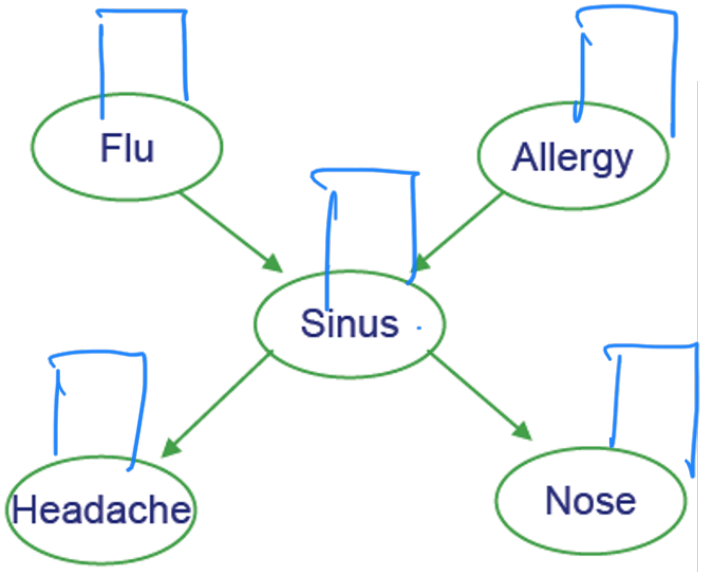
CS 4824/ECE 4424: Expectation Maximization

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.

Estimate θ from partly observed data — recap



- What if FAHN observed, but not S?
- Can't calculate MLE

$$\theta \leftarrow \arg\max_\theta \log \prod_k P(f_k, a_k, s_k, h_k, n_k|\theta)$$

- Let X be all *observed* variable values (over all examples)

X = { F, A, H, N }

- Let Z be all unobserved variable values

Z = { S }

- Can't calculate MLE

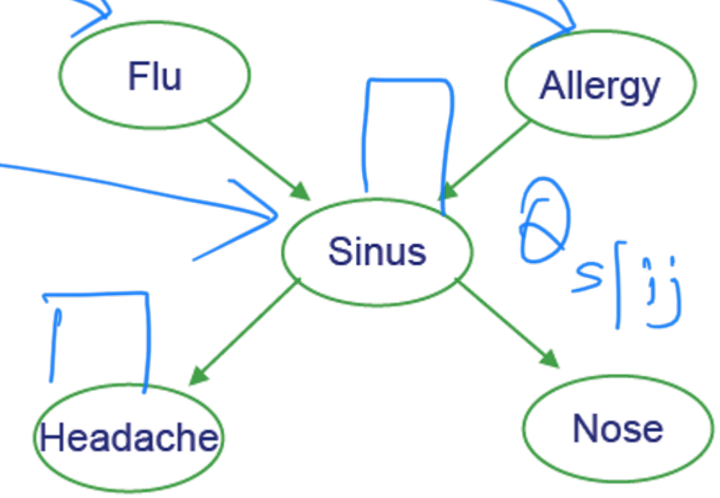
$$\boxed{\theta \leftarrow \arg\max_\theta \log P(X, Z|\theta)}$$

$$\theta \leftarrow \arg\max \mathbb{E}_{Z|X,\theta} \log P(X,Z|\theta)$$

- Estimate:

$$\boxed{\theta \leftarrow \arg\max_\theta E_{Z|X,\theta}[\log P(X, Z|\theta)]}$$

MLE estimate of $\theta_{s|ij}$ from fully observed data - recap



- Maximum likelihood estimate

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

- Our case

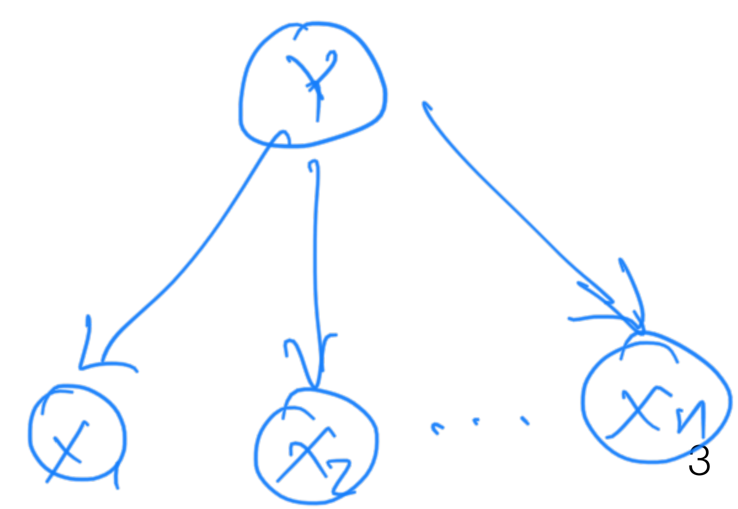
$$P(data|\theta) = \prod_{k=1}^{K} P(f_k, a_k, s_k, h_k, n_k)$$

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

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

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

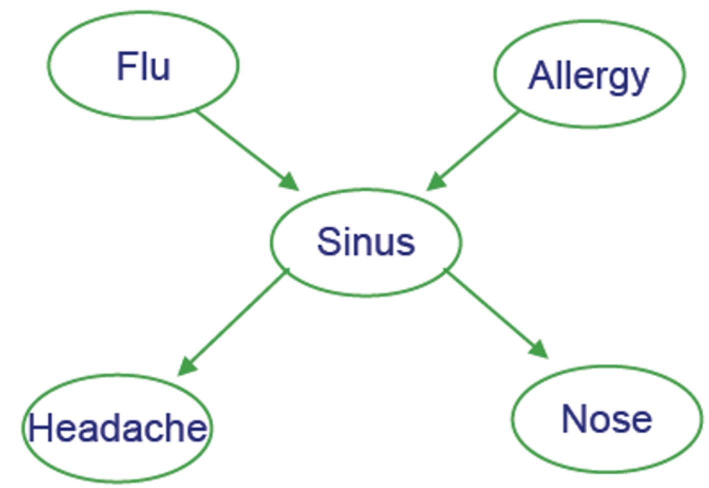
$$\theta_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$



Estimate θ from partly observed data

- What if FAHN observed, but not S?
- Can't calculate MLE

$$\theta \leftarrow \arg\max_{\theta} \log \prod_{k} P(f_k, a_k, s_k, h_k, n_k|\theta)$$

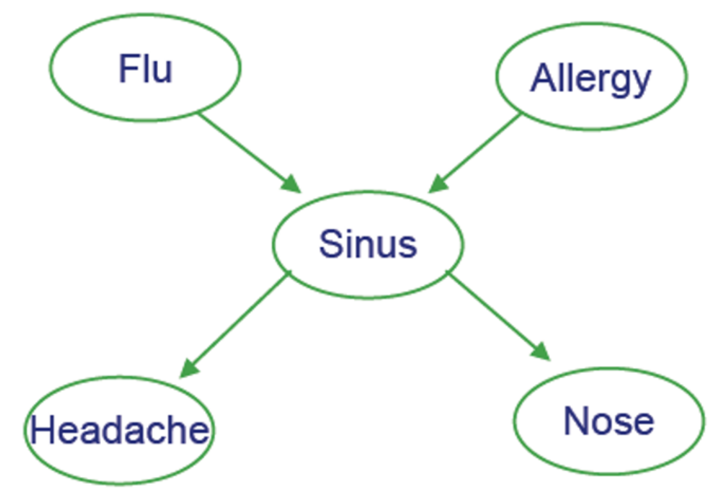


- Let X be all *observed* variable values (over all examples)
- Let Z be all unobserved variable values
- Can't calculate MLE

$$\theta \leftarrow \arg\max_{\theta} \log P(X, Z|\theta)$$

- What to do?

Estimate θ from partly observed data



- What if FAHN observed, but not S?
- Can't calculate MLE

$$\theta \leftarrow \arg\max_{\theta} \log \prod_k P(f_k, a_k, s_k, h_k, n_k | \theta)$$

- Let X be all *observed* variable values (over all examples)
- Let Z be all unobserved variable values
- Can't calculate MLE

$$\theta \leftarrow \arg\max_{\theta} \log P(X, Z | \theta)$$

- EM seeks* to estimate:

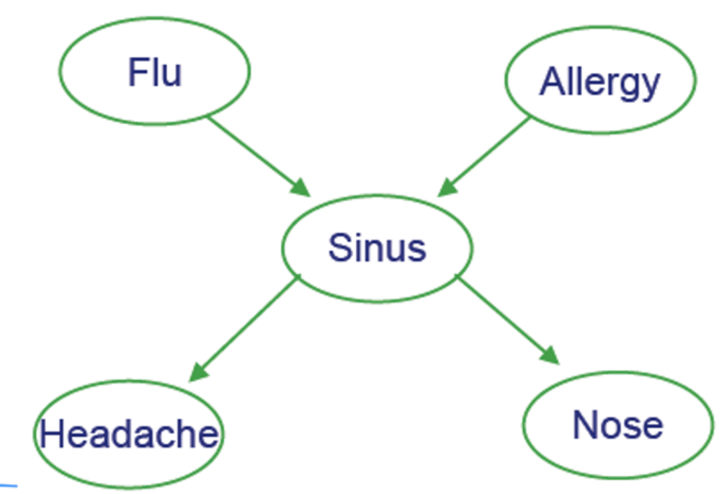
$$\boxed{\theta} \leftarrow \arg\max_{\theta} E_{Z|X,\theta}[\log P(X, Z|\theta)]$$

* EM guaranteed to find local optima

Estimate θ from partly observed data

- EM seeks to estimate:

$$\boxed{\theta \leftarrow \arg\max_{\theta} E_{Z|X,\theta}[\log P(X, Z|\theta)]}$$



- here, observed $\underbrace{X=\{F,A,H,N\}}$, unobserved $Z=\{S\}$

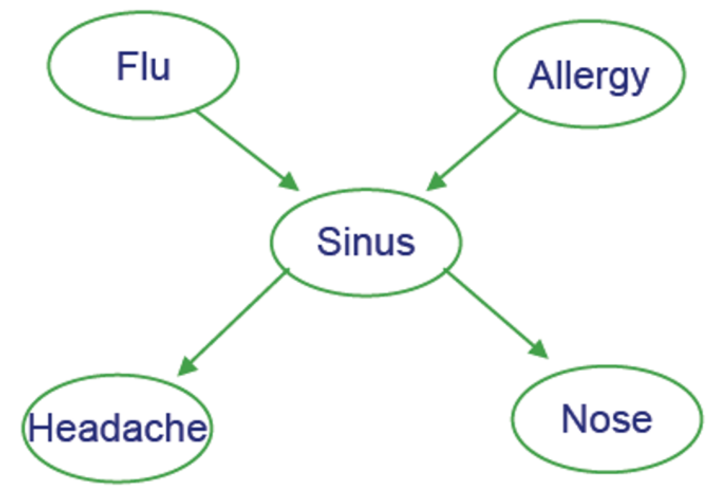
$$\log P(X, Z|\theta) = \sum_{k=1}^{K} \log P(f_k) + \log P(a_k) + \log P(s_k|f_k a_k) + \log P(h_k|s_k) + \log P(n_k|s_k)$$

$$E_{P(Z|X,\theta)} \log P(X, Z|\theta) = \sum_{k=1}^{K} \sum_{i=0}^{1} P(s_k = i \,|\, f_k, a_k, h_k, n_k) \Big[\log P(f_k) + \log P(a_k) + \log P(s_k|f_k, a_k) + \log (h_k|s_k) + \log (n_k|s_k)\Big]$$

Estimate θ from partly observed data

- EM seeks to estimate:

$$\theta \leftarrow \arg\max_{\theta} E_{Z|X,\theta}[\log P(X,Z|\theta)]$$



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

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

$$E_{P(Z|X,\theta)} \log P(X,Z|\theta) = \sum_{k=1}^{K} \sum_{i=0}^{1} P(s_k = i | f_k, a_k, h_k, n_k)$$

$$[log P(f_k) + \log P(a_k) + \log P(s_k|f_k a_k) + \log P(h_k|s_k) + \log P(n_k|s_k)]$$

EM algorithm — informally

- EM is a general procedure for learning from partly observed data
- Given observed variables X, unobserved Z (X={F,A,H,N}, Z={S})

Begin with arbitrary choice for parameters θ

Iterate until convergence:

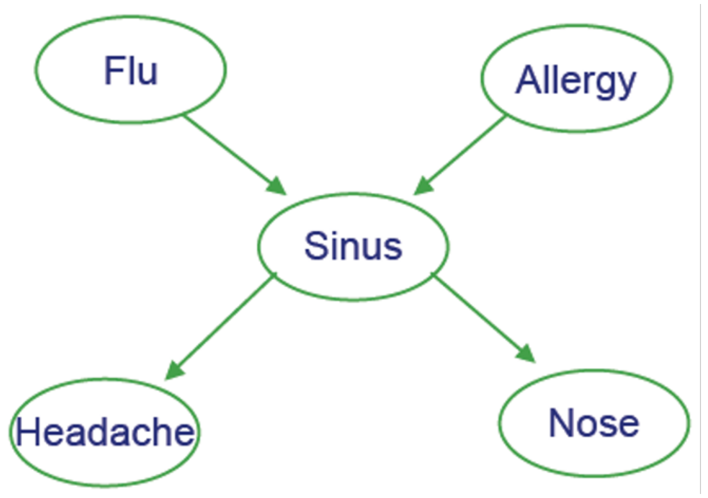
- E Step: estimate the values of unobserved Z conditioned on X using θ
- M Step: use observed values plus E-step estimates to derive a better θ

- Guaranteed to find local maximum. Each iteration increases

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

E Step: Use X, θ to Calculate $\boxed{P(Z|X,\theta)}$

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



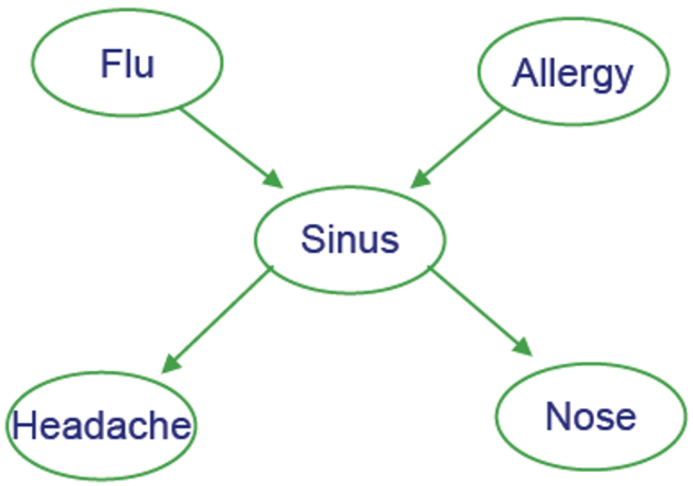
- How? Bayes net inference problem

$$\boxed{P(S_k = 1|f_k a_k h_k n_k, \theta)} = \frac{P(S_k=1, f_k, a_k, h_k, n_k | \theta)}{P(f_k, a_k, h_k, n_k | \theta)}$$

$$P(S=1, f_k, a_k, h_k, n_k | \theta) + P(S=0, f_k, a_k, h_k, n_k | \theta)$$

E Step: Use X, θ to Calculate P(Z|X,θ)

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

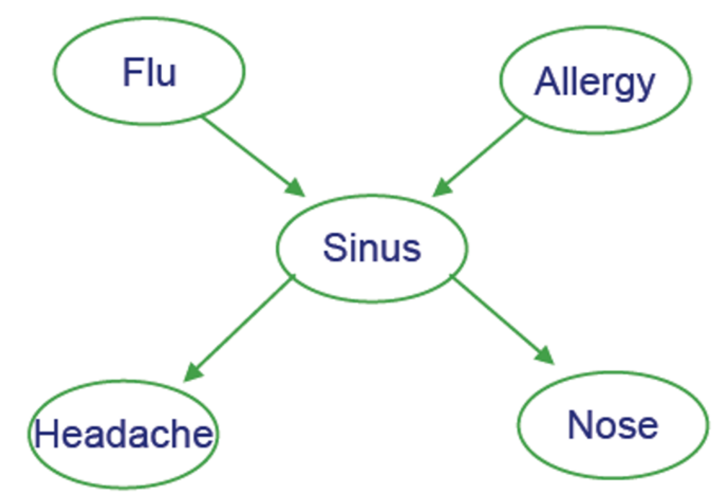


- How? Bayes net inference problem

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

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

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



- E Step: Calculate $P(Z_k | X_k; \theta)$ for each training example, k

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

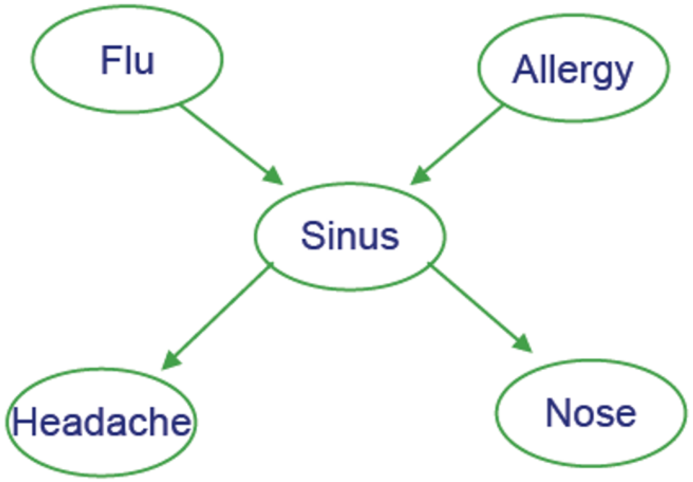
- M Step: update all relevant parameters.

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

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

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

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



- E Step: Calculate $P(Z_k | X_k; \theta)$ 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: update all relevant parameters. For example:

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

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

Generalizing: EM and estimating θ

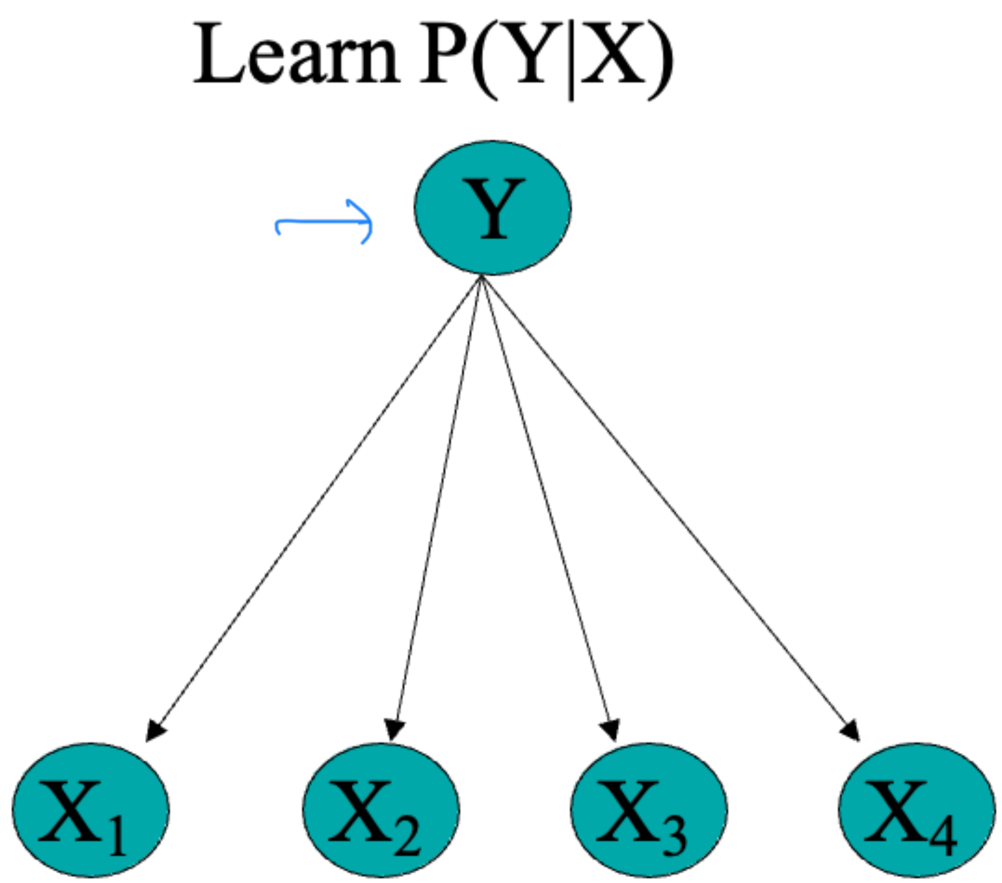
- More generally, given observed set X, unobserved set Z of boolean values

- E Step: Calculate for each training example, k the expected value of each unobserved variable

- M Step: Calculate estimates similar to MLE, but replacing each count by its <u>expected count</u>

$$\boxed{\delta(Y=1)} \rightarrow E_{Z|X,\theta}[Y] \qquad \delta(Y=0) \rightarrow (1 - E_{Z|X,\theta}[Y])$$

Using (partially) unlabeled data to help train naïve Bayes classifier

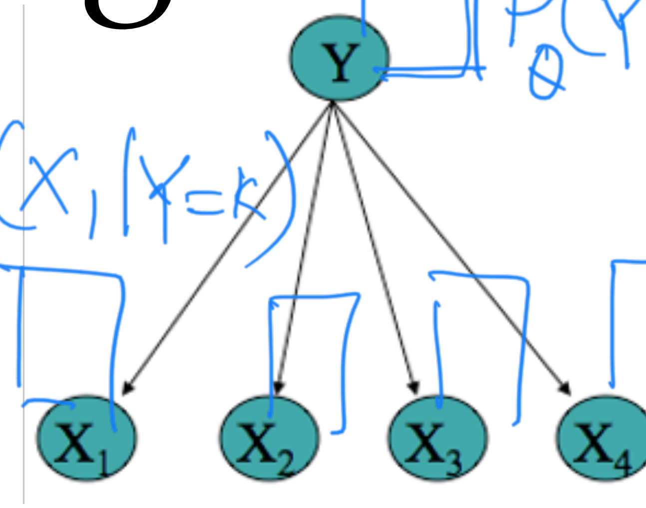


Y	X1	X2	X3	X4
1	0	0	1	1
0	0	1	0	0
0	0	0	1	0
? 0.6	0	1	1	0
? 0.7	0	1	0	1

semi-supervised learning

EM and estimating θ

- E step: Calculate for each training example k, the expected value of each unobserved variable Y



$P_\theta(Y=k)$

$P_\theta(X_1|Y=k)$

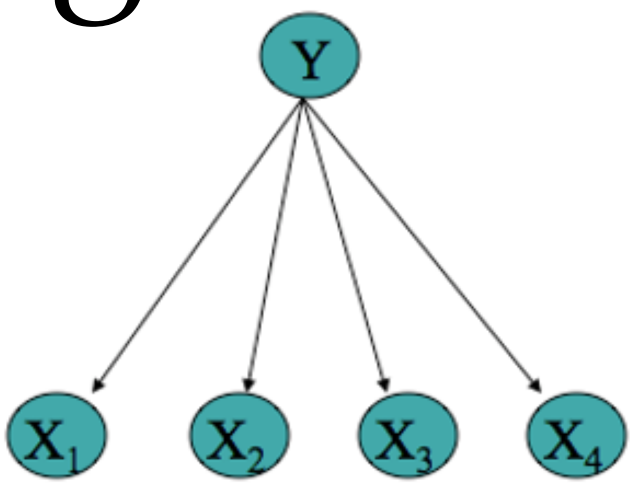
$$\boxed{\mathbb{E}[Y=k]} = \boxed{P_\theta(Y_k | x_1 \dots x_4)}$$

$$= \frac{P_\theta(x_1 \dots x_4 | Y=k)\, P_\theta(Y=k)}{\sum_j P_\theta(x_1 \dots x_4 | Y=j)\, P_\theta(Y=j)}$$

Bayes Rule

$$\prod_{i=1}^{4} P(x_i | Y=k)$$

EM and estimating θ



- Observed set X
- Partially unobserved set Y of boolean values

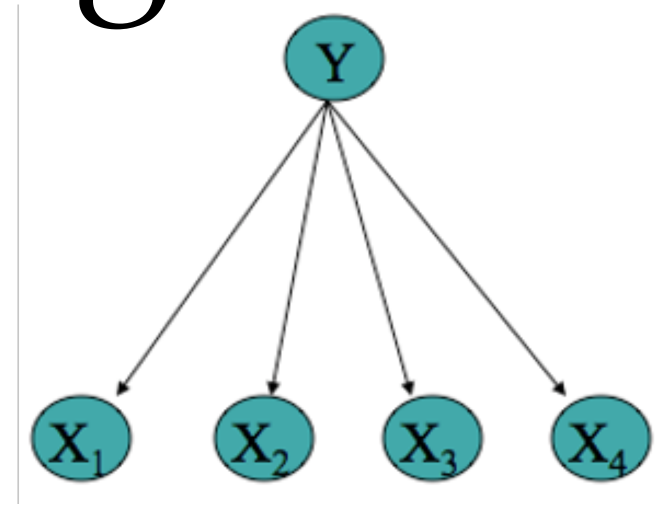
- E step: Calculate for each training example k, the expected value of each unobserved variable Y

$$E_{P(Y|X_1 \dots X_N)}[y(k)] = P(y(k)=1|x_1(k), \dots x_N(k); \theta) = \boxed{\frac{P(y(k)=1)\prod_i P(x_i(k)|y(k)=1)}{\sum_{j=0}^{1} P(y(k)=j)\prod_i P(x_i(k)|y(k)=j)}}$$

- M step: Calculate estimates similar to MLE, but replacing each count by its <u>expected count</u>

EM and estimating θ

- Observed set X
- Partially unobserved set Y of boolean values



- E step: Calculate for each training example k, the expected value of each unobserved variable Y

$$E_{P(Y|X_1 \dots X_N)}[y(k)] = P(y(k) = 1 | x_1(k), \dots x_N(k); \theta) = \frac{P(y(k)=1) \prod_i P(x_i(k)|y(k)=1)}{\sum_{j=0}^{1} P(y(k)=j) \prod_i P(x_i(k)|y(k)=j)}$$

- M step: Calculate estimates similar to MLE, but replacing each count by its <u>expected count</u>

$$\theta_{ij|m} = \hat{P}(X_i = j | Y = m) = \frac{\sum_k P(y(k) = m | x_1(k) \dots x_N(k)) \; \delta(x_i(k) = j)}{\sum_k P(y(k) = m | x_1(k) \dots x_N(k))}$$

MLE would be: $\hat{P}(X_i = j | Y = m) = \frac{\sum_k \delta((y(k) = m) \wedge (x_i(k) = j))}{\sum_k \delta(y(k) = m)}$

EM algorithm — summary

- EM is a general procedure for learning from partly observed data
- Given observed variables X, unobserved Z (X={F,A,H,N}, Z={S})
- Define $Q(\theta'|\theta) = E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$

Begin with arbitrary choice for parameters θ

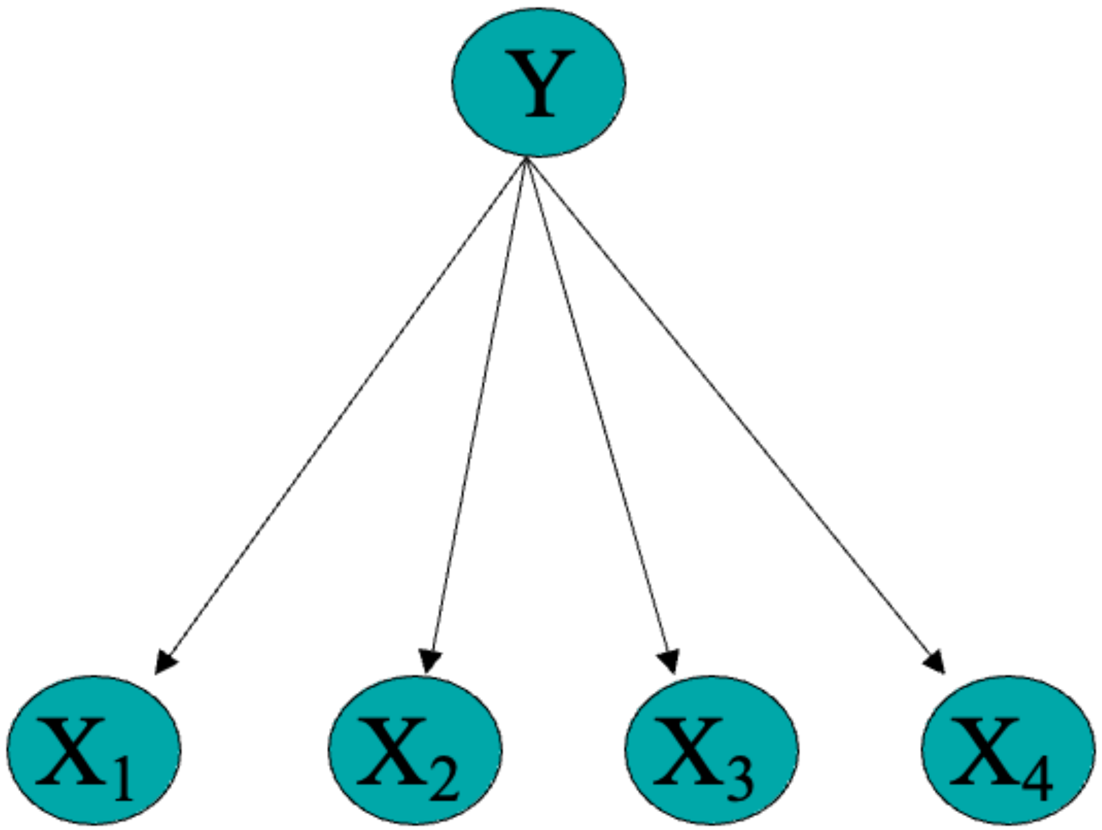
Iterate until convergence:

- E Step: Use X and current θ to calculate P(Z|X,θ)
- M Step: Replace current θ by $\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$

- Guaranteed to find local maximum. Each iteration increases

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

What if we have no labeled data at all ?



Y	X1	X2	X3	X4
?	0	0	1	1
?	0	1	0	0
?	0	0	1	0
?	0	1	1	0
?	0	1	0	1

un
~~semi-~~supervised learning

Unsupervised clustering

Just extreme case of EM with zero labeled examples…