



Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 25, 2015

Today:

- Graphical models
- Bayes Nets:
 - Inference
 - Learning
 - EM

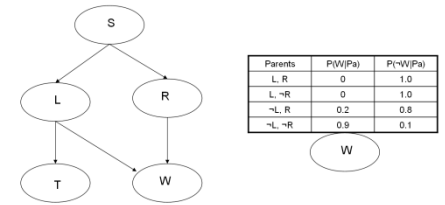
Readings:

- Bishop chapter 8
- Mitchell chapter 6

Midterm

- In class on Monday, March 2
- Closed book
- You may bring a 8.5x11 “cheat sheet” of notes
- Covers all material through today
- Be sure to come on time. We’ll start precisely at 12 noon

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 $P(X_i | Pa(X_i))$
- The joint distribution over all variables is defined to be

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

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

What You Should Know

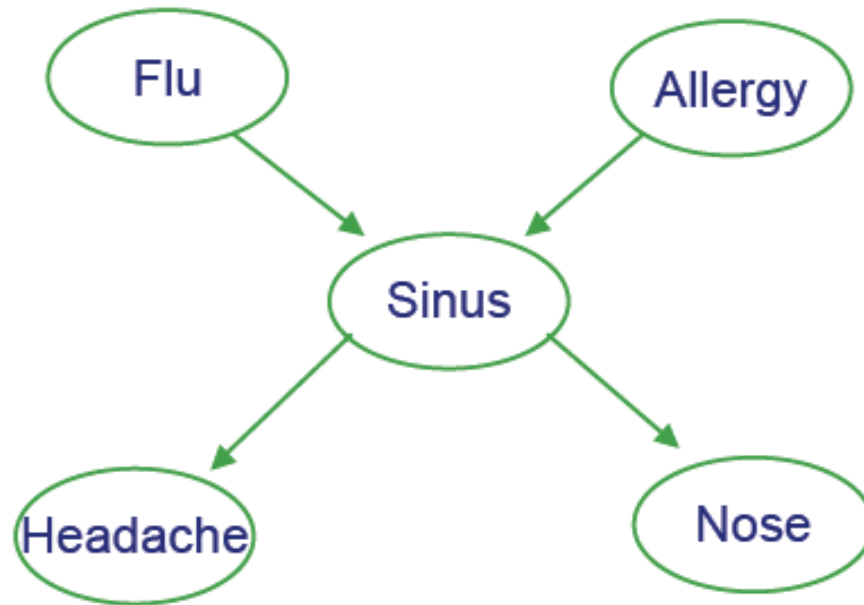
- Bayes nets are convenient representation for encoding dependencies / conditional independence
- BN = Graph plus parameters of CPD's
 - Defines joint distribution over variables
 - Can calculate everything else from that
 - Though inference may be intractable
- Reading conditional independence relations from the graph
 - Each node is cond indep of non-descendants, given only its parents
 - X and Y are conditionally independent given Z if Z D-separates every path connecting X to Y
 - Marginal independence : special case where $Z=\{\}$

Inference in Bayes Nets

- In general, intractable (NP-complete)
- For certain cases, tractable
 - Assigning probability to fully observed set of variables
 - Or if just one variable unobserved
 - Or for singly connected graphs (ie., no undirected loops)
 - Belief propagation
- Sometimes use Monte Carlo methods
 - Generate many samples according to the Bayes Net distribution, then count up the results
- Variational methods for tractable approximate solutions

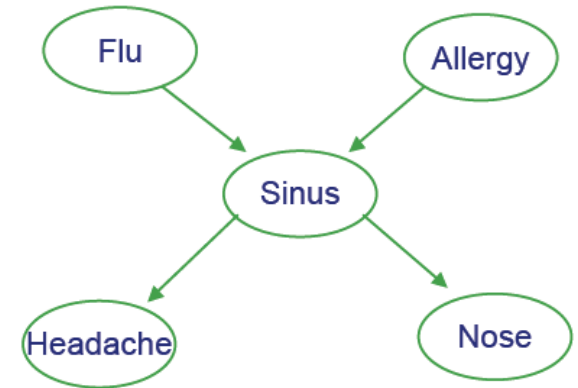
Example

- Bird flu and Allergies both cause Sinus problems
- Sinus problems cause Headaches and runny Nose



Prob. of joint assignment: easy

- Suppose we are interested in joint assignment $\langle F=f, A=a, S=s, H=h, N=n \rangle$

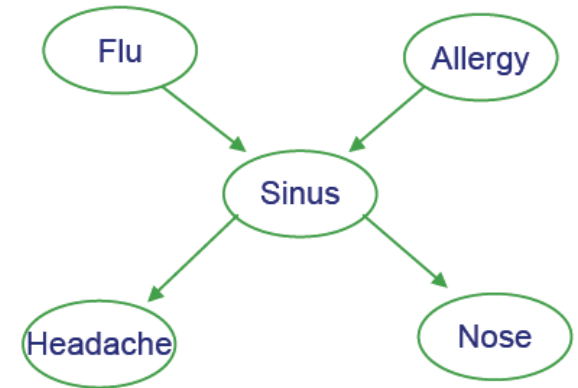


What is $P(f,a,s,h,n)$?

let's use $p(a,b)$ as shorthand for $p(A=a, B=b)$

Prob. of marginals: not so easy

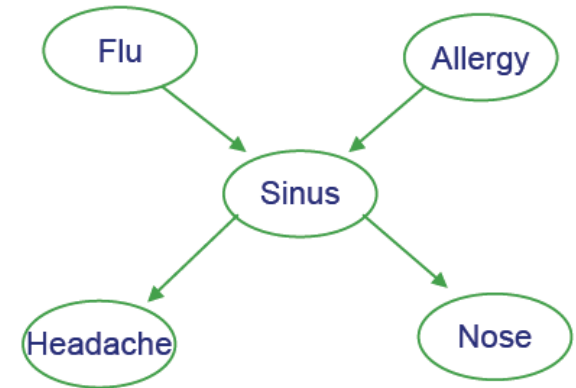
- How do we calculate $P(N=n)$?



let's use $p(a,b)$ as shorthand for $p(A=a, B=b)$

Generating a sample from joint distribution: easy

How can we generate random samples drawn according to $P(F,A,S,H,N)$?



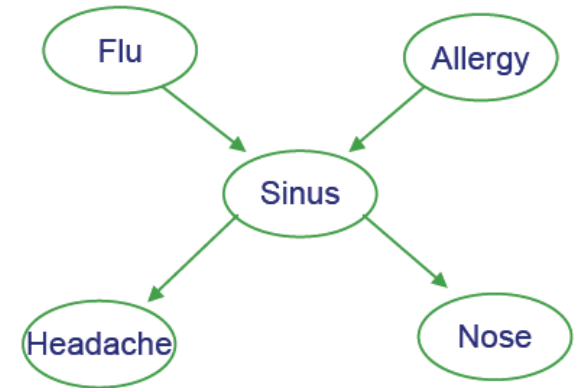
Hint: random sample of F according to $P(F=1) = \theta_{F=1}$:

- draw a value of r uniformly from $[0, 1]$
- if $r < \theta$ then output $F=1$, else $F=0$

let's use $p(a,b)$ as shorthand for $p(A=a, B=b)$

Generating a sample from joint distribution: easy

How can we generate random samples drawn according to $P(F,A,S,H,N)$?



Hint: random sample of F according to $P(F=1) = \theta_{F=1}$:

- draw a value of r uniformly from $[0,1]$
- if $r < \theta$ then output $F=1$, else $F=0$

Solution:

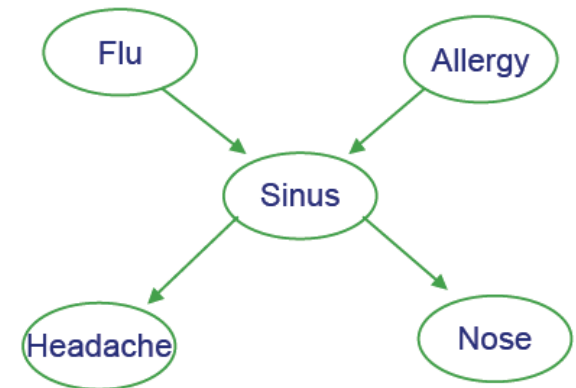
- draw a random value f for F , using its CPD
- then draw values for A , for $S|A,F$, for $H|S$, for $N|S$

Generating a sample from joint distribution: easy

Note we can estimate marginals like $P(N=n)$ by generating many samples from joint distribution, then count the fraction of samples for which $N=n$

Similarly, for anything else we care about $P(F=1|H=1, N=0)$

→ weak but general method for estimating any probability term...



Inference in Bayes Nets

- In general, intractable (NP-complete)
- For certain cases, tractable
 - Assigning probability to fully observed set of variables
 - Or if just one variable unobserved
 - Or for singly connected graphs (ie., no undirected loops)
 - Variable elimination
 - Belief propagation
- Often use Monte Carlo methods
 - e.g., Generate many samples according to the Bayes Net distribution, then count up the results
 - Gibbs sampling
- Variational methods for tractable approximate solutions

see Graphical Models course 10-708

Learning of Bayes Nets

- Four categories of learning problems
 - Graph structure may be known/unknown
 - Variable values may be fully observed / partly unobserved
- Easy case: learn parameters for graph structure is *known*, and data is *fully observed*
- Interesting case: graph *known*, data *partly known*
- Gruesome case: graph structure *unknown*, data *partly unobserved*

Learning CPTs from Fully Observed Data

- Example: Consider learning the parameter

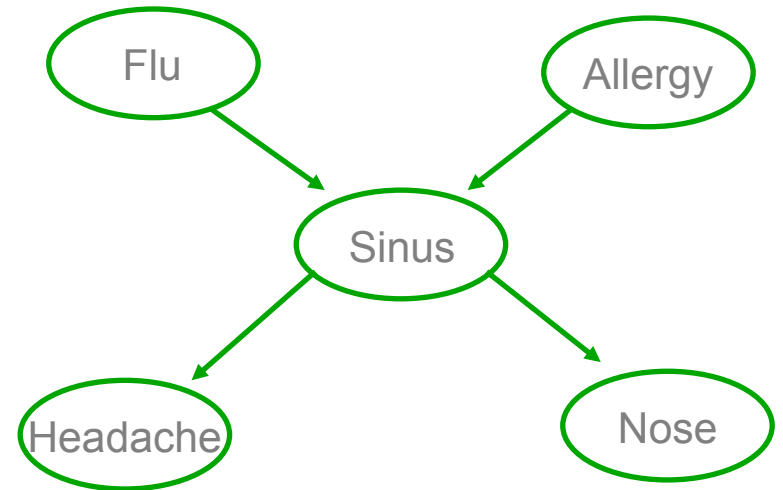
$$\theta_{s|ij} \equiv P(S = 1 | F = i, A = j)$$

- Max Likelihood Estimate is

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

kth training example

$\delta(x) = 1$ if x =true,
= 0 if x =false



- Remember why?

let's use $p(a,b)$ as shorthand for $p(A=a, B=b)$

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

- Maximum likelihood estimate

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

- Our case:

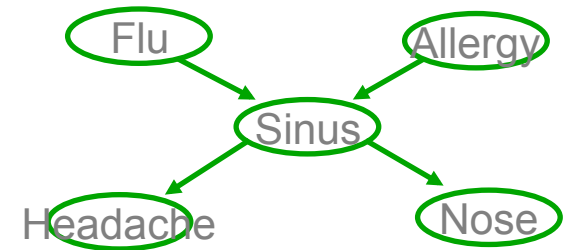
$$P(\text{data}|\theta) = \prod_{k=1}^K P(f_k, a_k, s_k, h_k, n_k)$$

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

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

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

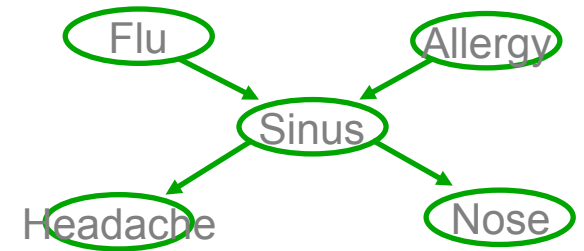
$$\theta_{s|ij} = \frac{\sum_{k=1}^K \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^K \delta(f_k = i, a_k = j)}$$



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:

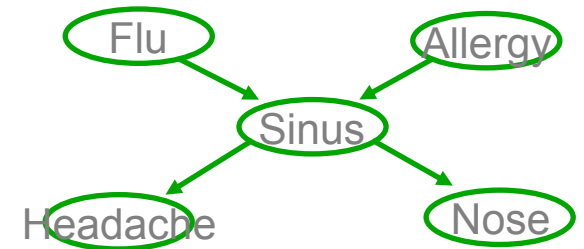
$$\theta \leftarrow \arg \max_{\theta} \log P(X, Z | \theta)$$

- 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:

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

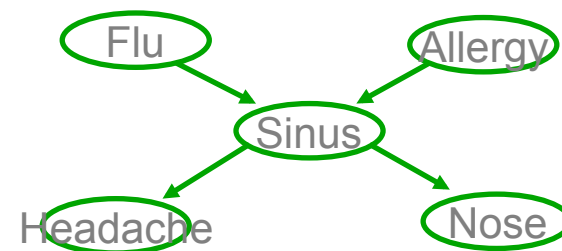
- EM seeks* to estimate:

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

* EM guaranteed to find local maximum

- EM seeks estimate:

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



- 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 , using θ
- M Step: use observed values plus E-step estimates to derive a better θ

Guaranteed to find local maximum.

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

EM Algorithm - Precisely

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 $Q(\theta'|\theta) = E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$
↑ current *↖ M step new*

Iterate until convergence:

- E Step: Use X and current θ to calculate $P(Z|X,\theta)$
- M Step: Replace current θ by

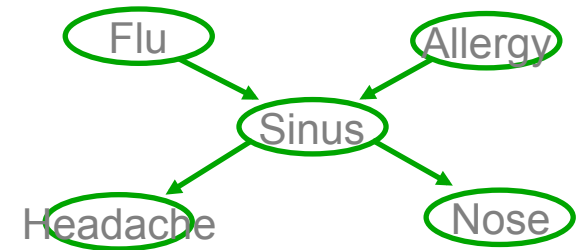
$$\theta \leftarrow \arg \max_{\theta'} Q(\theta'|\theta)$$

Guaranteed to find local maximum.

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

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

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



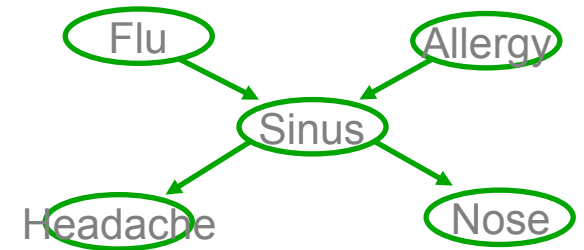
- How? Bayes net inference problem.

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

let's use $p(a,b)$ as shorthand for $p(A=a, B=b)$

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

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



- How? Bayes net inference problem.

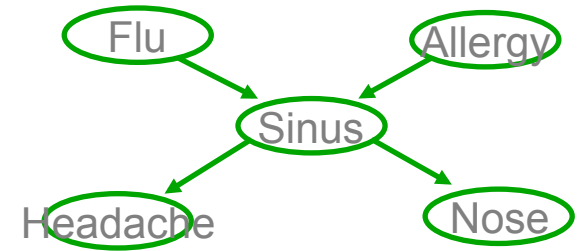
$$P(S_k = 1 | f_k a_k h_k n_k, \theta) =$$

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

let's use $p(a,b)$ as shorthand for $p(A=a, B=b)$

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

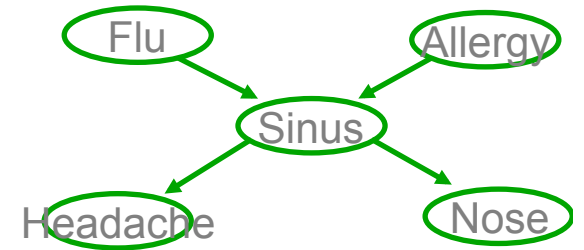
$$P(S_k = 1 | f_k a_k h_k n_k, \theta) = \frac{P(S_k = 1, f_k a_k h_k n_k | \theta)}{P(S_k = 1, f_k a_k h_k n_k | \theta) + P(S_k = 0, f_k a_k h_k n_k | \theta)}$$

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

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

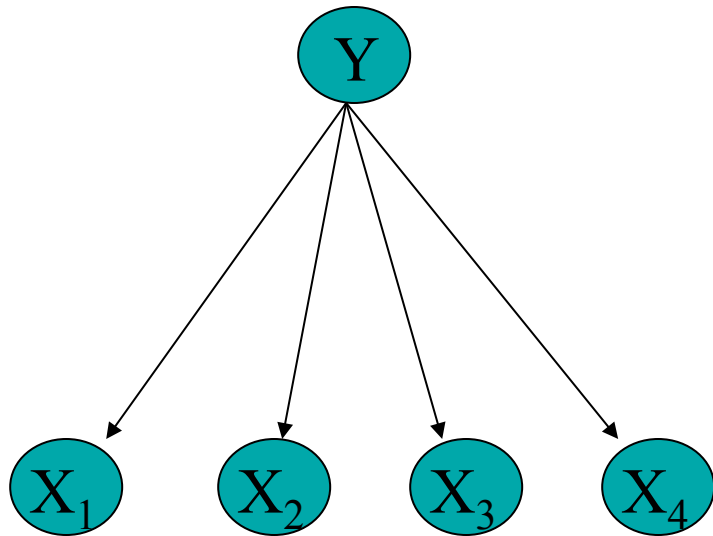
M step:

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

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

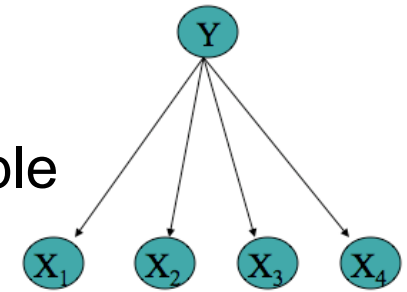
Using Unlabeled Data to Help Train Naïve Bayes Classifier

Learn $P(Y|X)$

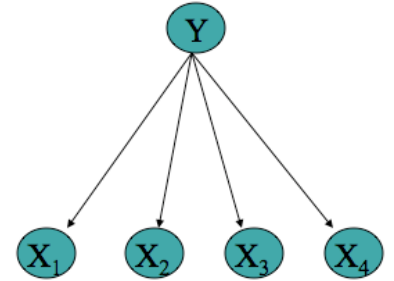


Y	X1	X2	X3	X4
1	0	0	1	1
0	0	1	0	0
0	0	0	1	0
?	0	1	1	0
?	0	1	0	1

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



EM and estimating θ



Given observed set X , unobserved set Y of boolean values

E step: Calculate for each training example, k

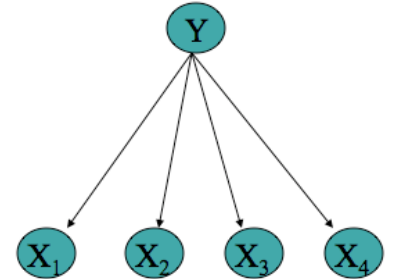
the expected value of each unobserved variable Y

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

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

let's use $y(k)$ to indicate value of Y on k th example

EM and estimating θ



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

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

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

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

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

-
- **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

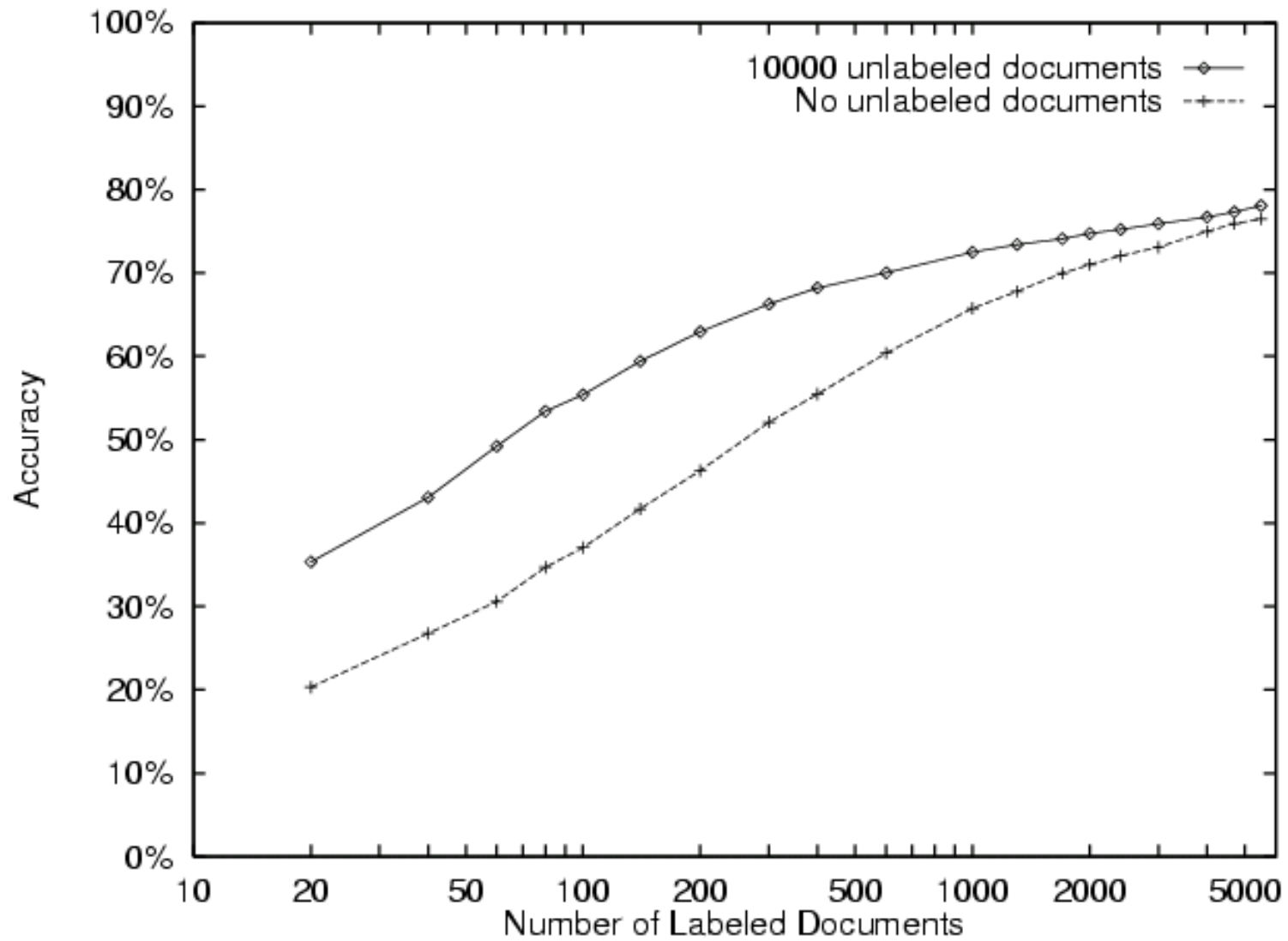
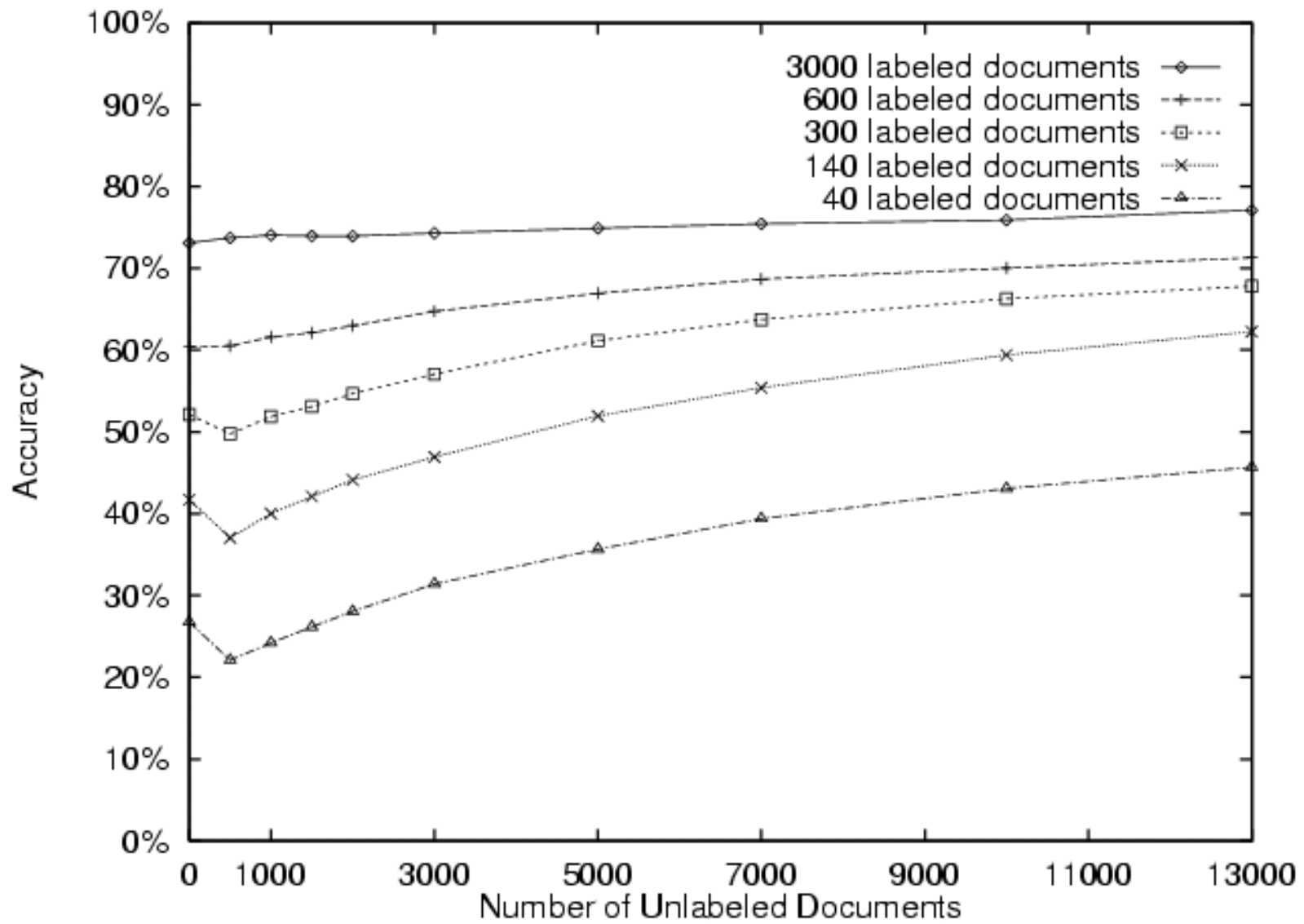


Table 3. Lists of the words most predictive of the course class in the WebKB data set, as they change over iterations of EM for a specific trial. By the second iteration of EM, many common course-related words appear. The symbol D indicates an arbitrary digit.

Iteration 0		Iteration 1	Iteration 2
intelligence	word w ranked by $P(w Y=\text{course}) /$ $P(w Y \neq \text{course})$	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

20 Newsgroups



Bayes Nets – What You Should Know

- Representation
 - Bayes nets represent joint distribution as a DAG + Conditional Distributions
 - D-separation lets us decode conditional independence assumptions
- Inference
 - NP-hard in general
 - For some graphs, some queries, exact inference is tractable
 - Approximate methods too, e.g., Monte Carlo methods, ...
- Learning
 - Easy for known graph, fully observed data (MLE's, MAP est.)
 - EM for partly observed data, known graph