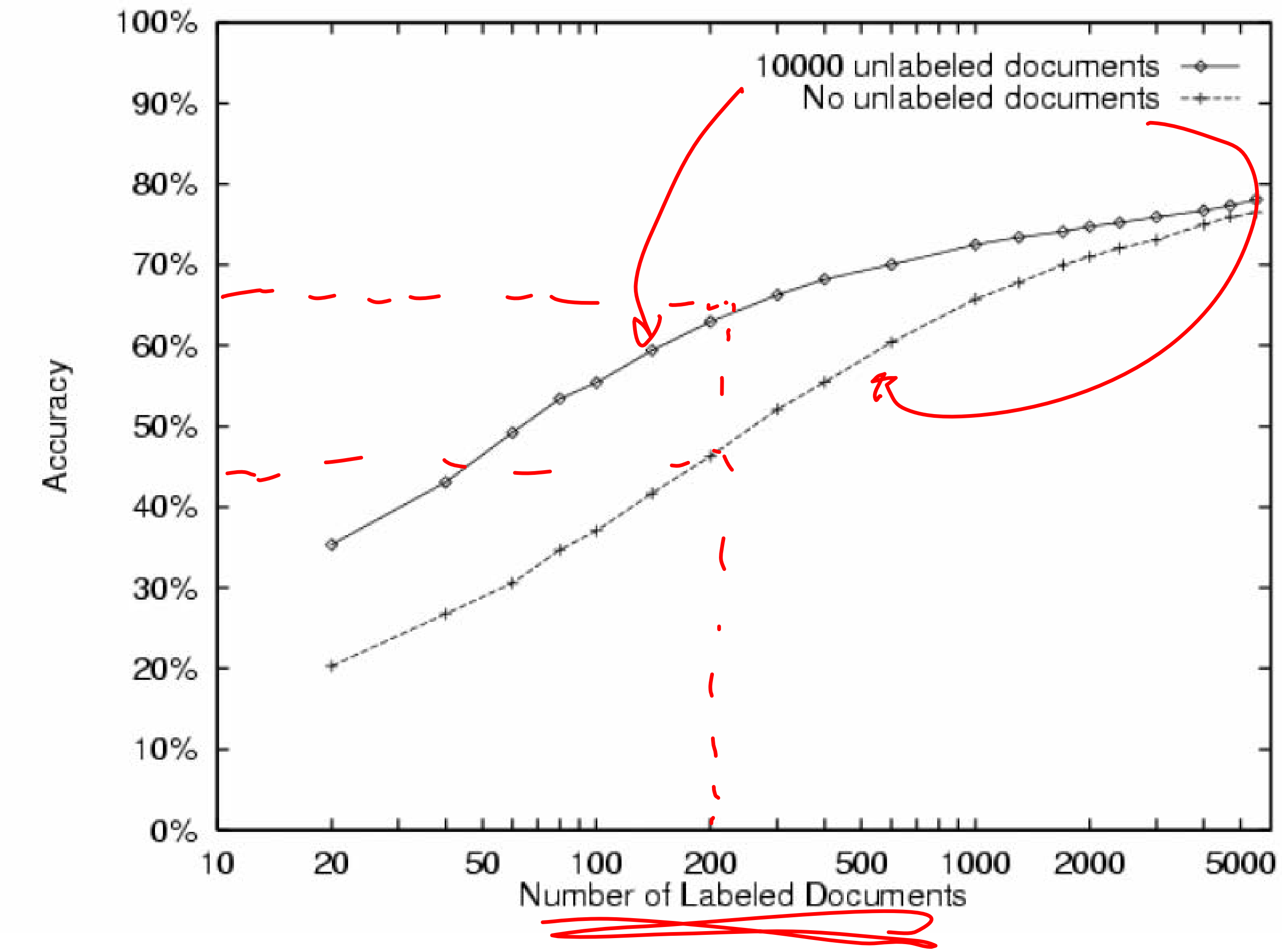


Table 3. Lists of the words most predictive of the course class in the WebKB data set, as they change over iterations of EM for a specific trial. By the second iteration of EM, many common course-related words appear. The symbol D indicates an arbitrary digit.

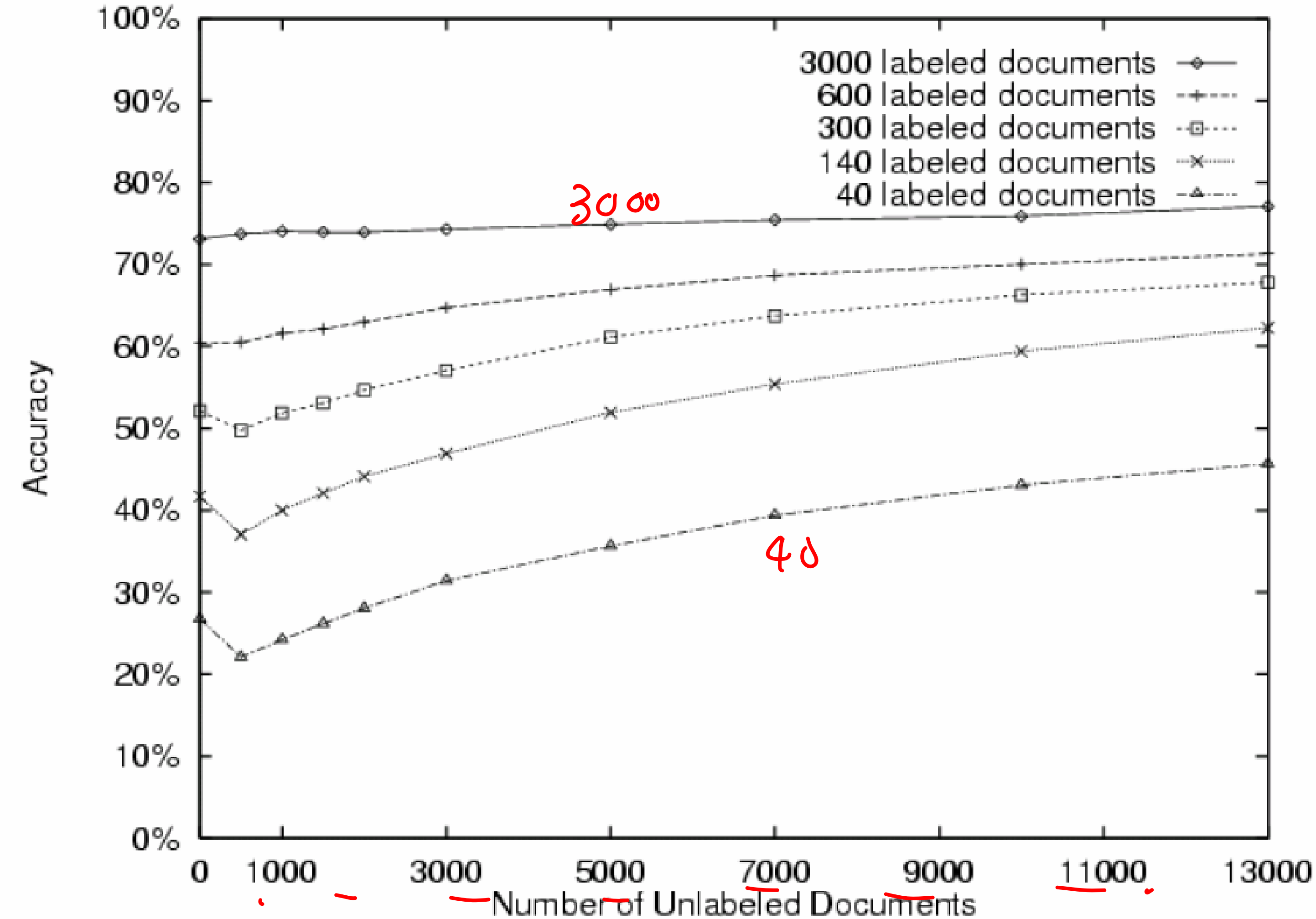
Iteration 0	Iteration 1	Iteration 2
intelligence	DD	D
DD	D	DD
artificial	lecture	lecture
understanding	cc	cc
DDw	$D^\star$	$DD{:}DD$
dist	$DD{:}DD$	due
identical	handout	$D^\star$
rus	due	homework
arrange	problem	assignment
games	set	handout
dartmouth	tay	set
natural	DDam	hw
cognitive	yurttas	exam
logic	homework	problem
proving	kfoury	DDam
prolog	sec	postscript
knowledge	postscript	solution
human	exam	quiz
representation	solution	chapter
field	assaf	ascii

Using one labeled example per class

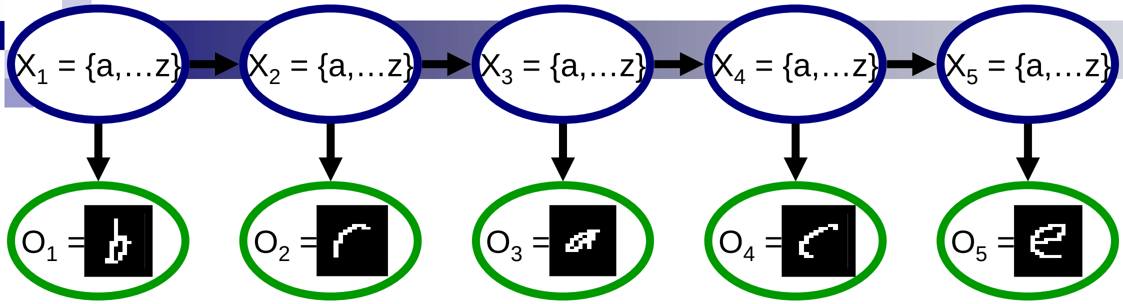
20 Newsgroups data – advantage of adding unlabeled data



20 Newsgroups data – Effect of additional unlabeled data



Exploiting unlabeled data in HMMs



- A few data points are labeled
  - <x,o>
- Most points are unlabeled
  - <?,o>
- In the E-step of EM:
  - If i'th point is unlabeled:
    - compute $Q(X|o_i)$ as usual
  - If i'th point is labeled:
    - set $Q(X=x|o_i)=1$ and $Q(X\neq x|o_i)=0$
- M-step as usual
  - Speed up by remembering counts for labeled data

What you need to know

- Baum-Welch = EM for HMMs
- E-step:
  - Inference using forwards-backwards
- M-step:
  - Use weighted counts
- Exploiting unlabeled data:
  - Some unlabeled data can help classification
  - Small change to EM algorithm
    - In E-step, only use inference for unlabeled data

Acknowledgements

- Experiments combining labeled and unlabeled data provided by Tom Mitchell