# Machine Learning 10-601, 10-301

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

April 19, 2021

Today:

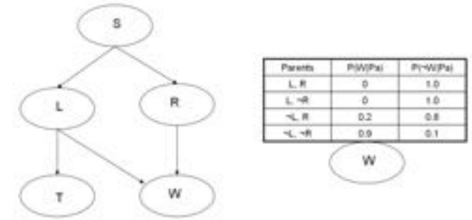
- Graphical models 3
  - Learning Bayesian Networks

Readings:

- Bishop chapter 8

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/Bishop-PRML-sample.pdf

# Bayesian Networks Definition




A Bayes network represents the joint probability distribution over a collection of random variables

A Bayes network is a directed acyclic graph and a set of Conditional Probability Distributions (CPD's)

- Each node denotes a random variable
- Edges denote dependencies
- For each node $X_i$ its CPD defines $P(X_i \mid Pa(X_i))$
- The joint distribution over all variables is defined to be

$$P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

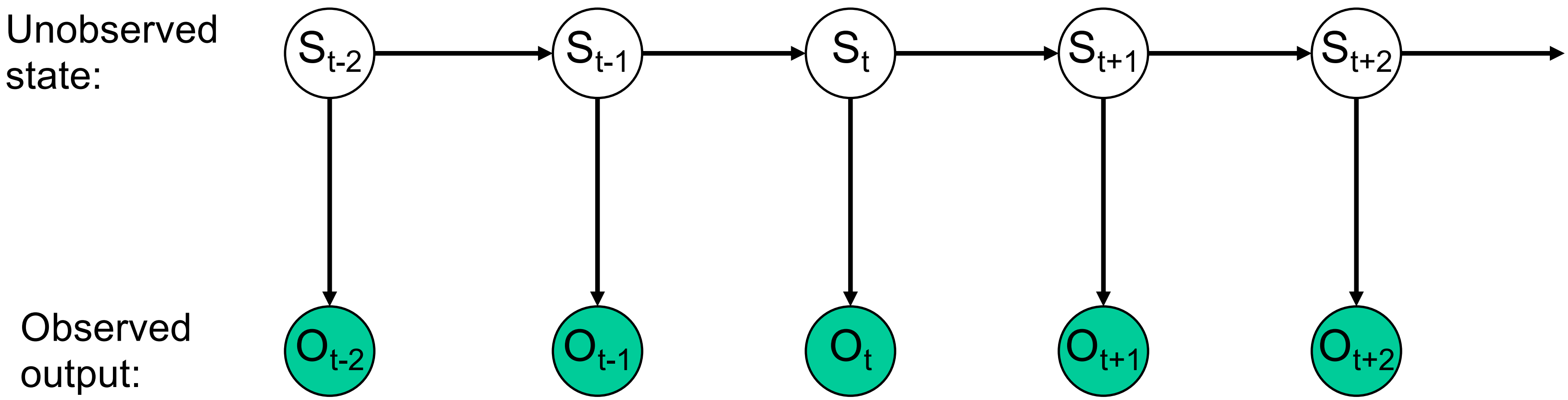
Pa(X) = immediate parents of X in the graph

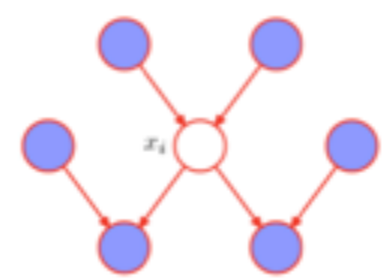
Poll:  Answer Question 1

What is the graphical model for a Naïve Bayes classifier?

# Bayes Network for a Hidden Markov Model

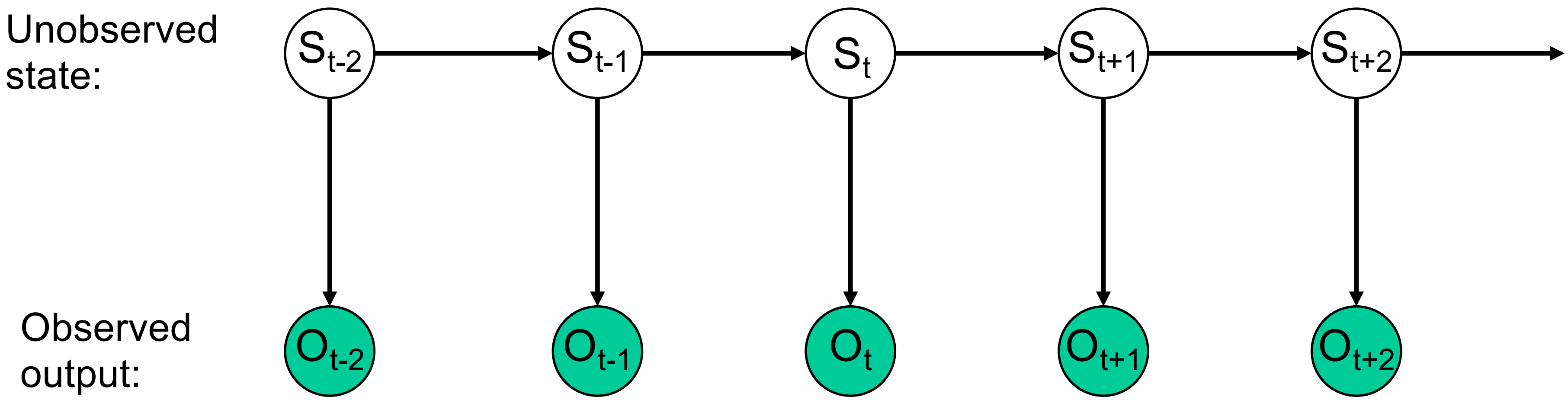
Implies the future is conditionally independent of the past, given the present



# Bayes Network for a Hidden Markov Model

Implies the future is conditionally independent of the past, given the present




$$P(S_{t-2}, O_{t-2}, S_{t-1}, \ldots, O_{t+2}) =$$

# Learning of Bayes Nets

Four types of learning problems

- Graph structure may be known/unknown
- Variable values may be fully observed / partly unobserved

1. Easy case: learn parameters when graph structure is *known*, and training data is *fully observed*
2. Interesting case: graph *known*, data *partly observed*
3. Interesting case: graph un*known,* data *fully observed*
4. Gruesome case: graph structure *unknown*, data *partly unobserved*

# Easy: Graph Known, Fully Observed Data

- Example: Consider learning the parameter

$$\theta_{K=1|S=0,A=1} \equiv P(K=1|S=0,A=1)$$

- Max Likelihood Estimate is

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

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} \delta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} \delta(s_m = i, a_m = j)}$$

$m^{th}$ training example

$\delta(X)$ = 1 if X is true
0 otherwise


| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

let's use $a_m$ to represent value of A on the mth example

# Easy: Graph Known, Fully Observed Data

- Example: Consider learning the parameter

$$\theta_{K=1|S=0,A=1} \equiv P(K=1|S=0,A=1)$$


| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

Poll: Answer Question 2

What is the Maximum Likehood estimate of P(K=1|S=0,A=1)

let's use $a_m$ to represent value of $A$ on the mth example

# Easy: Graph Known, Fully Observed Data

- Example: Consider learning the parameter

$$\theta_{K=1|S=0,A=1} \equiv P(K=1|S=0,A=1)$$


| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

- Max Likelihood Estimate is

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

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} \delta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} \delta(s_m = i, a_m = j)}$$

$m^{th}$ training example

δ(X) = 1 if X is true
0 otherwise

let's use $a_m$ to represent value of A on the mth example


## Fully Observed

| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

## Partially Observed

| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | ? | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | ? | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

- Max Likelihood Estimate is

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

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} \delta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} \delta(s_m = i, a_m = j)}$$

$m^{th}$ training example

δ(X) = 1 if X is true
0 otherwise

let's use $a_m$ to represent value of A on the mth example

## Fully Observed

| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

## Partially Observed

| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | ? | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | ? | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

## EM Approach

| Pr | S | A | K | E | H |
|---|---|---|---|---|---|
| 1.0 | 1 | 0 | 1 | 1 | 0 |
| 0.6 | 0 | 1 | 1 | 0 | 1 |
| 0.4 | 0 | 1 | 0 | 0 | 1 |
| 1.0 | 0 | 1 | 1 | 0 | 0 |
| 0.2 | 0 | 1 | 0 | 1 | 0 |
| 0.8 | 0 | 0 | 0 | 1 | 0 |
| 1.0 | 1 | 1 | 1 | 1 | 1 |

- Max Likelihood Estimate is

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

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} \delta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} \delta(s_m = i, a_m = j)}$$

$m^{th}$ training example

δ(X) = 1 if X is true
0 otherwise

- EM Estimate is:

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j)}$$

let's use $a_m$ to represent value of A on the mth example

## Fully Observed

| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

$$\theta_{K=1|S=0,A=1} =$$

## Partially Observed

| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | ? | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | ? | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

## EM Approach

| Pr | S | A | K | E | H |
|---|---|---|---|---|---|
| 1.0 | 1 | 0 | 1 | 1 | 0 |
| 0.6 | 0 | 1 | 1 | 0 | 1 |
| 0.4 | 0 | 1 | 0 | 0 | 1 |
| 1.0 | 0 | 1 | 1 | 0 | 0 |
| 0.2 | 0 | 1 | 0 | 1 | 0 |
| 0.8 | 0 | 0 | 0 | 1 | 0 |
| 1.0 | 1 | 1 | 1 | 1 | 1 |

- Max Likelihood Estimate is

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

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} \delta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} \delta(s_m = i, a_m = j)}$$

$m^{th}$ training example

δ(X) = 1 if X is true
0 otherwise

- EM Estimate

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j)}$$

$$\theta_{K=1|S=0,A=1} =$$

let's use $a_m$ to represent value of A on the mth example

## Fully Observed

| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

$$\theta_{K=1|S=0,A=1} = \frac{1}{1+1}$$

## Partially Observed

| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | ? | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | ? | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

## EM Approach

| Pr | S | A | K | E | H |
|---|---|---|---|---|---|
| 1.0 | 1 | 0 | 1 | 1 | 0 |
| 0.6 | 0 | 1 | 1 | 0 | 1 |
| 0.4 | 0 | 1 | 0 | 0 | 1 |
| 1.0 | 0 | 1 | 1 | 0 | 0 |
| 0.2 | 0 | 1 | 0 | 1 | 0 |
| 0.8 | 0 | 0 | 0 | 1 | 0 |
| 1.0 | 1 | 1 | 1 | 1 | 1 |

- Max Likelihood Estimate is

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

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} \delta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} \delta(s_m = i, a_m = j)}$$

$m^{th}$ training example

$\delta(X)$ = 1 if X is true
0 otherwise

- EM Estimate $\theta_{K=1|S=0,A=1} = \frac{0.6+1+0.2}{0.6+0.4+1+0.2}$

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j)}$$

- Fractional examples
- Replace $\delta(X)$ by $P(X)$
- Replaces counts by expected values
- iterate!

let's use $a_m$ to represent value of A on the mth example

# EM algorithm


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

## EM algorithm:

- Iterate until convergence:
  - E step: use current Bayes net parameters θ to estimate unobserved Z values

| S | A | K | E | H |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | ? | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | ? | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

| Pr | S | A | K | E | H |
|---|---|---|---|---|---|
| 1.0 | 1 | 0 | 1 | 1 | 0 |
| 0.6 | 0 | 1 | 1 | 0 | 1 |
| 0.4 | 0 | 1 | 0 | 0 | 1 |
| 1.0 | 0 | 1 | 1 | 0 | 0 |
| 0.2 | 0 | 1 | 0 | 1 | 0 |
| 0.8 | 0 | 0 | 0 | 1 | 0 |
| 1.0 | 1 | 1 | 1 | 1 | 1 |

  - M step: use estimated values of Z to retrain Bayes net params θ

$$\theta_{K=1|S=i,A=j} = \frac{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j)}$$

Expected value =
probability weighted average

$$E_{P(X)}[f(X)] = \sum_{x} P(X = x) f(x)$$

# Expected value = probability weighted average

$$E_{P(X)}[f(X)] = \sum_x P(X = x) f(x)$$

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

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

$$E_{P(Z|X,\theta)}[\log P(X, Z|\theta)] = \sum_z P(Z = z|X, \theta) \; \log(P(X, Z = z|\theta)$$

* EM guaranteed to find local maximum

- EM seeks estimate:

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


- suppose for every example, observed X={S,A,E,H}, unobserved Z={K}
- how do we calculate $E_{P(Z|X,\theta)} \log P(X, Z|\theta)$ over our M training examples?

$$\log P(X, Z|\theta) = \sum_{m=1}^{M} \log P(s_m) + \log P(a_m) + \log P(k_m|s_m, a_m) + \log P(e_m|k_m) + \log P(h_m|k_m)$$

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

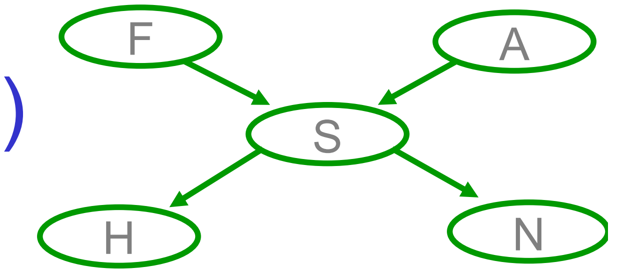
$$= \sum_{m=1}^{M} \sum_{i=0}^{i=1} P(k=i|s_m, a_m, e_m, h_m) \left[\log P(s_m) + \log P(a_m) + \log P(k=i|s_m, a_m) + \log P(e_m|k=i) + \log P(h_m|k=i)\right]$$

$m^{th}$ training example

This corresponds to splitting each training example into two probabilistically weighted examples

let's use $a_m$ to represent value of A on the mth example

# E Step: Use X, $\theta$, to Calculate $P(Z|X,\theta)$



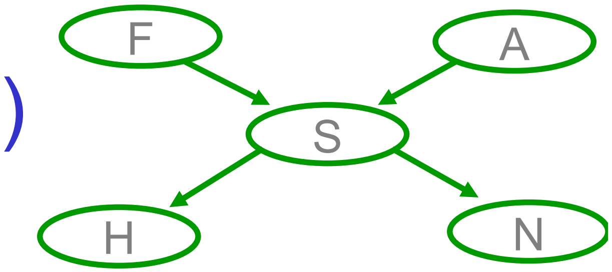

observed X={F,A,H,N},
unobserved Z={S}

How? Bayes net inference problem.

$$P(S_k = 1 | f_k a_k h_k n_k, \theta) =$$

let's use $a_k$ to represent value of $A$ on the $k$th example

# 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.

$$P(S_k = 1 | f_k a_k h_k n_k, \theta) =$$

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

let's use $a_k$ to represent value of *A* on the kth example

# EM in general

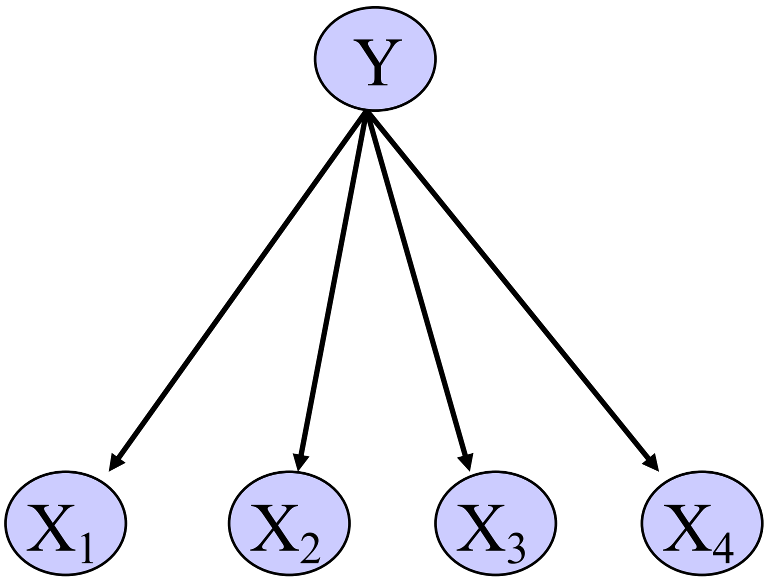
- Unobserved data points can be any combination of variables sometimes observed , sometimes not

- You can build a MAP version instead of MLE version of EM

- Basis for many important algorithms
  - Hidden markov models
  - Unsupervised clustering

# Using Unlabeled Data to Help Train Naïve Bayes Classifier

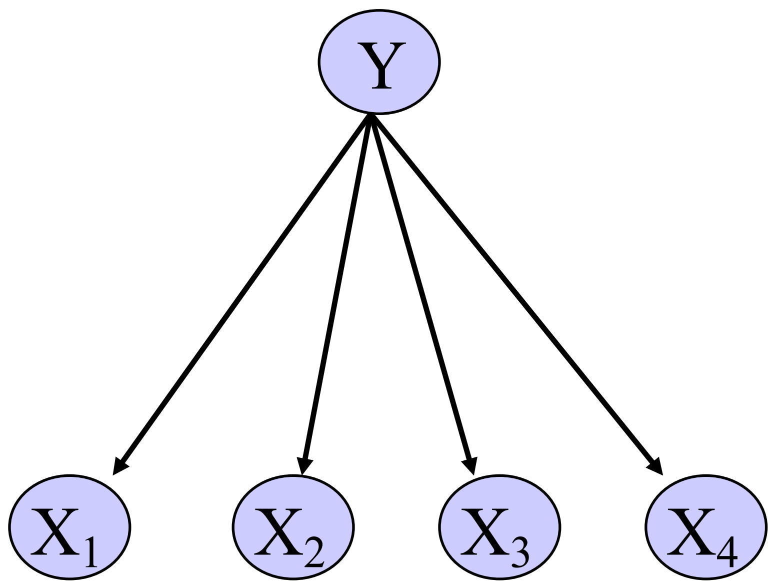
Learn P(Y|X)




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

# Using Unlabeled Data to Help Train Naïve Bayes Classifier

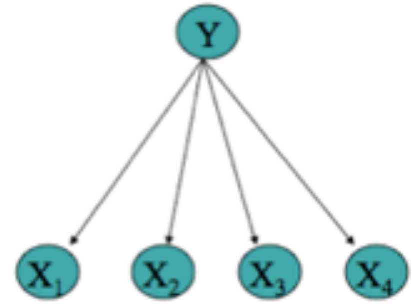
Learn P(Y|X)




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

$$P(Y=1 \mid x_1, x_2, x_3, x_4) = \frac{P(Y=1) \prod_i P(x_i \mid Y=1)}{\sum_k P(Y=k) \prod_i P(x_i \mid Y=k)}$$

# EM and estimating $\theta$




Given observed set X, unobserved set Y of boolean values

E step: Calculate for each training example, k

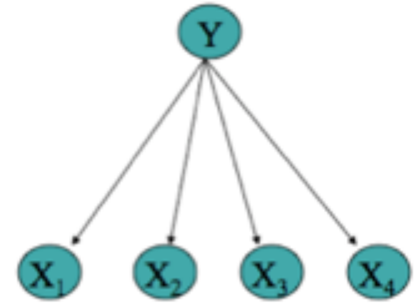
the expected value of each unobserved value of variable Y

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

$k^{th}$ training example

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

# EM and estimating $\theta$




Given observed set X, unobserved set Y of boolean values

E step: Calculate for each training example, k

the expected value of each unobserved variable Y

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

k[th] training example

M step: Calculate estimates similar to MLE, but replacing each count by its expected count

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

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

- **Inputs:** Collections $\mathcal{D}^l$ of labeled documents and $\mathcal{D}^u$ of unlabeled documents.
- Build an initial naive Bayes classifier, $\hat{\theta}$, from the labeled documents, $\mathcal{D}^l$, only. Use maximum a posteriori parameter estimation to find $\hat{\theta} = \arg\max_\theta P(\mathcal{D}|\theta)P(\theta)$ (see Equations 5 and 6).
- Loop while classifier parameters improve, as measured by the change in $l_c(\theta|\mathcal{D};\mathbf{z})$ (the complete log probability of the labeled and unlabeled data
  - **(E-step)** Use the current classifier, $\hat{\theta}$, to estimate component membership of each unlabeled document, *i.e.*, the probability that each mixture component (and class) generated each document, $P(c_j|d_i;\hat{\theta})$ (see Equation 7).
  - **(M-step)** Re-estimate the classifier, $\hat{\theta}$, given the estimated component membership of each document. Use maximum a posteriori parameter estimation to find $\hat{\theta} = \arg\max_\theta P(\mathcal{D}|\theta)P(\theta)$ (see Equations 5 and 6).
- **Output:** A classifier, $\hat{\theta}$, that takes an unlabeled document and predicts a class label.

From [Nigam et al., 2000]

# Experimental Evaluation

- Newsgroup postings
  - 20 newsgroups, 1000/group
- Web page classification
  - student, faculty, course, project
  - 4199 web pages
- Reuters newswire articles
  - 12,902 articles
  - 90 topics categories

# 20 Newsgroups

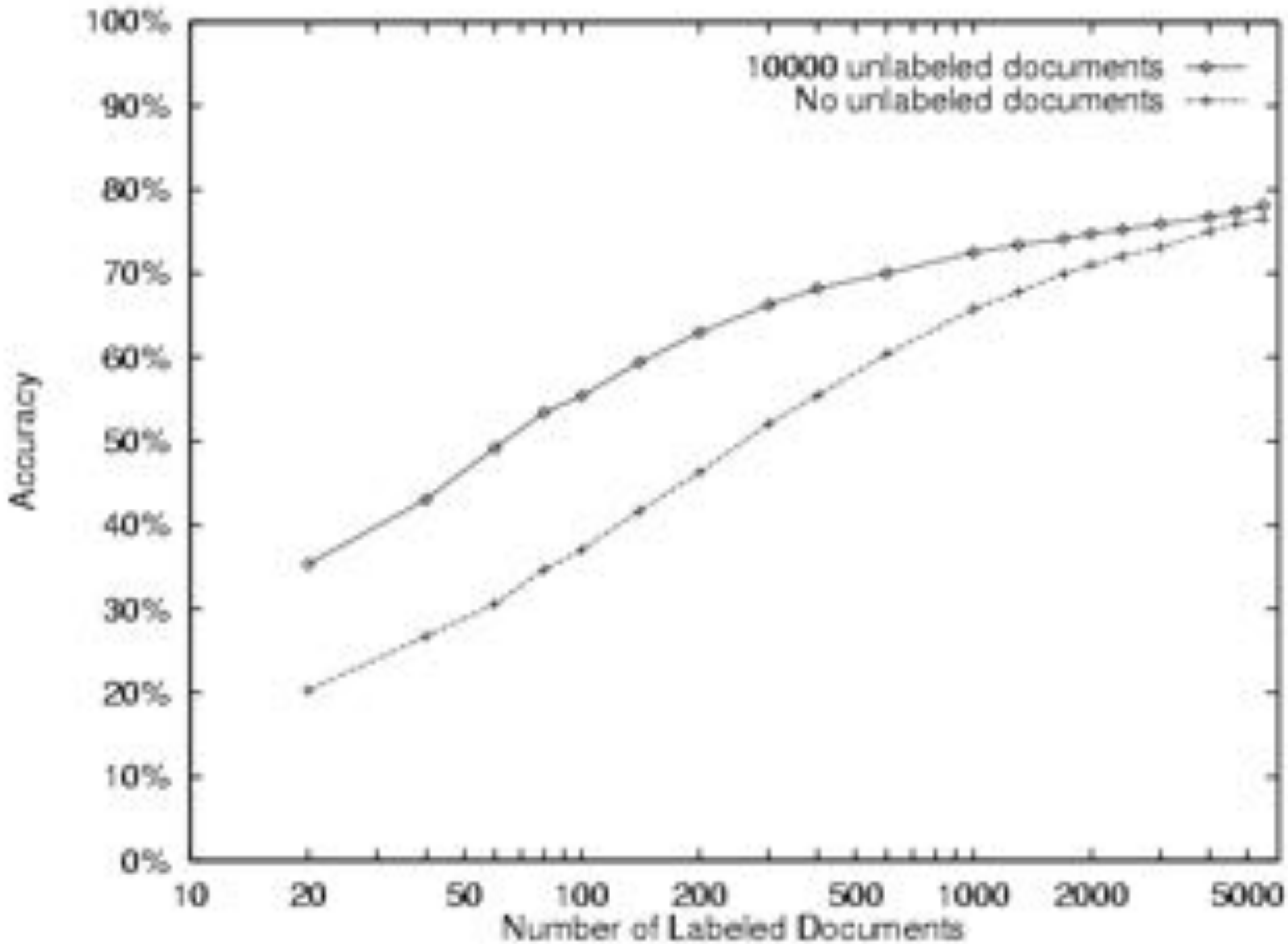


Table 3. Lists of the words most predictive of the course class in the WebKB data set, as they change over iterations of EM for a specific trial. By the second iteration of EM, many common course-related words appear. The symbol $D$ indicates an arbitrary digit.

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

word w ranked by P(w|Y=course) /P(w|Y ≠ course)

Using one labeled example per class

# What you should know about EM

- For learning from partly unobserved data
- MLE 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

- EM for training Bayes networks
- Can also develop MAP version of EM
- Can also derive your own EM algorithm for your own problem
  - write out expression for $E_{Z|X,\theta}[\log P(X, Z|\theta)]$
  - E step: for each training example $X^k$, calculate $P(Z^k \mid X^k, \theta)$
  - M step: chose new $\theta$ to maximize $E_{Z|X,\theta}[\log P(X, Z|\theta)]$

# Usupervised clustering

Just extreme case for EM with zero labeled examples…

# Clustering

- Given set of data points, group them
- Unsupervised learning
- Which documents are similar? (or which patients, earthquakes, customers, faces, molecules, …)

# Mixture Distributions

Model joint $P(X_1 \ldots X_n)$ as mixture of multiple distributions.

Use discrete-valued random var Z to indicate which distribution is being use for each random draw

So

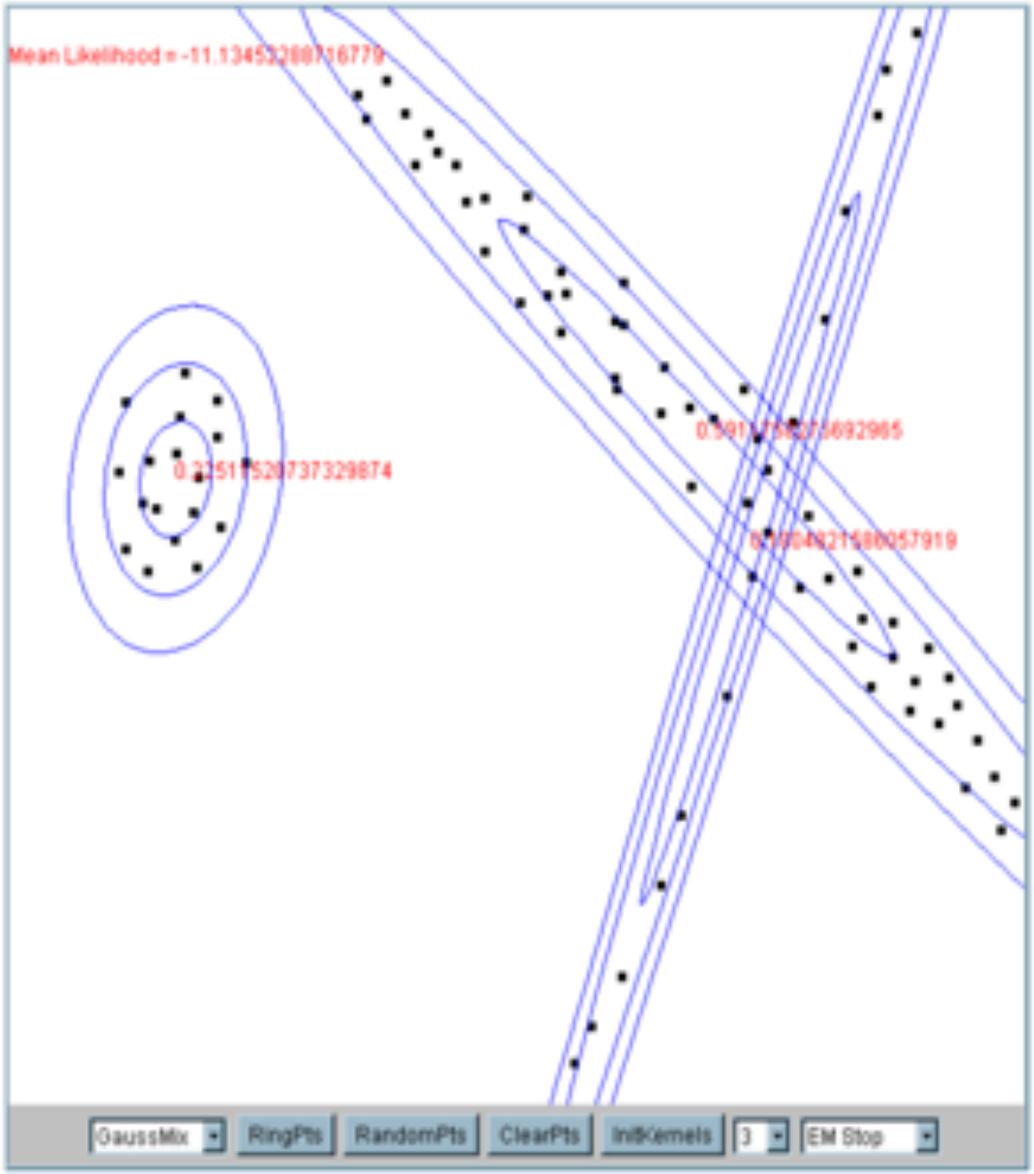
$$P(X_1 \ldots X_n) = \sum_i P(Z = i) \;\; P(X_1 \ldots X_n | Z)$$

Mixture of *Gaussians*:

- Assume each data point X=<X1, … Xn> is generated by one of several Gaussians, as follows:

1. randomly choose Gaussian i, according to P(Z=i)
2. randomly generate a data point <x1,x2 .. xn> according to $N(\mu_i, \Sigma_i)$

# Mixture of Gaussians



# EM for Mixture of Gaussian Clustering

Let's simplify to make this easier:

1. assume $X = \langle X_1 \ldots X_n \rangle$, and the $X_i$ are conditionally independent given $Z$.

$$P(X|Z=j) = \prod_i N(X_i|\mu_{ji}, \sigma_{ji})$$

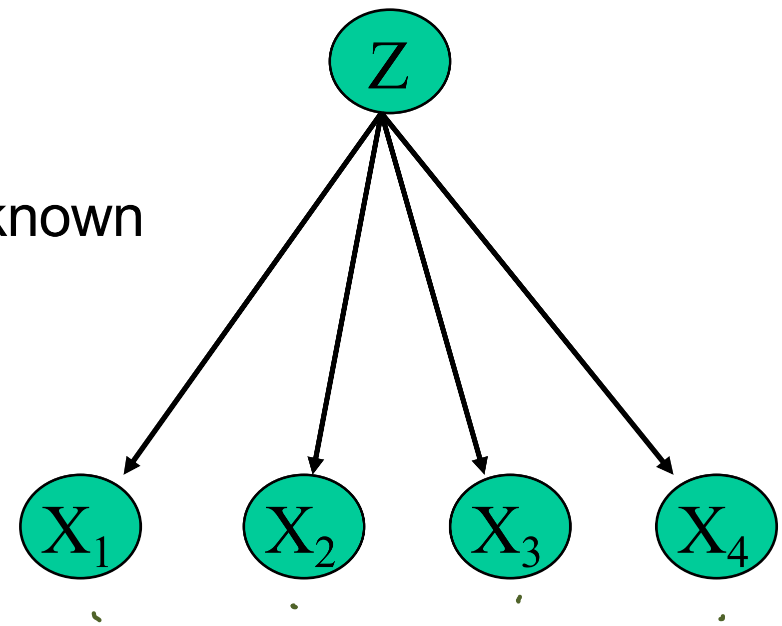
2. assume only 2 clusters (values of Z), and $\forall i, j, \sigma_{ji} = \sigma$

$$P(\mathbf{X}) = \sum_{j=1}^{2} P(Z=j|\pi) \prod_i N(x_i|\mu_{ji}, \sigma)$$

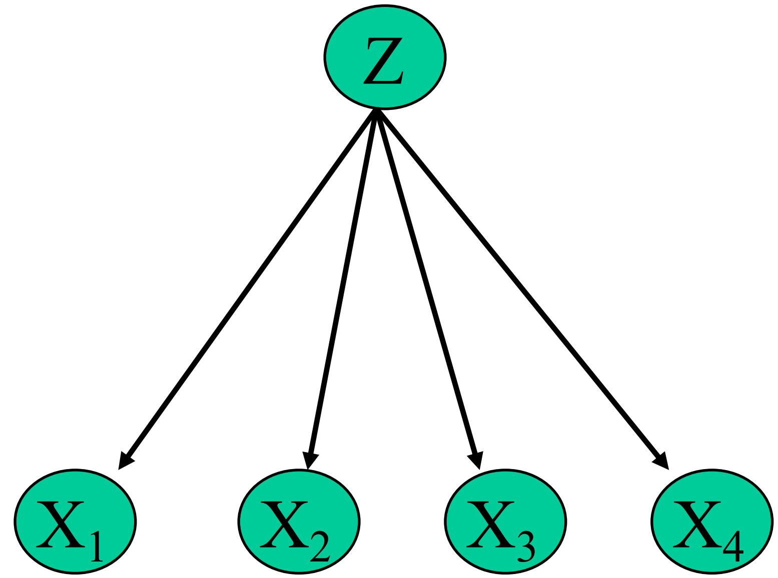
3. Assume σ known, $\pi_1 \ldots \pi_K, \mu_{1i} \ldots \mu_{Ki}$ unknown

Observed: $X = \langle X_1 \ldots X_n \rangle$

Unobserved: $Z$



# EM




Given observed variables X, unobserved Z

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

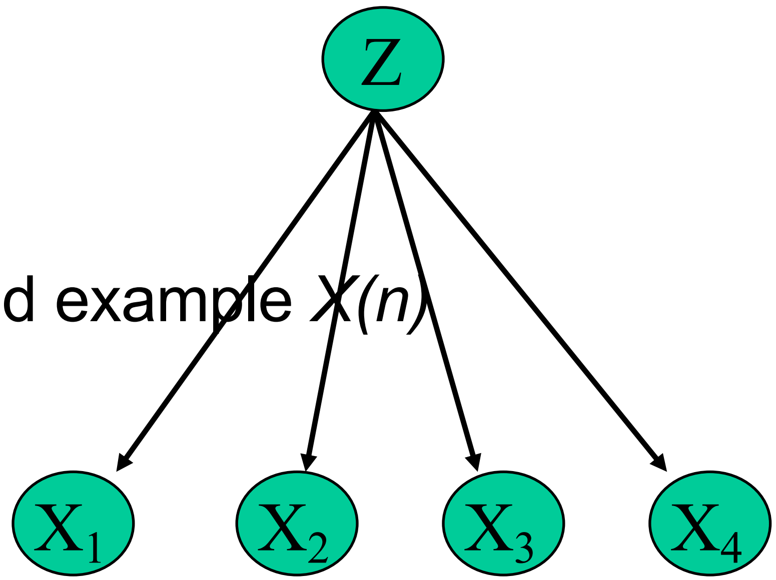
where $\theta = \langle \pi, \mu_{ji} \rangle$

Iterate until convergence:

- E Step: Calculate $P(Z(n)|X(n), \theta)$ for each example $X(n)$. Use this to construct $Q(\theta'|\theta)$
- M Step: Replace current $\theta$ by

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

# EM – E Step




Calculate $P(Z(n)|X(n),\theta)$ for each observed example $X(n)$

$X(n)=<x_1(n), x_2(n), \ldots x_T(n)>$.

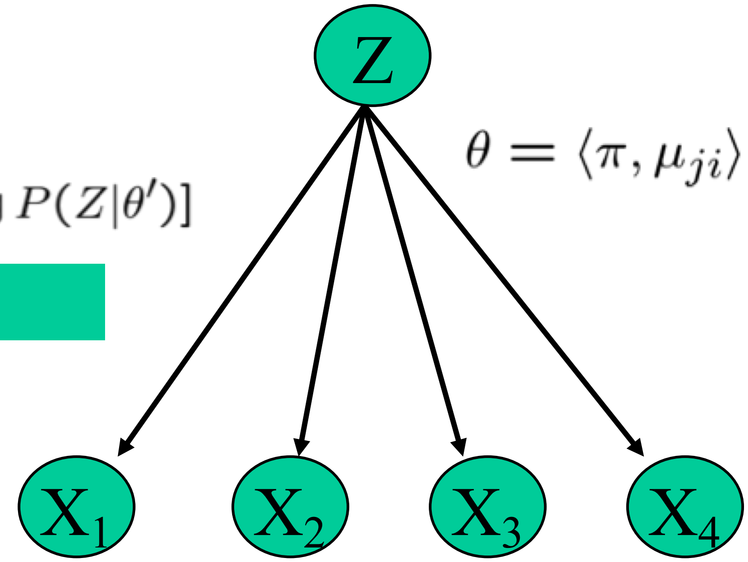
$$P(z(n)=k|x(n),\theta) = \frac{P(x(n)|z(n)=k,\theta) \quad P(z(n)=k|\theta)}{\sum_{j=0}^{1} p(x(n)|z(n)=j,\theta) \; P(z(n)=j|\theta)}$$

$$P(z(n)=k|x(n),\theta) = \frac{\prod_i P(x_i(n)|z(n)=k,\theta)] \quad P(z(n)=k|\theta)}{\sum_{j=0}^{1} \prod_i P(x_i(n)|z(n)=j,\theta) \; P(z(n)=j|\theta)}$$

$$P(z(n)=k|x(n),\theta) = \frac{\prod_i N(x_i(n)|\mu_{k,i},\sigma)] \quad (\pi^k(1-\pi)^{(1-k)})}{\sum_{j=0}^{1}[\prod_i N(x_i(n)|\mu_{j,i},\sigma)] \quad (\pi^j(1-\pi)^{(1-j)}))}$$

# EM – M Step

First consider update for π




$\theta = \langle \pi, \mu_{ji} \rangle$

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

π' has no influence

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

z=1 for nth example

$$E_{Z|X,\theta}\left[\log P(Z|\pi')\right] = E_{Z|X,\theta}\left[\log\left(\pi'^{\sum_n z(n)}(1-\pi')^{\sum_n (1-z(n))}\right)\right]$$

$$= E_{Z|X,\theta}\left[\left(\sum_n z(n)\right)\log \pi' + \left(\sum_n (1-z(n))\right)\log(1-\pi')\right]$$

$$= \left(\sum_n E_{Z|X,\theta}[z(n)]\right)\log \pi' + \left(\sum_n E_{Z|X,\theta}[(1-z(n))]\right)\log(1-\pi')$$

$$\frac{\partial E_{Z|X,\theta}[\log P(Z|\pi')]}{\partial \pi'} = \left(\sum_n E_{Z|X,\theta}[z(n)]\right)\frac{1}{\pi'} + \left(\sum_n E_{Z|X,\theta}[(1-z(n))]\right)\frac{(-1)}{1-\pi'}$$

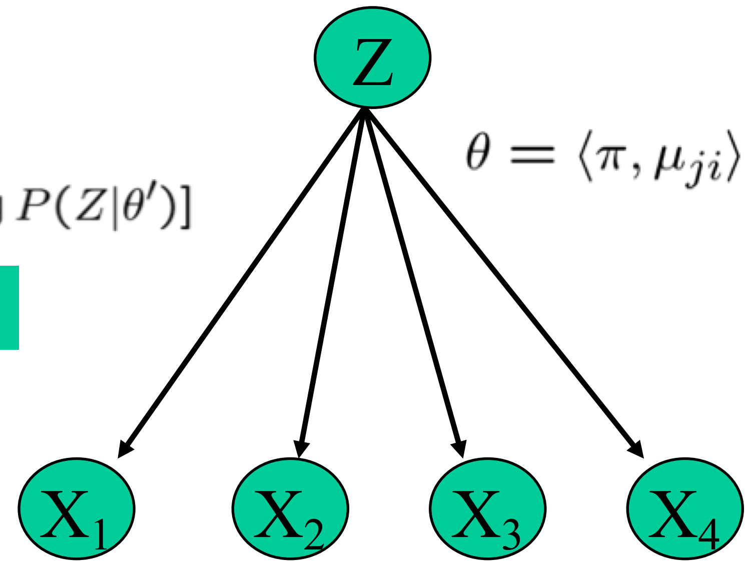
$$\pi \leftarrow \frac{\sum_{n=1}^N E[z(n)]}{\left(\sum_{n=1}^N E[z(n)]\right) + \left(\sum_{n=1}^N (1-E[z(n)])\right)} = \frac{1}{N}\sum_{n=1}^N E[z(n)]$$

# EM – M Step

Now consider update for $\mu_{ji}$

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

$\mu_{ji}$' has no influence




$\theta = \langle \pi, \mu_{ji} \rangle$

$$\mu_{ji} \leftarrow \arg\max_{\mu'_{ji}} E_{Z|X,\theta}[\log P(X|Z,\theta')]$$

…

…

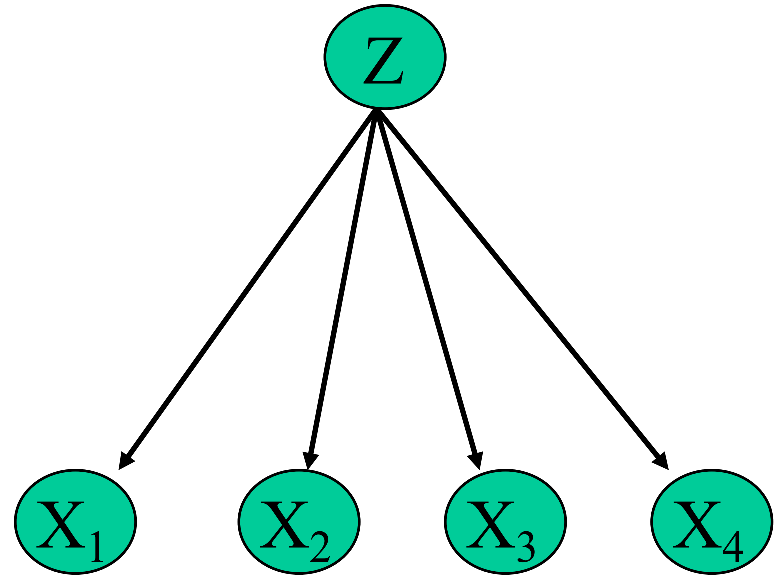
…

$$\mu_{ji} \leftarrow \frac{\sum_{n=1}^{N} P(z(n)=j|x(n),\theta) \quad x_i(n)}{\sum_{n=1}^{N} P(z(n)=j|x(n),\theta)}$$

Compare above to MLE if Z were observable:

$$\mu_{ji} \leftarrow \frac{\sum_{n=1}^{N} \delta(z(n)=j) \quad x_i(n)}{\sum_{n=1}^{N} \delta(z(n)=j)}$$

# EM – putting it together




Given observed variables X, unobserved Z

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

where $\theta = \langle \pi, \mu_{ji} \rangle$

Iterate until convergence:

- E Step: For each observed example X(n), calculate $P(Z(n)|X(n), \theta)$

$$P(z(n) = k \mid x(n), \theta) = \frac{[\prod_i N(x_i(n)|\mu_{k,i}, \sigma)] \quad (\pi^k (1-\pi)^{(1-k)})}{\sum_{j=0}^{1} [\prod_i N(x_i(n)|\mu_{j,i}, \sigma)] \quad (\pi^j (1-\pi)^{(1-j)})}$$

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

P(z=1)

$$\pi \leftarrow \frac{1}{N} \sum_{n=1}^{N} E[z(n)]$$

$$\mu_{ji} \leftarrow \frac{\sum_{n=1}^{N} P(z(n) = j|x(n), \theta) \quad x_i(n)}{\sum_{n=1}^{N} P(z(n) = j|x(n), \theta)}$$