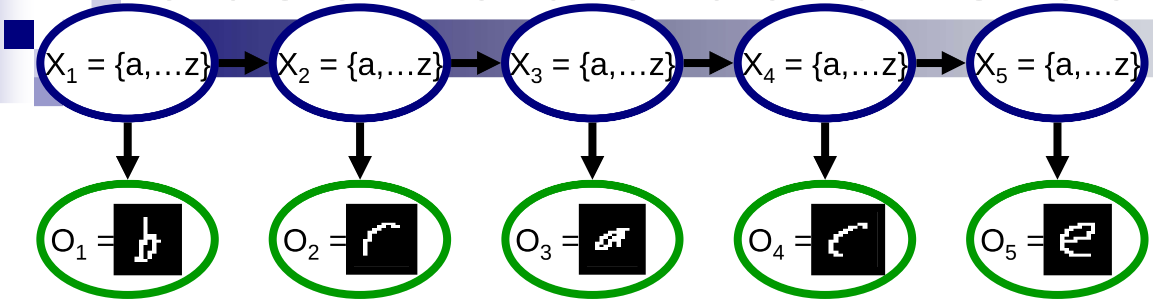
Recommended reading:
"An Introduction to HMMs and Bayesian Networks,"
Z. Ghahramani, *Int. Journal of Pattern Recognition and AI*, 15(1):9-42, (2001)
Especially Section 4

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

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

April 12th, 2006

Learning HMMs from fully observable data is easy



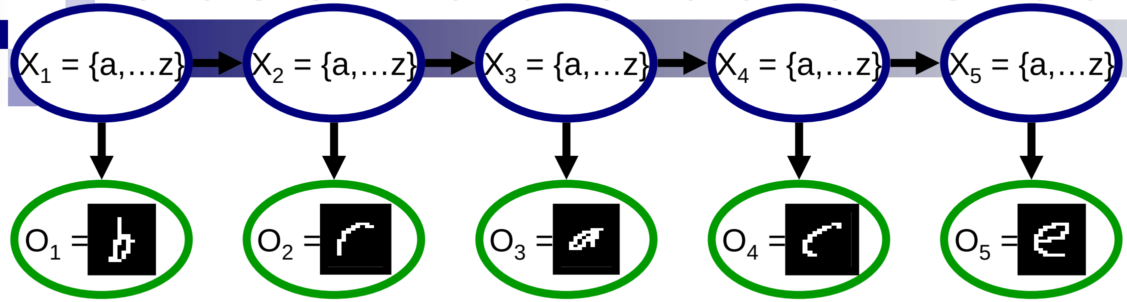
Learn 3 distributions:

$$P(X_1 {=} a) = \frac{\text{Count}(X_i = a)}{m}$$

$$P(O_i {=} \text{pixel17} = \text{on} \mid X_i {=} a) = \frac{\text{Count}(A = a,\ \text{Pixel17} = \text{on})}{\text{count}(A = a)}$$

$$P(X_i {=} a \mid X_{i-1} {=} b) = \frac{\text{Count}(X_{i-1} = b,\ X_i = a)}{\text{count}(X_{i-1} = b,\ X_i = ?)}$$

Learning HMMs from fully observable data is easy



Learn 3 distributions:

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

N = data set size

select training data where the letter was a

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

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

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

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

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

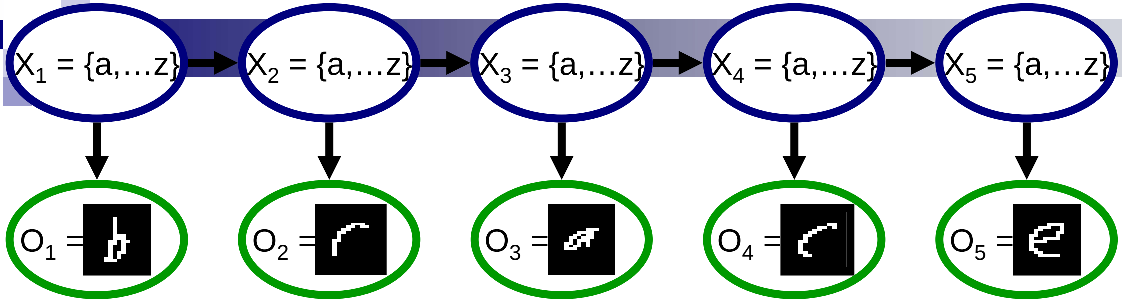
$$
\begin{aligned}
\ell(\theta : \mathcal{D}) &= \log P(\mathbf{o} \mid \theta) \\
&= \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o} \mid \theta)
\end{aligned}
$$

marginal likelihood

  - 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



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

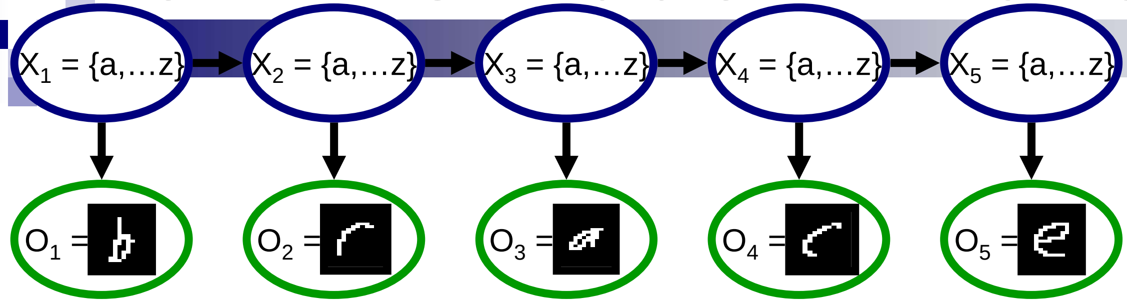
sum naively

k – letters

n – positions (word length)

k^n – terms

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



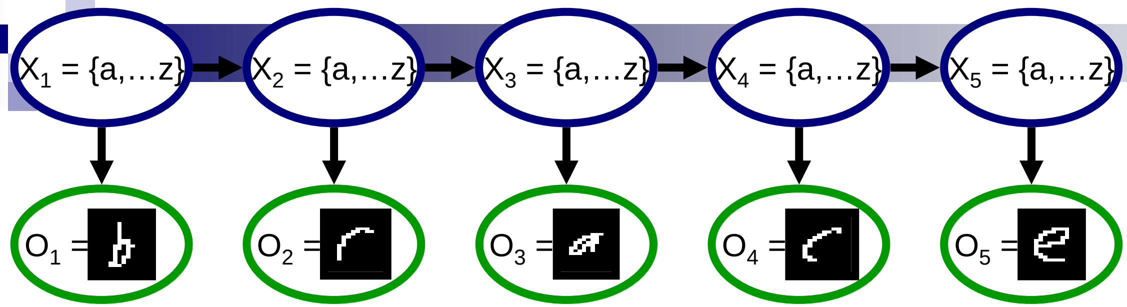
- Can compute efficiently with variable elimination:

$$\ell(\theta : \mathcal{D}) = \log P(\mathbf{o} \mid \theta)$$
$$= \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o} \mid \theta) = \log \sum_{x_1 \dots x_n} P(x_1) \cdot P(O_1 | x_1) \prod_{t=2}^{n} P(x_t | x_{t-1}) P(O_t | x_t)$$

$$= \log \sum_{x_1 \dots x_{n-1}} P(x_1) P(O_1 | x_1) \prod_{t=2}^{n-1} P(x_t | x_{t-1}) P(O_t | x_t) \underbrace{\sum_{x_n} P(x_n | x_{n-1}) P(O_n | x_n)}_{g(x_{n-1})}$$

eliminate x_{n-1}
x_{n-2}
$\vdots$

EM for HMMs when **X** is hidden



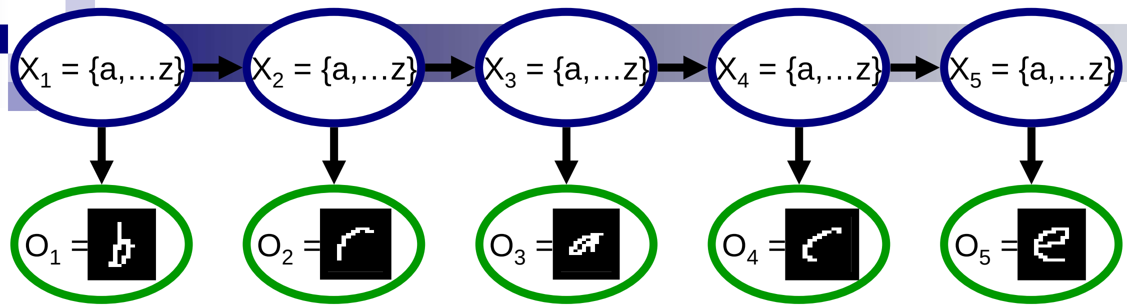
- E-step: Use inference (forwards-backwards algorithm)

$$p(X_3 = a \mid O = \boxed{b \mid r \mid a \mid c \mid k})$$

- M-step: Recompute parameters with weighted data

learn weighted data!! :)

E-step



- E-step computes probability of hidden vars **x** given **o**

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

iteration t+1

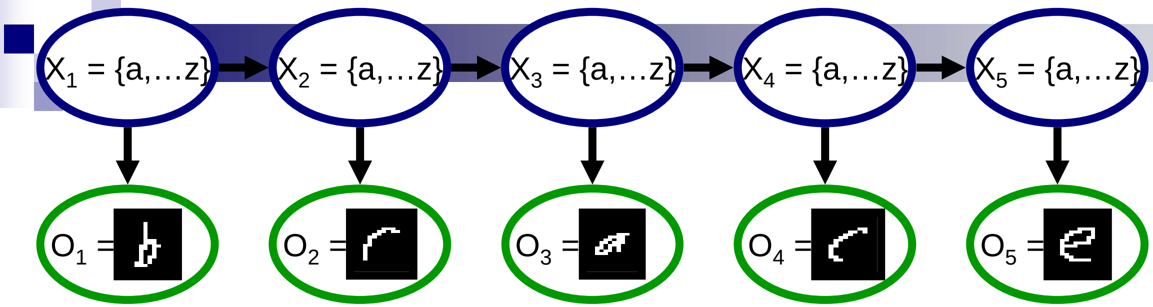
- Will correspond to inference
  - use forward-backward algorithm!

The M-step

Starting state: $P(X_1) \leftarrow \theta_{X_1=a}, \theta_{X_1=b} \ldots$

transition model: $P(X_i \mid X_{i-1}) \leftarrow \theta_{X_i=a \mid X_{i-1}=b}$

obs. model: $P(O_t \mid X_t) \leftarrow \theta_{O\,pixel7=on \mid X=a}$



- Maximization step:

likelihood completed data

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

weighted

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

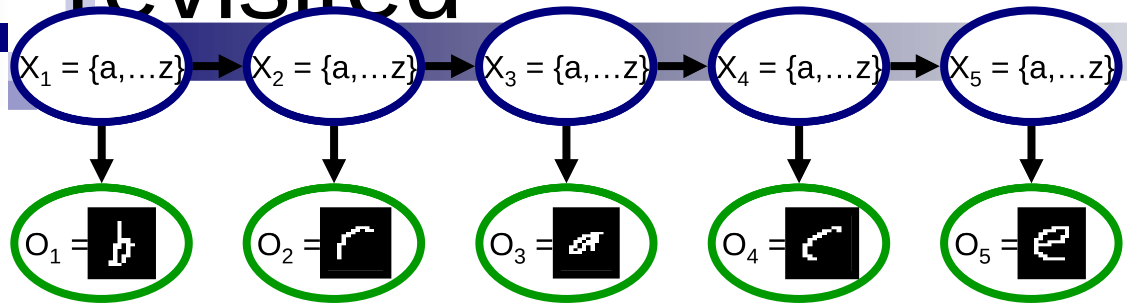
Decomposition of likelihood revisited

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

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

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

$\log a \cdot b = \log a + \log b$



- Likelihood optimization decomposes:

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

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

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

learn start state dist.

learn obs. model

learn transition model

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

Starting state probability $P(X_1)$

$P(a,b) = P(a) \cdot P(b|a)$

chain rule

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

$Q \in$ prob. dist. ✓

$Q(x_1 \ldots x_n | o) = Q(x_1|o) \cdot Q(x_2 \ldots x_n | x_1, o)$

$$\max_{\theta_{X_1}} \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}) \log P(x_1 \mid \theta_{X_1}) = \max_{\theta_{X_1}} \sum_{x_1, \ldots x_n} Q(x_1 \ldots x_n | o) \log P(x_1 | \theta_{x_1})$$

$$= \max_{\theta_{X_1}} \sum_{x_1 \ldots x_n} Q(x_1|o) \cdot Q(x_2 \ldots x_n | x_1, o) \log P(x_1 | \theta_{x_1}) =$$

$$= \max_{\theta_{X_1}} \sum_{x_1} Q(x_1|o) \log P(x_1|\theta_{X_1}) \cdot \underbrace{\sum_{x_2 \ldots x_n} Q(x_2 \ldots x_n | \overset{x_1,}{o})}_{1}$$

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

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

$P(X_1 = x_1 \mid \theta_{X_1}) = \theta_{X_1 = x_1}$

$$\frac{\partial}{\partial \theta_{X_1}} \sum_{j=1}^{m} \sum_{x_1} Q(x_1 \mid o^{(j)}) \log P(X_1 \mid \theta_{X_1}) = 0$$

$P(X_1 = t \mid \theta_{X_1}) = \theta_{X_1 = t}$

$P(X_1 = f \mid \theta_{X_1}) = 1 - \theta_{X_1 = t}$

$\frac{\partial}{\partial x} \log x = \frac{1}{x}$

$$= \sum_{j=1}^{m} \sum_{x_1} Q(x_1 \mid o^{(j)}) \frac{\partial}{\partial \theta_{x_1}} \log \theta_{x_1}$$

$$= \sum_{j=1}^{m} \left[Q(x_1 = t \mid o^{(j)}) \frac{1}{\theta_{x_1 = t}} + Q(x_1 = f \mid o^{(j)}) \frac{1}{1 - \theta_{x_1 = t}} \right] = 0$$

rearrange

$$\theta_{x_1 = t} = \frac{\sum_{j=1}^{m} Q(x_1 = t \mid o^{(j)})}{\sum_{j=1}^{m} \left[Q(x_1 = t \mid o^{(j)}) + Q(x_1 = f \mid o^{(j)}) \right]}$$

$m =$

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

$\log \prod = \sum \log$

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

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

$$= \max_{\theta_{X_t|X_{t-1}}} \sum_{j} \sum_{t=2}^{n} \sum_{x_1 \cdots x_n} Q(x_t, x_{t-1}|o) \cdot Q(x_1 \cdots x_{t-2}, x_{t+1} \cdots x_n | x_t, x_{t-1}, o) \log P(x_t|x_{t-1}, \theta)$$

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

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

Observation probability $P(O_t|X_t)$

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

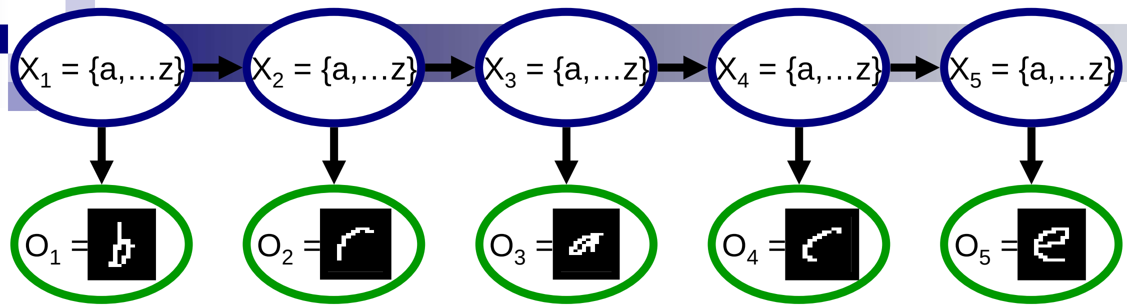
$$\max_{\theta_{O|X}} \sum_{\mathbf{x}} Q(\mathbf{x} \mid \mathbf{o}) \log \prod_{t=1}^{n} P(o_t \mid x_t, \theta_{O|X}) = \max_{\theta_{O|X}} \sum_{t=1}^{n} \sum_{x_t} Q(x_t \mid o) \log P(O_t \mid x_t, \theta_{O|X})$$

$$\delta(o_t^{(j)} = a) \leftarrow \begin{cases} 1; & \text{if } t\text{'th obs. of } j\text{'th training example} = a \\ 0; & \text{otherwise} \end{cases}$$

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

E-step revisited

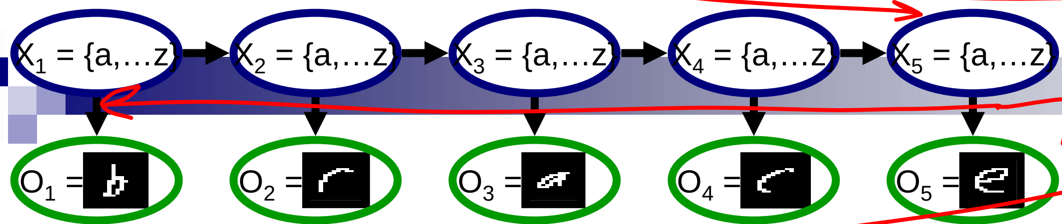
$$Q^{(t+1)}(\mathbf{x} \mid \mathbf{o}) \;=\; P(\mathbf{x} \mid \mathbf{o}, \theta^{(t)})$$



- E-step computes probability of hidden vars **x** given **o**
- Must compute: $P(X_t = a \mid O)$
  - $\boxed{Q(x_t=a|\mathbf{o})}$ – marginal probability of each position

    $P(X_{t+1} = a, X_t = b \mid O)$
  - $\boxed{Q(x_{t+1}=a,x_t=b|\mathbf{o})}$ – joint distribution between pairs of positions

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

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

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

$\alpha_n(X_n)$ 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}

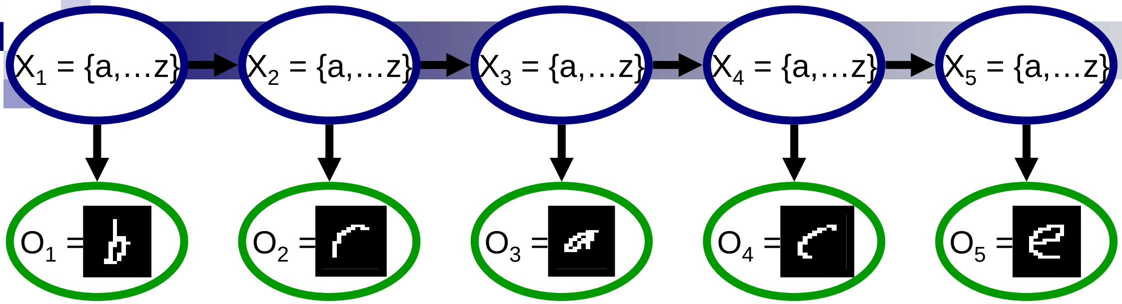
$\forall x'_i$

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

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

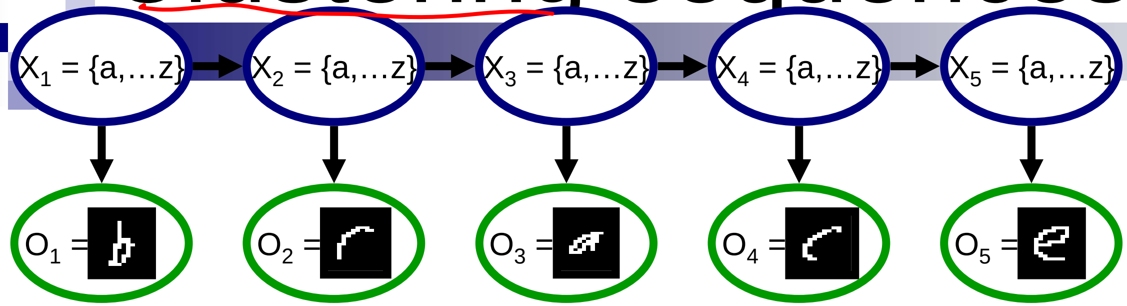
E-step revisited

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

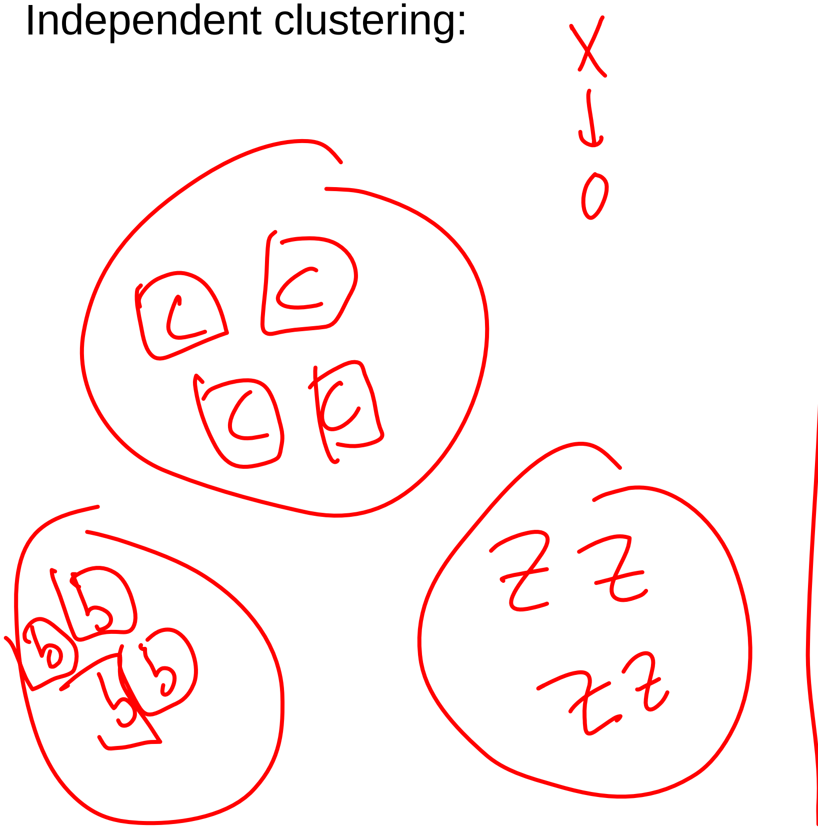


- E-step computes probability of hidden vars **x** given **o**
- Must compute:
  - $Q(x_t=a|\mathbf{o})$ – marginal probability of each position
    - Just forwards-backwards! $P(X_t = a \mid o_1 \ldots o_n)$
  - $Q(x_{t+1}=a,x_t=b|\mathbf{o})$ – joint distribution between pairs of positions $P(x_{t+1}=a, x_t=b \mid o_1 \ldots o_n)$
    - Homework! ☺

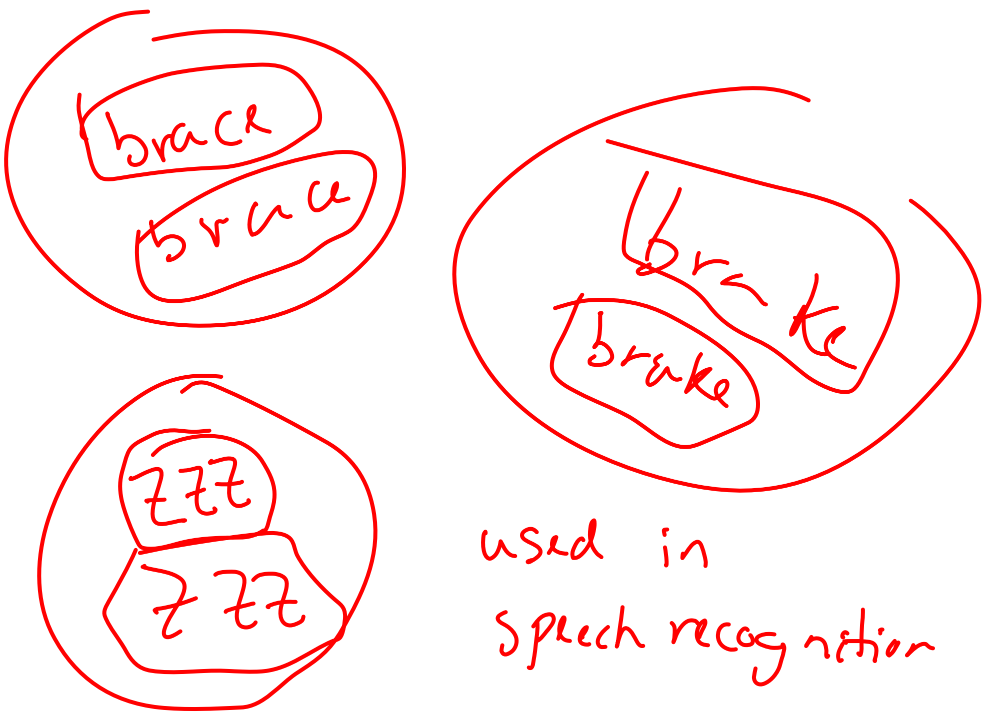
What can you do with EM for HMMs? 1 – Clustering sequences



Independent clustering:

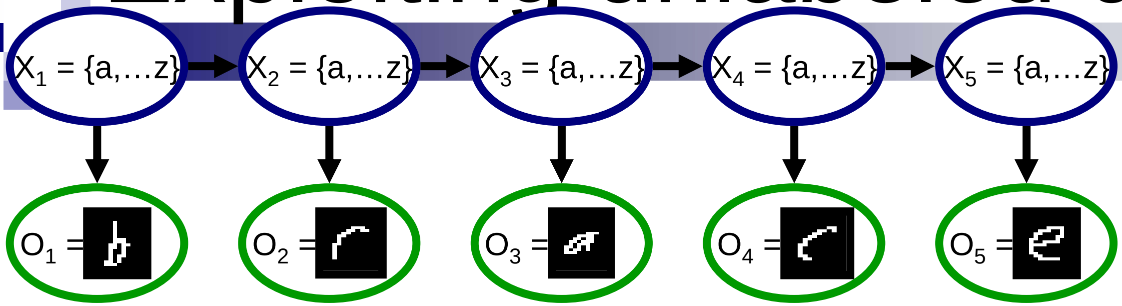


Sequence clustering:



used in speech recognition

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=b r a c e>
    - ⋮
  - Unlabeled data:
    - <X=?????,O=b r a c e>

Exploiting unlabeled data in clustering

- A few data points are labeled
  - <x,o> $p_1 = \langle x=1,\ o=\{0.5, 0.8\}\rangle$ (o_1, o_2)
- Most points are unlabeled
  - <?,o> $p_2\ \langle x=?,\ o=\{0.4, 0.9\}\rangle$ (o_1, o_2)
- In the E-step of EM:
  - If i'th point is unlabeled: $p(x=j|o_i)$ point generate by clusters
    - 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

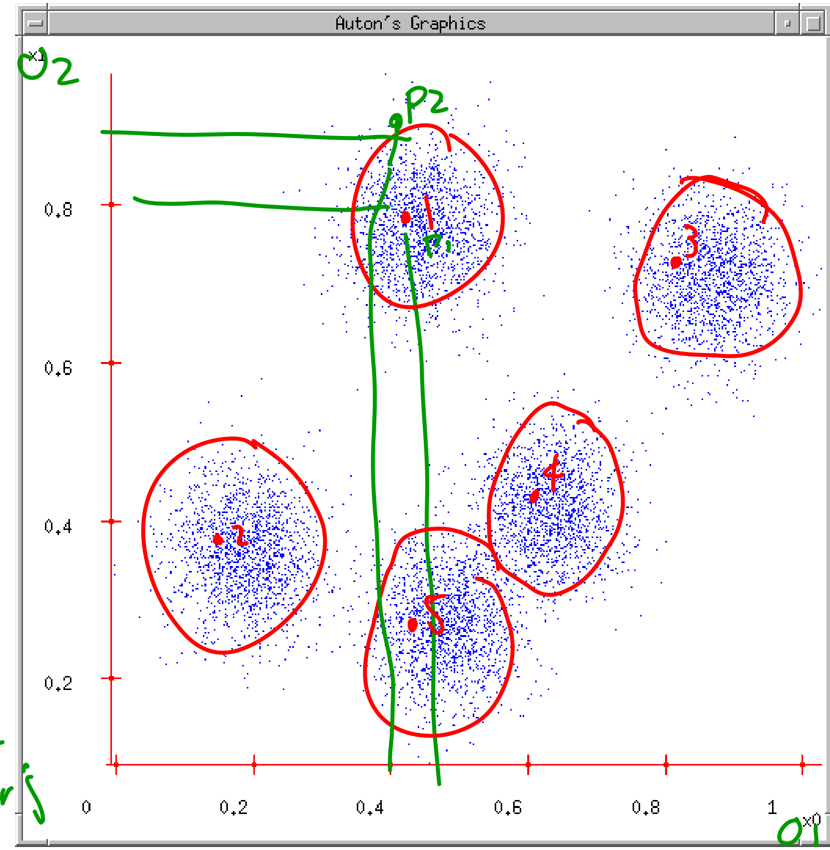


Table 3. Lists of the words most predictive of the course class in the WebKB data set, as they change over iterations of EM for a specific trial. By the second iteration of EM, many common course-related words appear. The symbol D indicates an arbitrary digit.

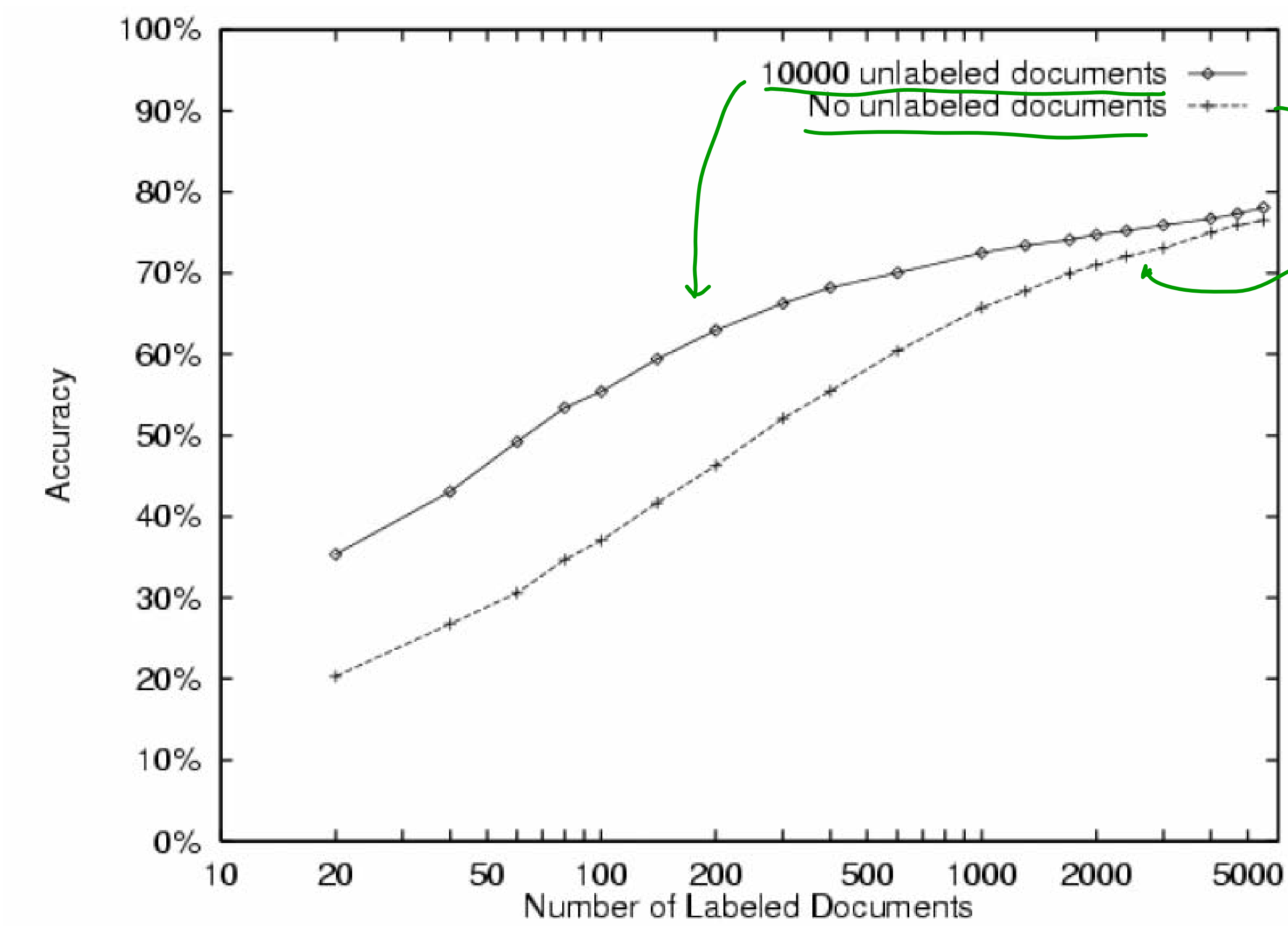
Iteration 0	Iteration 1	Iteration 2
intelligence	DD	D
DD	D	DD
artificial	lecture	lecture
understanding	cc	cc
DDw	D^*	$DD{:}DD$
dist	$DD{:}DD$	due
identical	handout	D^*
rus	due	homework
arrange	problem	assignment
games	set	handout
dartmouth	tay	set
natural	DDam	hw
cognitive	yurttas	exam
logic	homework	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

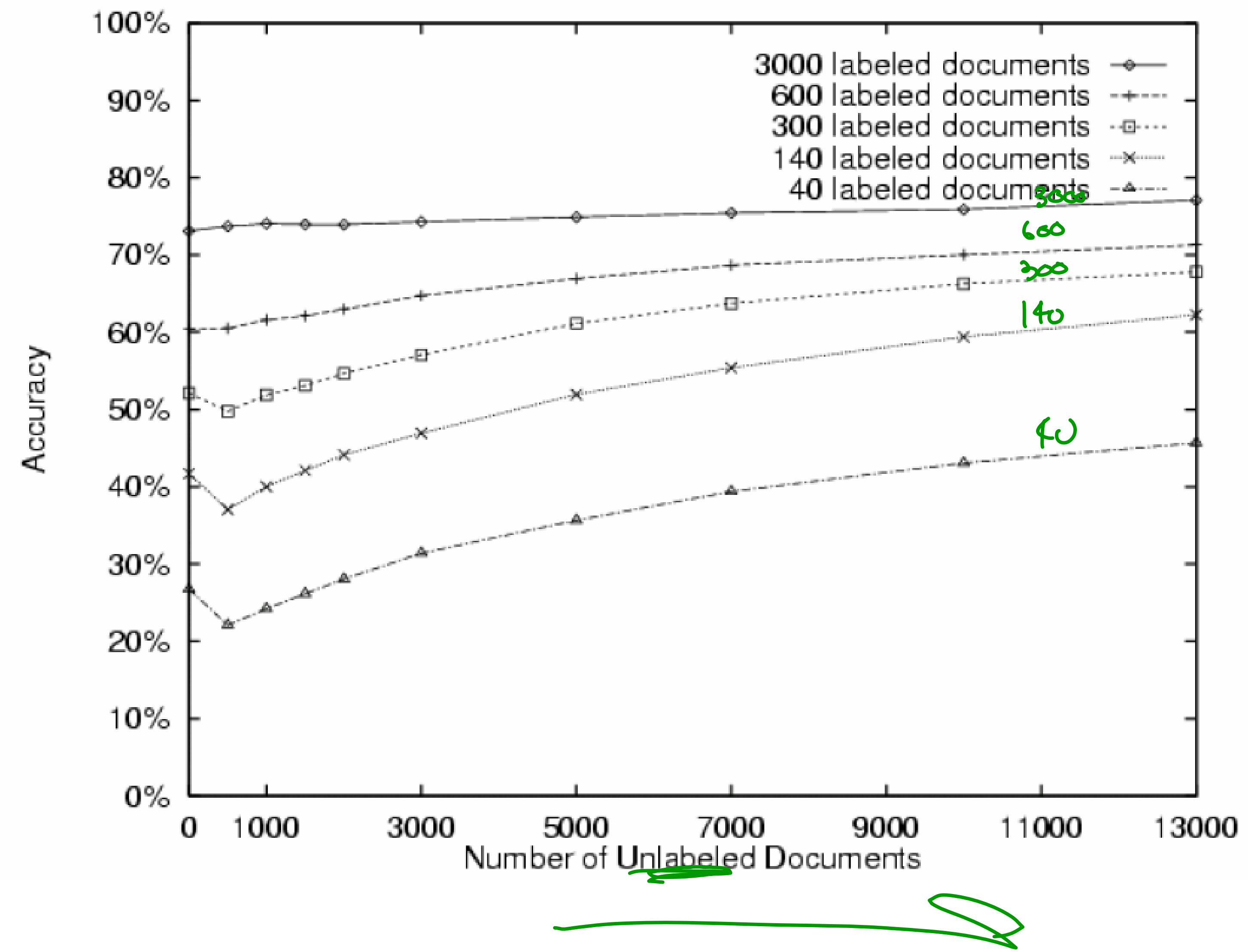
words likely associated with class = course

P(word | X=course) is high

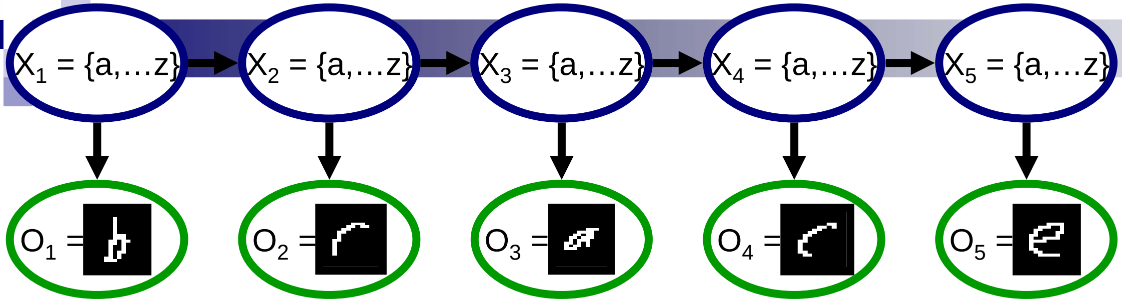
20 Newsgroups data – advantage of adding unlabeled data



20 Newsgroups data – Effect of additional unlabeled data



Exploiting unlabeled data in HMMs



- A few data points are labeled
  - <x,o> <x="brace", o= b r a c e>
- Most points are unlabeled
  - <?,o> <x="?????", o= b r a k e>
- 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 $Q(X\neq x|o_i)=0$

$Q(X_3=a \mid \text{b r a c e}) = 1$
$Q(X_3\neq a \mid \text{''}) = 0$

- M-step as usual
  - Speed up by remembering counts for labeled data (pseudo-counts)

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

EM for Bayes Nets

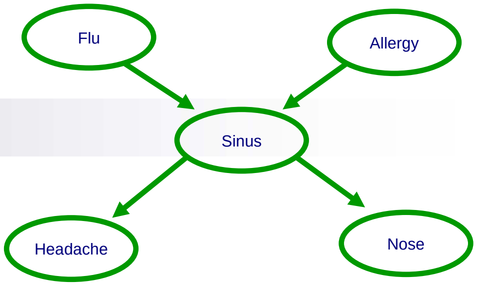
Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

April 12th, 2006

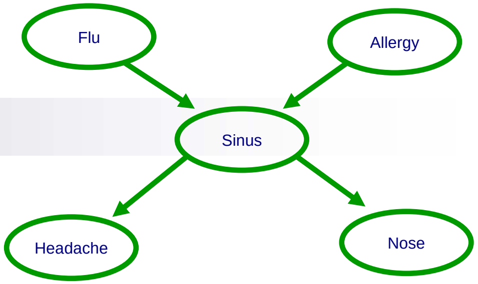
Data likelihood for BNs



- Given structure, log likelihood of fully observed data:

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G})$$

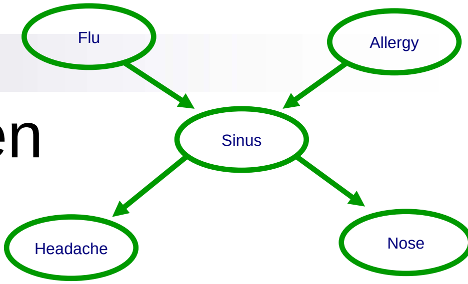
Marginal likelihood



- What if S is hidden?

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G})$$

Log likelihood for BNs with hidden data

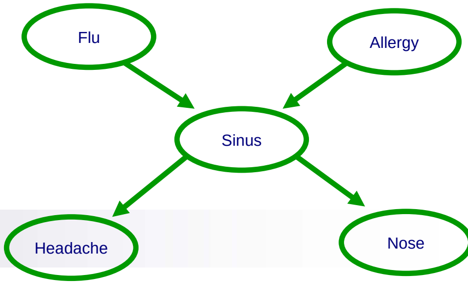


- Marginal likelihood – **O** is observed, **H** is hidden

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

$$= \sum_{j=1}^{m} \log \sum_{\mathbf{h}} P(\mathbf{h}, \mathbf{o}^{(j)} \mid \theta)$$

E-step for BNs

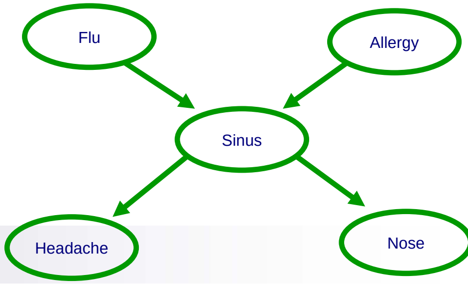


- E-step computes probability of hidden vars **h** given **o**

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

- Corresponds to inference in BN

The M-step for BNs

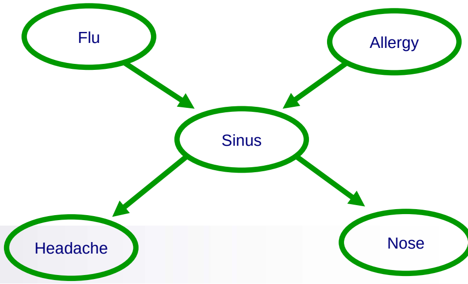


- Maximization step:

$$\theta^{(t+1)} \leftarrow \arg\max_{\theta} \sum_{\mathbf{x}} Q^{(t+1)}(\mathbf{h} \mid \mathbf{o}) \log P(\mathbf{h}, \mathbf{o} \mid \theta)$$

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

M-step for each CPT



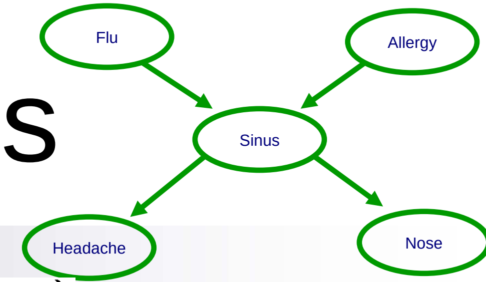
- M-step decomposes per CPT
  - Standard MLE:

$$P(X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{z}) = \frac{\text{Count}(X_i = x_i, \mathbf{Pa}_{X_i} = \mathbf{z})}{\text{Count}(\mathbf{Pa}_{X_i} = \mathbf{z})}$$

  - M-step uses expected counts:

$$P(X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{z}) = \frac{\text{ExCount}(X_i = x_i, \mathbf{Pa}_{X_i} = \mathbf{z})}{\text{ExCount}(\mathbf{Pa}_{X_i} = \mathbf{z})}$$

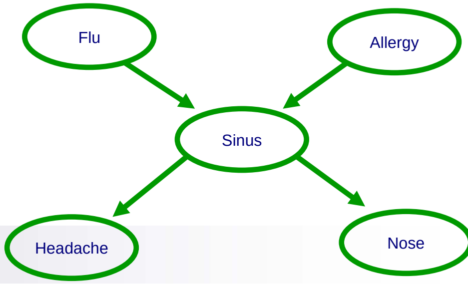
Computing expected counts



$$P(X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{z}) = \frac{\text{ExCount}(X_i = x_i, \mathbf{Pa}_{X_i} = \mathbf{z})}{\text{ExCount}(\mathbf{Pa}_{X_i} = \mathbf{z})}$$

- M-step requires expected counts:
  - For a set of vars **A**, must compute ExCount(**A**=**a**)
  - Some of **A** in example *j* will be observed
    - denote by $\mathbf{A}_O = \mathbf{a}_O^{(j)}$
  - Some of **A** will be hidden
    - denote by $\mathbf{A}_H$
- Use inference (E-step computes expected counts):
  - $\text{ExCount}^{(t+1)}(\mathbf{A}_O = \mathbf{a}_O^{(j)}, \mathbf{A}_H = \mathbf{a}_H) \leftarrow P(\mathbf{A}_H = \mathbf{a}_H \mid \mathbf{A}_O = \mathbf{a}_O^{(j)}, \theta^{(t)})$

Data need not be hidden in the same way



- When data is fully observed
  - A data point is
- When data is partially observed
  - A data point is
- But unobserved variables can be different for different data points
  - e.g.,
- Same framework, just change definition of expected counts
  - $\text{ExCount}^{(t+1)}(\mathbf{A_O = a_O}^{(j)}\mathbf{, A_H = a_H}) \leftarrow P(\mathbf{A_H = a_H} \mid \mathbf{A_O = a_O}^{(j)}, \theta^{(\mathbf{t})})$

What you need to know

- EM for Bayes Nets
- E-step: inference computes expected counts
  - Only need expected counts over X_i and $\mathbf{Pa}_{x_i}$
- M-step: expected counts used to estimate parameters
- Hidden variables can change per datapoint
- Use labeled and unlabeled data $\rightarrow$ some data points are complete, some include hidden variables