Expectation Maximization, and Learning from Partly Unobserved Data

Recommended reading:

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

Machine Learning 10-701
March 2005

Tom M. Mitchell
Carnegie Mellon University

Outline

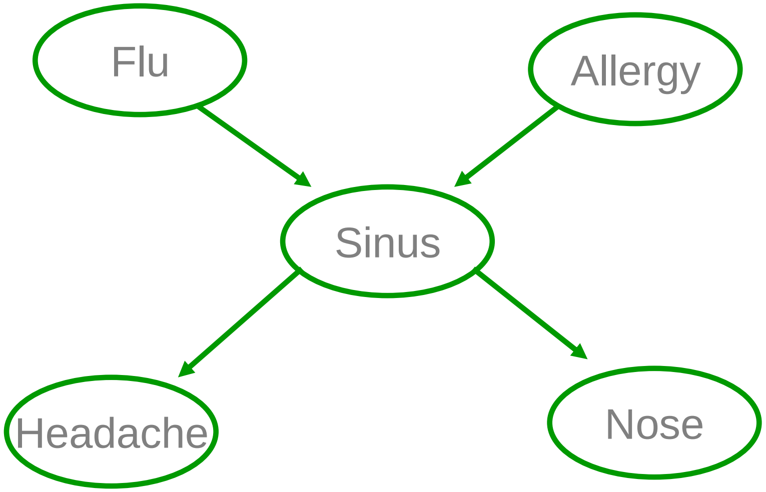
- EM_1: Learning Bayes network CPT's from partly unobserved data
- EM_2: Learning HMM's with unobserved hidden states
- EM_3: Mixture of Gaussians – clustering
- EM: the general story

1. Learning Bayes net parameters from partly unobserved data

Learning CPTs from Fully Observed Data

- Example: Consider learning the parameter

$$\theta_{s|ij} \equiv P(S=1|F=i, A=j)$$



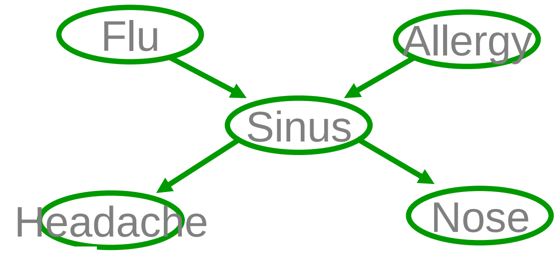
- MLE (Max Likelihood Estimate) is

$$\theta_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$

k^{th} training example

- Remember why?

MLE estimate of from fully observed data



- Maximum likelihood estimate

$$\theta \leftarrow \arg\max_{\theta} \log P(data|\theta)$$

- Our case:

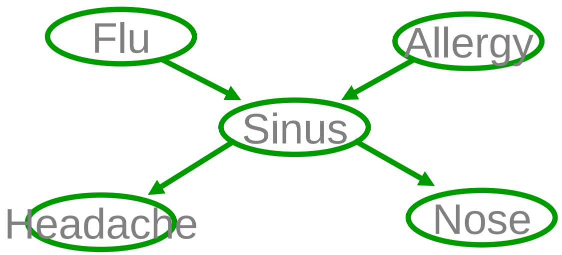
$$P(data|\theta) = \prod_{k=1}^{K} P(f_k)P(a_k)P(s_k|f_k a_k)P(h_k|s_k)P(n_k|s_k)$$

$$\log P(data|\theta) = \sum_{k=1}^{K} \log P(f_k) + \log P(a_k) + \log P(s_k|f_k a_k) + \log P(h_k|s_k) + \log P(n_k|s_k)$$

$$\frac{\partial \log P(data|\theta)}{\partial \theta_{s|ij}} = \sum_{k=1}^{K} \frac{\partial \log P(s_k|f_k a_k)}{\partial \theta_{s|ij}}$$

$$\theta_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$

Estimate θ when S unobservable, FAHN observed



- Can't calculate maximum likelihood estimate $\theta \leftarrow \arg\max_\theta \log P(F, A, S, H, N|\theta)$

- Chicken and egg problem

- What do we want to maximize in order to choose $\theta_{s|ij}$??

$$\arg\max_\theta \sum_i P(S = i|F, A, H, N, \theta) \log P(F, A, S = i, H, N|\theta)]$$

$$= \arg\max_\theta E_{S|F,A,H,N,\theta}[\log P(F, A, S, H, N|\theta)]$$

EM

$$\theta \leftarrow \arg\max_{\theta} \sum_{i} P(S = i | F, A, H, N, \theta) \log P(F, A, S = i, H, N | \theta)$$

$$= \arg\max_{\theta} E_{S|F,A,H,N,\theta}[\log P(F, A, S, H, N | \theta)]$$

EM is a general procedure for solving such problems

Given observed variables X, unobserved Z (X={F,A,H,N}, Z={S})

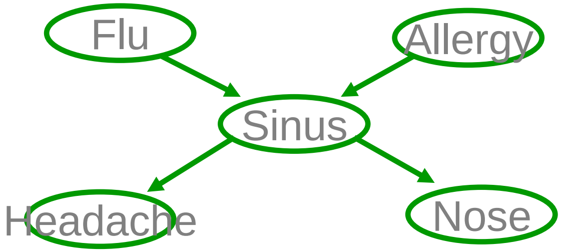
Define $Q(\theta' | \theta) = E_{Z|X,\theta}[\log P(X, Z | \theta')]$

Iterate until convergence:

- E Step: Calculate $Q(\theta' | \theta)$ by using X and current θ to estimate $P(Z|X,\theta)$
- M Step: Replace current θ by

$$\theta \leftarrow \arg\max_{\theta'} Q(\theta' | \theta)$$

EM and estimating $\theta_{s|ij}$



observed X = {F,A,H,N}, unobserved Z={S})

E step: Calculate for each training example, k

$$P(S_k=1|f_k a_k h_k n_k,\theta) = E[s_k] = \frac{P(S_k=1, f_k a_k h_k n_k|\theta)}{P(S_k=1, f_k a_k h_k n_k|\theta) + P(S_k=0, f_k a_k h_k n_k|\theta)}$$

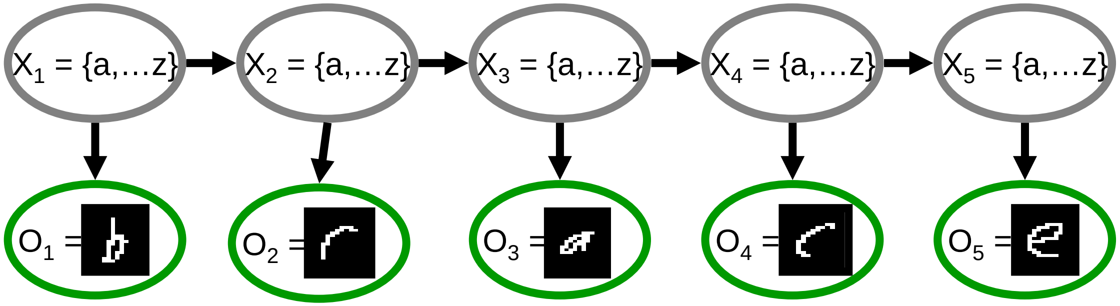
M step:

$$\theta_{s|ij} \leftarrow \frac{\sum_{k=1}^{K} \delta(f_k=i, a_k=j) P(s_k=1|f_k a_k h_k n_k,\theta)}{\sum_{k=1}^{K} \delta(f_k=i, a_k=j)}$$

Recall MLE was: $$\theta_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k=i, a_k=j, s_k=1)}{\sum_{k=1}^{K} \delta(f_k=i, a_k=j)}$$

2. Learning HMM’s with EM

Learning HMMs from fully observed data is easy



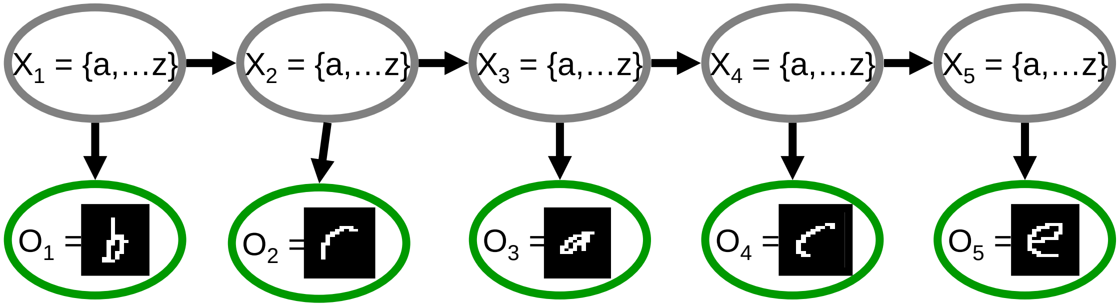
Learn 3 distributions:

$$P(X_1)$$

$$P(O_i \mid X_i)$$

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

Learning HMMs from fully observed data is easy



Data

$\langle x_1^{(j)}, o_1^{(j)}, x_2^{(j)}, o_2^{(j)}, \ldots, x_n^{(j)}, o_n^{(j)} \rangle$

Learn 3 distributions:

$$P(X_1 = x_1) = \frac{\text{Count}(X_1 = x_1)}{m}$$

use all i's in counts

each data "point", contributes ... count

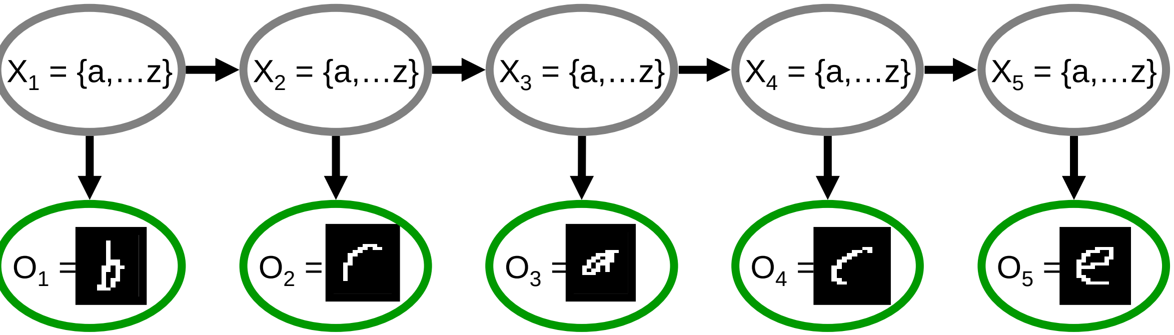
$$P(O_i = o_i \mid X_i = x_i) \propto \frac{\text{Count}(O_i = \ldots}{\text{Count}(X \ldots}$$

$$P(X_i = x_i \mid X_{i-1} = x_{i-1}) \propto \frac{\text{Count}(\ldots}{\text{Count}(X_{i-1} = x_{i-1})}$$

What if we need to learn from data with observed O's, unobserved X's ?

Parameter sharing / tying

Learning HMMs with EM



Just 3 distributions:

$$\left.\begin{array}{l} P(X_1) \\ P(X_i \mid X_{i-1}) \\ P(O_i \mid X_i) \end{array}\right\} \theta$$

Observed data: $O \equiv O_1, \ldots O_n$

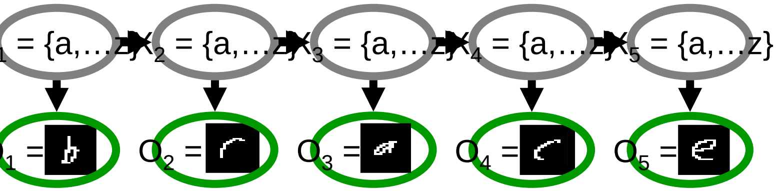
Unobserved data: $X \equiv X_1, \ldots X_n$

EM: $Q(\theta'|\theta) = E_{X|O,\theta}[\log P(X,O|\theta')]$

E step: compute $P(X|O,\theta)$

M step: $\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$

Learning HMMs: E step



Observed data: $O \equiv O_1, \ldots O_n$

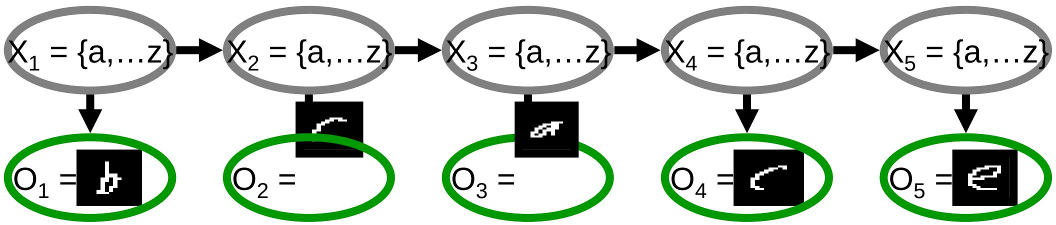
Unobserved data: $X \equiv X_1, \ldots X_n$

$$\left.\begin{array}{l} P(X_1) \\ P(X_i \mid X_{i-1}) \\ P(O_i \mid X_i) \end{array}\right\} \theta$$

E step: compute $P(X|O,\theta)$

use the Forward-Backward algorithm!

The forward-backward algorithm



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

Complexity O(n)

- 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}

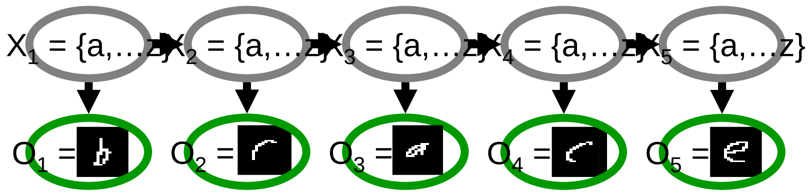
$$\alpha_i(X_i) = \sum_{x_{i-1}} P(o_i \mid X_i)P(X_i \mid X_{i-1} = x_{i-1})\alpha_{i-1}(x_{i-1})$$

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

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

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

Learning HMMs



Observed data: $O \equiv O_1, \ldots O_n$

Unobserved data: $X \equiv X_1, \ldots X_n$

$$\left. \begin{array}{l} P(X_1) \\ P(X_i \mid X_{i-1}) \\ P(O_i \mid X_i) \end{array} \right\} \theta$$

E step: compute $P(X|O,\theta)$

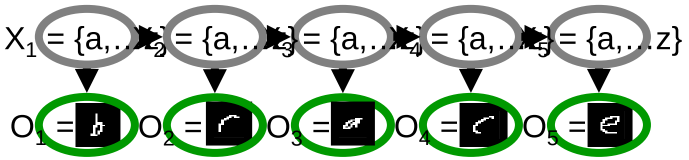
- using forward/backward algorithm

M step: $\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$

$$Q(\theta'|\theta) = E_{X|O,\theta}[\log P(X,O|\theta')]$$

$$\log P(X,O|\theta') = \log[P(X_1|\theta')P(O_1|X_1,\theta') \prod_{i=2}^{n} P(X_i|X_{i-1},\theta')P(O_i|X_i,\theta')]$$

HMMs: M step: $\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$



$$Q(\theta'|\theta) = E_{X|O,\theta}[\log P(X, O|\theta')]$$

$$\log P(X, O|\theta') = \log[P(X_1|\theta')P(O_1|X_1, \theta') \prod_{i=2}^{n} P(X_i|X_{i-1}, \theta')P(O_i|X_i, \theta')]$$

$$= \log P(X_1|\theta') + \sum_{i=2}^{n} \log P(X_i|X_{i-1}, \theta') + \sum_{i=1}^{n} \log P(O_i|X_i, \theta')$$

$$E_{X|O,\theta}[\log P(X, O|\theta')] = E_{X_1|O,\theta}[\log P(X_1|\theta')] + \sum_{i=2}^{n} E_{X_i,X_{i-1}|O,\theta}[\log P(X_i|X_{i-1}, \theta')]$$

$$+ \sum_{i=1}^{n} E_{X_i|O,\theta}[\log P(O_i|X_i, \theta')]]$$

HMM's: M Step: $\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$

$$Q(\theta'|\theta) \equiv E_{X|O,\theta}[\log P(X,O|\theta')]$$

$$E_{X|O,\theta}[\log P(X,O|\theta')] = E_{X_1|O,\theta}[\log P(X_1|\theta')] + \sum_{i=2}^{n} E_{X_i,X_{i-1}|O,\theta}[\log P(X_i|X_{i-1},\theta')]$$
$$+ \sum_{i=1}^{n} E_{X_i|O,\theta}[\log P(O_i|X_i,\theta')]]$$

$$\frac{\partial Q(\theta'|\theta)}{\partial \theta'} = 0 \quad \rightarrow$$

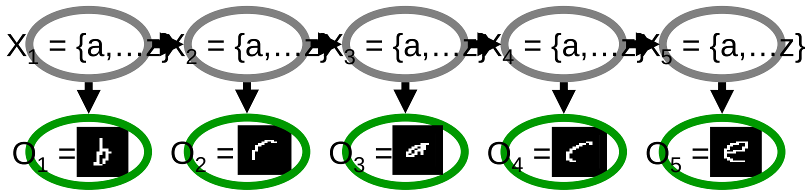
$$\pi'_i \equiv \widehat{P}(X_1 = i) \leftarrow P(X_1 = i|O,\theta)$$

$$\phi'_{ij} \equiv \widehat{P}(X_t = i|X_{t-1} = j) \leftarrow \frac{\sum_{t=2}^{T} P(X_t = i, X_{t-1} = j|O,\theta)}{\sum_{t=2}^{T} P(X_{t-1} = j|O,\theta)}$$

$$\lambda'_{ij} \equiv \widehat{P}(O_t = i|X_t = j) \leftarrow \frac{\sum_{t=1}^{T} \delta(O_t = i) P(X_t = j|O,\theta)}{\sum_{t=1}^{T} P(X_t = j|O,\theta)}$$

$$\theta = \langle \pi, \phi, \lambda \rangle$$

Learning HMMs



Observed data: $O \equiv O_1, \ldots O_n$

Unobserved data: $X \equiv X_1, \ldots X_n$

$$\left.\begin{array}{l} P(X_1) \\ P(X_i \mid X_{i-1}) \\ P(O_i \mid X_i) \end{array}\right\} \theta$$

E step: compute $P(X|O,\theta)$

• using forward/backward algorithm

M step: $\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$ $\quad \theta = \langle \pi, \phi, \lambda \rangle$

$$\pi'_i \equiv \hat{P}(X_1 = i) \leftarrow P(X_1 = i|O,\theta)$$

$$\phi'_{ij} \equiv \hat{P}(X_t = i|X_{t-1} = j) \leftarrow \frac{\sum_{t=2}^{T} P(X_t = i, X_{t-1} = j|O,\theta)}{\sum_{t=2}^{T} P(X_{t-1} = j|O,\theta)}$$

$$\lambda'_{ij} \equiv \hat{P}(O_t = i|X_t = j) \leftarrow \frac{\sum_{t=1}^{T} \delta(O_t = i) P(X_t = j|O,\theta)}{\sum_{t=1}^{T} P(X_t = j|O,\theta)}$$

Repeat until converged

What you should know about EM

- For learning from partly unobserved data
- MLEst of $\theta = \arg\max_{\theta} \log P(data|\theta)$
- EM estimate: $\theta = \arg\max_{\theta} E_{Z|X,\theta}[\log P(X,Z|\theta)]$

  Where X is observed part of data, Z is unobserved

  $Q(\theta'|\theta) = E_{Z|X,\theta}[\log P(X,Z|\theta')]$
- EM for training Bayes networks
- EM for training HMMs
- Be able to derive your own EM algorithm for your own problem