

CS 331: Artificial Intelligence

Bayesian Networks (Inference)

Thanks to Andrew Moore for some of the slides.

Inference

- Suppose you are given a Bayesian network with the graph structure and the parameters all figured out
- Now you would like to use it to do inference
- You need inference to make predictions or classifications with a Bayes net

Another Example

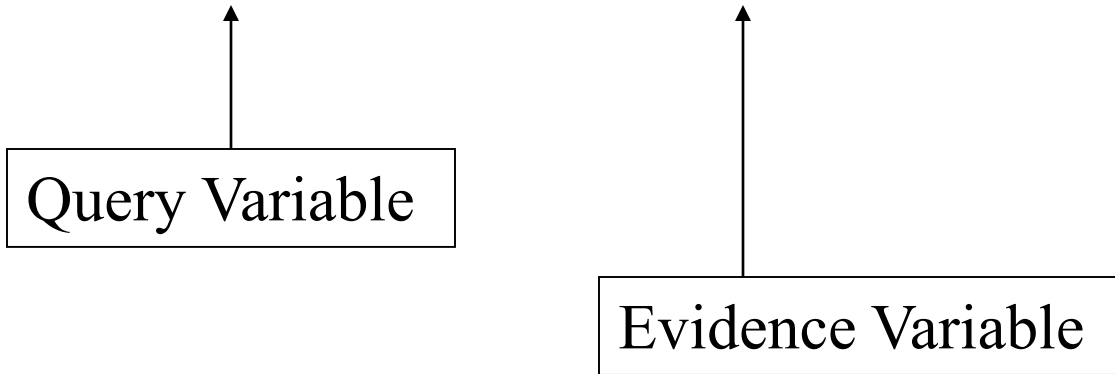
- You are very sick and you visit your doctor.
- The doctor is able to get the following information from you:
 - *HasFever = true*
 - *HasCough = true*
 - *HasBreathingProblems = true*
 - *AteBaconRecently = true*
- What's the probability you have *SwineFlu* given the above?

Another Example

- Need to compute
 $P(\text{SwineFlu} = \text{true} \mid \text{HasFever} = \text{true}, \text{HasCough} = \text{true}, \text{HasBreathingProblems} = \text{true}, \text{AteBaconRecently} = \text{true})$
- Suppose you pass out before you say a word to the doctor. The doctor is only able to determine you have a fever. What is $P(\text{SwineFlu} = \text{true} \mid \text{HasFever} = \text{true})$?

Query Example

$$P(\textit{SwineFlu} = \textit{true} \mid \textit{HasFever} = \textit{true})$$



Unobserved variables: *HasCough*, *HasBreathingProblems*,
AteBaconRecently

Queries Formalized

We will use the following notation:

- X = query variable
- $E = \{E_1, \dots, E_m\}$ is the set of evidence variables
- e = observed event
- $Y = \{Y_1, \dots, Y_l\}$ are the non-evidence (or hidden) variables
- The complete set of variables $\mathbf{X} = \{X\} \cup \mathbf{E} \cup \mathbf{Y}$

Need to calculate the query $P(X \mid e)$

Inference by Enumeration

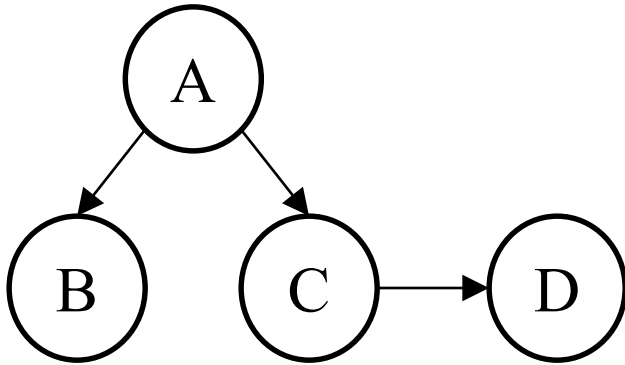
- Recall that:

$$P(X \mid e) = \alpha P(X, e) = \alpha \sum_y P(X, e, y)$$

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i \mid \text{parents}(X_i))$$

This means you can answer queries by computing sums of products of conditional probabilities from the network

Example #1

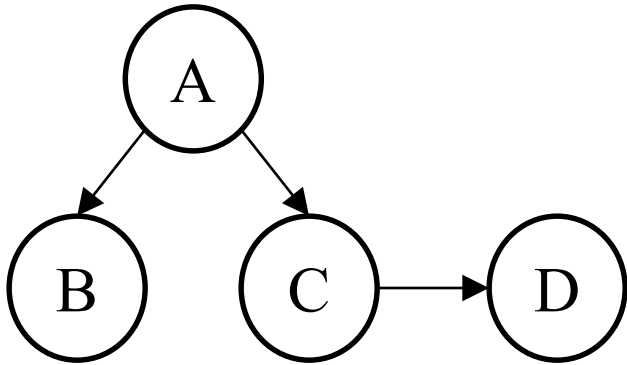


Query: $P(B=\text{true} \mid C=\text{true})$

How do you solve this? 2 steps:

1. Express it in terms of the joint probability distribution $P(A, B, C, D)$
2. Express the joint probability distribution in terms of the entries in the CPTs of the Bayes net

Example #1

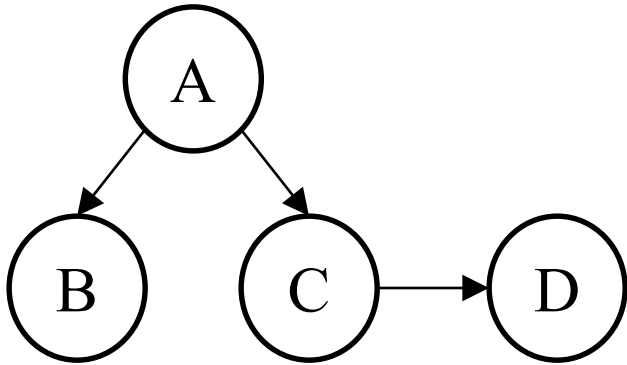


Whenever you see a conditional like $P(B = \text{true} \mid C = \text{true})$, use the Chain Rule:

$$P(B \mid C) = P(B, C) / P(C)$$

$$\begin{aligned} &P(B = \text{true} \mid C = \text{true}) \\ &= \frac{P(B = \text{true}, C = \text{true})}{P(C = \text{true})} \end{aligned}$$

Example #1



$$P(B = \text{true} \mid C = \text{true})$$

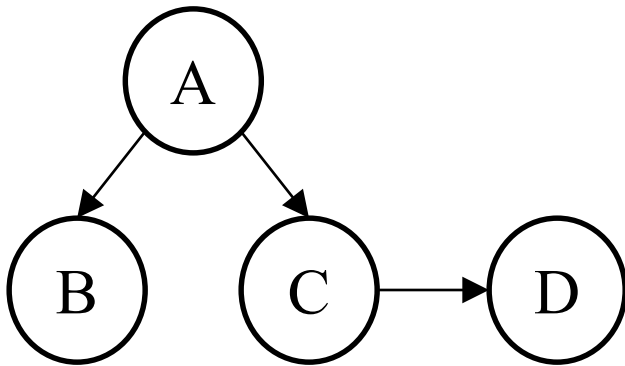
$$= \frac{P(B = \text{true}, C = \text{true})}{P(C = \text{true})}$$

$$= \frac{\sum_a \sum_d P(A = a, B = \text{true}, C = \text{true}, D = d)}{\sum_a \sum_b \sum_d P(A = a, B = b, C = \text{true}, D = d)}$$

Whenever you need to get a subset of the variables e.g. $P(B, C)$ from the full joint distribution $P(A, B, C, D)$, use marginalization:

$$P(X) = \sum_y P(X, Y = y)$$

Example #1

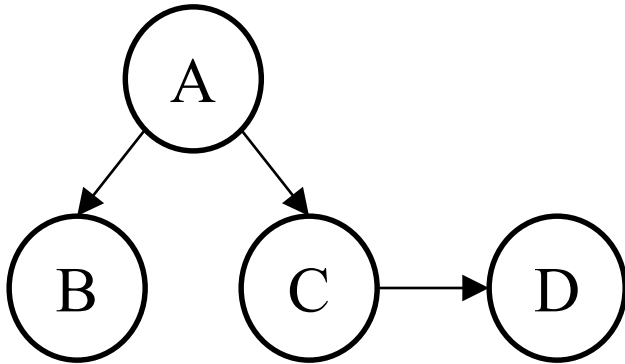


To express the joint probability distribution as the entries in the CPTs, use:

$$P(X_1, \dots, X_N) \\ = \prod_{i=1}^N P(X_i \mid \text{Parents}(X_i))$$

$$\begin{aligned} & \sum_a \sum_d P(A = a, B = \text{true}, C = \text{true}, D = d) \\ = & \frac{\sum_a \sum_b \sum_d P(A = a, B = b, C = \text{true}, D = d)}{\sum_a \sum_b \sum_d P(A = a)P(B = b \mid A = a)P(C = \text{true} \mid A = a)P(D = d \mid C = \text{true})} \\ = & \frac{\sum_a \sum_b \sum_d P(A = a)P(B = b \mid A = a)P(C = \text{true} \mid A = a)P(D = d \mid C = \text{true})}{\sum_a \sum_b \sum_d P(A = a)P(B = b \mid A = a)P(C = \text{true} \mid A = a)P(D = d \mid C = \text{true})} \end{aligned}$$

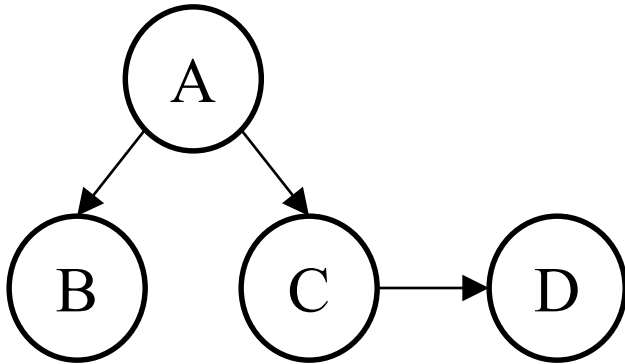
Example #1



Take the probabilities that don't depend on the terms in the summation and move them outside the summation

$$\begin{aligned} & \sum_a \sum_d P(A = a) P(B = \text{true} \mid A = a) P(C = \text{true} \mid A = a) P(D = d \mid C = \text{true}) \\ = & \frac{\sum_a \sum_b \sum_d P(A = a) P(B = b \mid A = a) P(C = \text{true} \mid A = a) P(D = d \mid C = \text{true})}{\sum_a \sum_b \sum_d 1} \\ = & \frac{\sum_a P(A = a) P(B = \text{true} \mid A = a) P(C = \text{true} \mid A = a) \sum_d P(D = d \mid C = \text{true})}{\sum_a \sum_b \sum_d 1} \\ = & \frac{\sum_a P(A = a) \sum_b P(B = b \mid A = a) P(C = \text{true} \mid A = a) \sum_d P(D = d \mid C = \text{true})}{\sum_a \sum_b \sum_d 1} \end{aligned}$$

Example #1



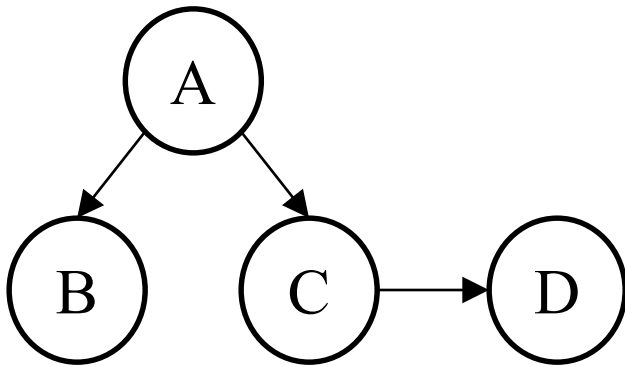
Take the probabilities that don't depend on the terms in the summation and move them outside the summation

Sums to 1

$$\begin{aligned}
 & \sum_a \sum_d P(A = a) P(B = \text{true} \mid A = a) P(C = \text{true} \mid A = a) P(D = d \mid C = \text{true}) \\
 = & \frac{\sum_a \sum_b \sum_d P(A = a) P(B = b \mid A = a) P(C = \text{true} \mid A = a) P(D = d \mid C = \text{true})}{\sum_a \sum_b \sum_d P(A = a) P(B = b \mid A = a) P(C = \text{true} \mid A = a) P(D = d \mid C = \text{true})} \\
 & \sum_a P(A = a) P(B = \text{true} \mid A = a) P(C = \text{true} \mid A = a) \sum_d P(D = d \mid C = \text{true}) \\
 = & \frac{\sum_a P(A = a) \sum_b P(B = b \mid A = a) P(C = \text{true} \mid A = a) \sum_d P(D = d \mid C = \text{true})}{\sum_a P(A = a) \sum_b P(B = b \mid A = a) P(C = \text{true} \mid A = a) \sum_d P(D = d \mid C = \text{true})}
 \end{aligned}$$

Sums to 1

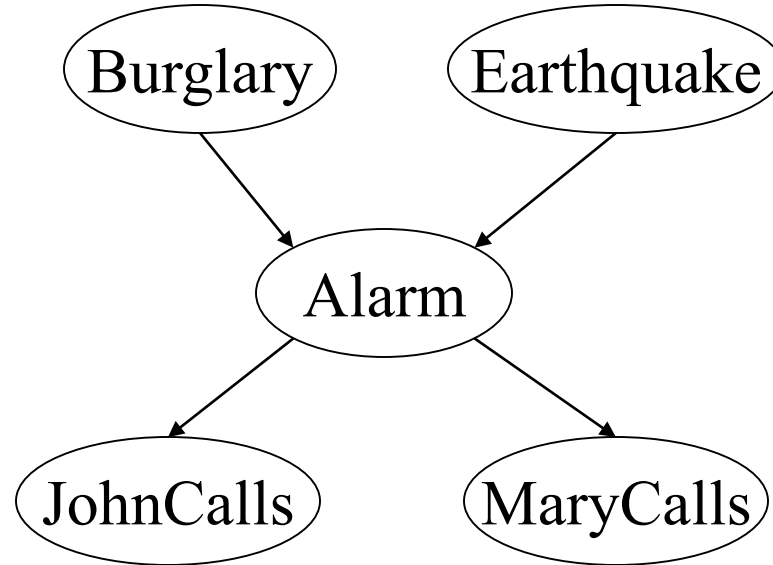
Example #1



Take the probabilities that don't depend on the terms in the summation and move them outside the summation

$$\begin{aligned}
 & \sum_a P(A = a) P(B = \text{true} \mid A = a) P(C = \text{true} \mid A = a) \\
 = & \frac{\sum_a P(A = a) \sum_b P(B = b \mid A = a) \underline{P(C = \text{true} \mid A = a)}}{\sum_a P(A = a) \sum_b P(B = b \mid A = a)} \quad \leftarrow \text{Doesn't depend on } b. \text{ Can move to the left} \\
 = & \frac{\sum_a P(A = a) P(B = \text{true} \mid A = a) P(C = \text{true} \mid A = a)}{\sum_a P(A = a) P(C = \text{true} \mid A = a) \underline{\sum_b P(B = b \mid A = a)}} \quad \leftarrow \text{Sums to 1} \\
 = & \frac{\sum_a P(A = a) P(B = \text{true} \mid A = a) P(C = \text{true} \mid A = a)}{\sum_a P(A = a) P(C = \text{true} \mid A = a)}
 \end{aligned}$$

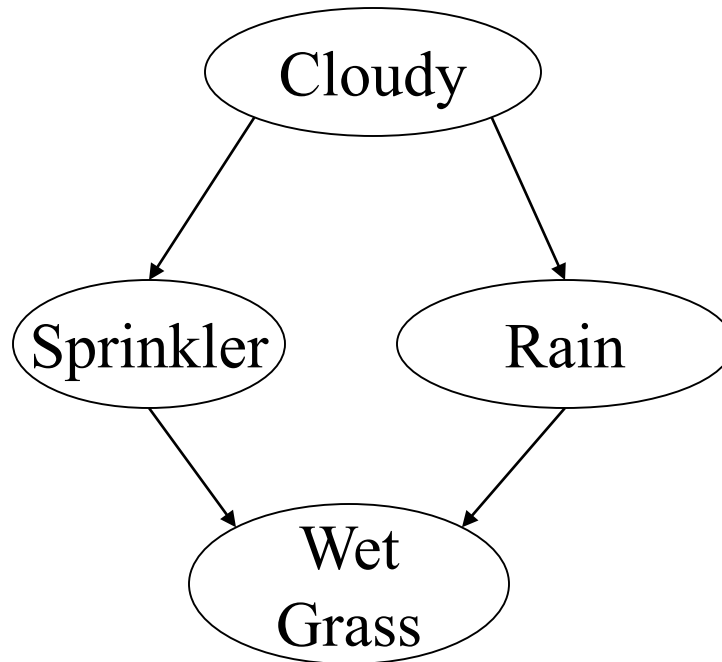
Complexity of Exact Inference



- The Burglary/Earthquake Bayesian network is an example of a polytree
- Singly connected networks (aka polytrees) have at most one undirected path between any two nodes in the network

Complexity of Exact Inference

- Polytrees have a nice property:
The time and space complexity of exact inference in polytrees is linear in the number of variables
- What about multiply connected networks?



Complexity of Exact Inference

- What about for multiply connected networks?
- Exponential time and space complexity in the number of variables in the worst case
- Bad news: Inference in Bayesian networks is NP-hard
- Even worse news: inference is #P-hard (strictly harder than NP-complete problems)

The Good News

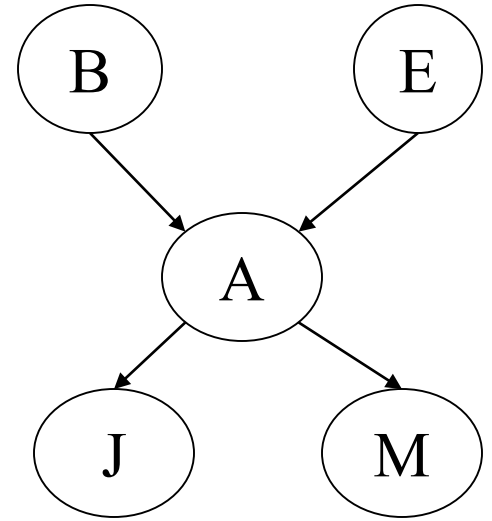
- Although **exact** inference is NP-hard, **approximate** inference is tractable
 - Lots of promising methods like sampling, MCMC, variational methods, etc.
- Approximate inference is a current research topic in Machine Learning

What You Should Know

- How to do exact inference in probabilistic queries of Bayes nets
- The complexity of inference for polytrees and multiply connected networks

Example #2

$$P(B = \text{true} \mid J = \text{true}, M = \text{true})$$



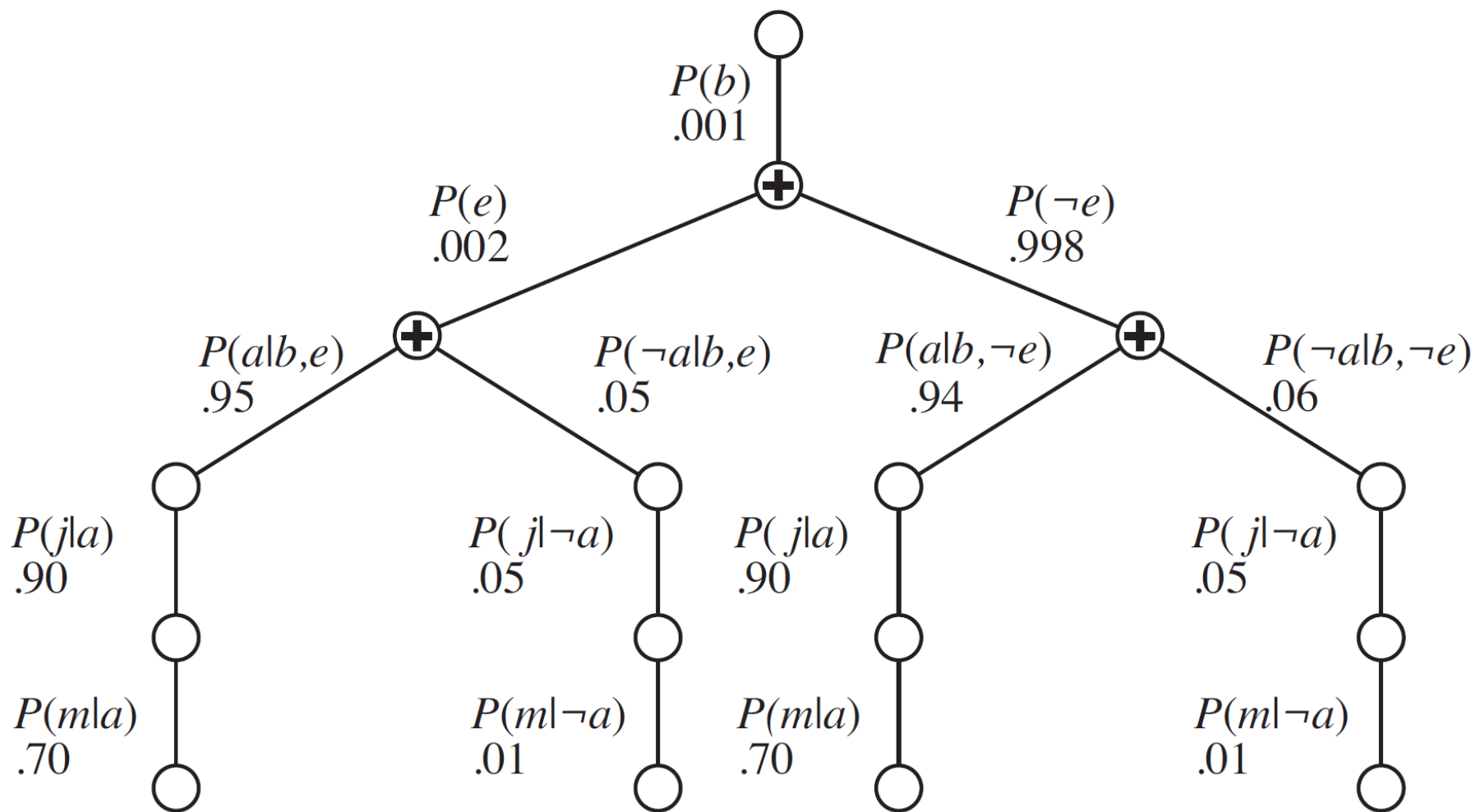
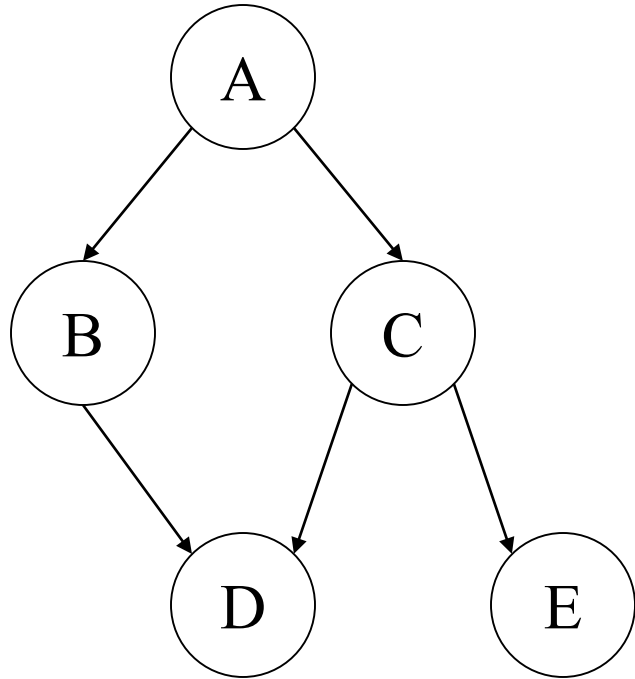


Figure 14.8 FILES: figures/enumeration-tree.eps (Tue Nov 3 16:22:41 2009). The structure of the expression shown in Equation (??). The evaluation proceeds top down, multiplying values along each path and summing at the “+” nodes. Notice the repetition of the paths for j and m .

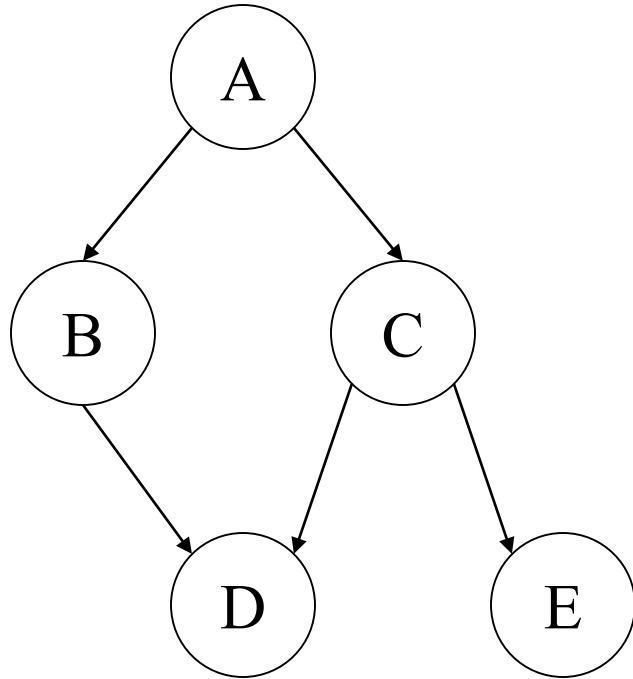
Practice



Write out the equations for the following probabilities using probabilities you can obtain from the Bayesian network. You will have to leave it in symbolic form because the CPTs are not shown, but simplify your answer as much as possible.

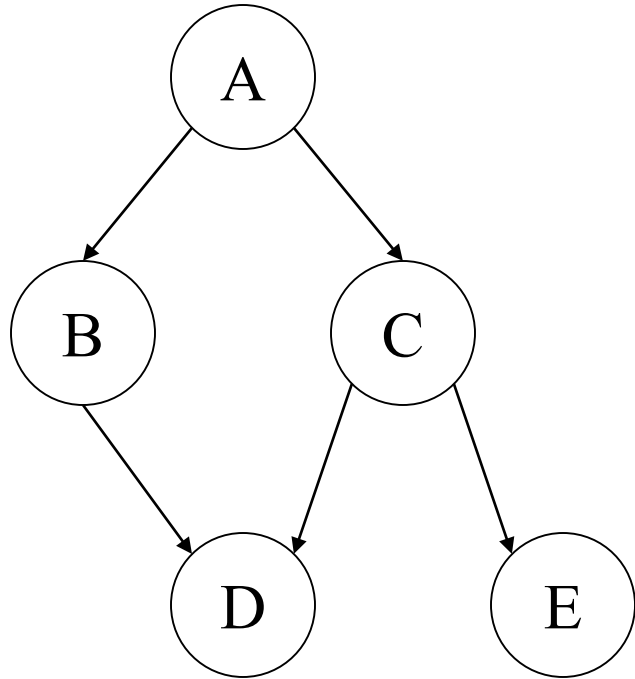
1. $P(A=\text{true}, B=\text{true}, C=\text{true}, D=\text{true}, E=\text{true})$

CW: Practice



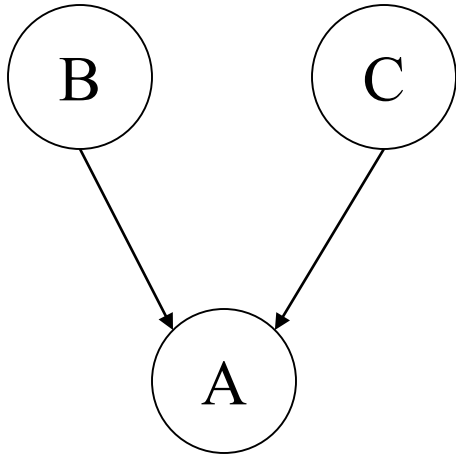
2. $P(B=\text{true} \mid D=\text{true})$

CW: Practice



3. $P(A=\text{true}, D=\text{true}, E=\text{true} \mid B=\text{true}, C=\text{true})$

CW: Practice



B	P(B)
false	0.25
true	0.75

C	P(C)
false	0.1
true	0.9

C	B	A	P(A B,C)
false	false	false	0.1
false	false	true	0.9
false	true	false	0.2
false	true	true	0.8
true	false	false	0.3
true	false	true	0.7
true	true	false	0.4
true	true	true	0.6

4. What is $P(B=\text{false}, C=\text{false})$?

5. Can you come up with another Bayes net structure (using only the 3 nodes above) that represents the same joint probability distribution?

Naïve Bayes

- A special type of Bayesian network
- Makes a conditional independence assumption
- Typically used for classification

Classification

Suppose you are trying to classify situations that determine whether or not Canvas will be down. You've come up with the following list of variables (which are all Boolean):

Monday	Is a Monday
Assn	CS331 assignment due
Grades	CS331 instructor needs to enter grades
Win	The Beavers won the football game

We also have a Boolean variable called CD which stands for “Canvas down”

Classification

These are called
features or attributes

This is called the “class” variable
(because we’re trying to classify it)

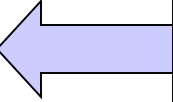
Monday	Assn	Grades	Win	CD
true	true	true	false	true
false	true	true	true	false
true	false	false	false	false
false	true	true	false	true
true	true	true	false	true
false	false	true	false	true
true	true	false	true	false

These entries in the
CD column are
called “class labels”

Classification

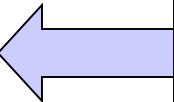
Monday	Assn	Grades	Win	CD
true	true	true	false	true
false	true	true	true	false
true	false	false	false	false
false	true	false	false	true
true	true	true	false	true
false	false	true	false	true
true	true	false	true	false

You create a dataset out of your past experience. This is called “training data”.

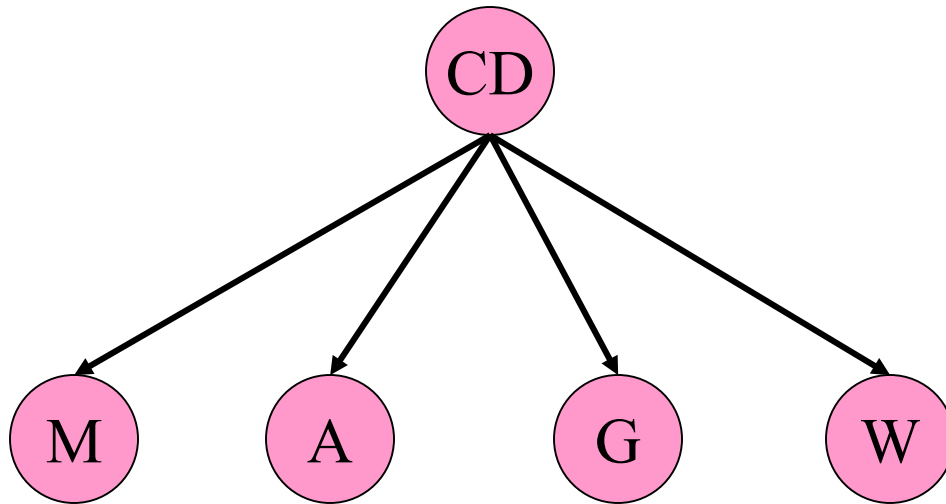


Monday	Assn	Grades	Win
true	true	true	true
false	true	true	false

You now have 2 new situations and you would like to predict if Canvas will go down. This is called “test data”.



Naïve Bayes Structure

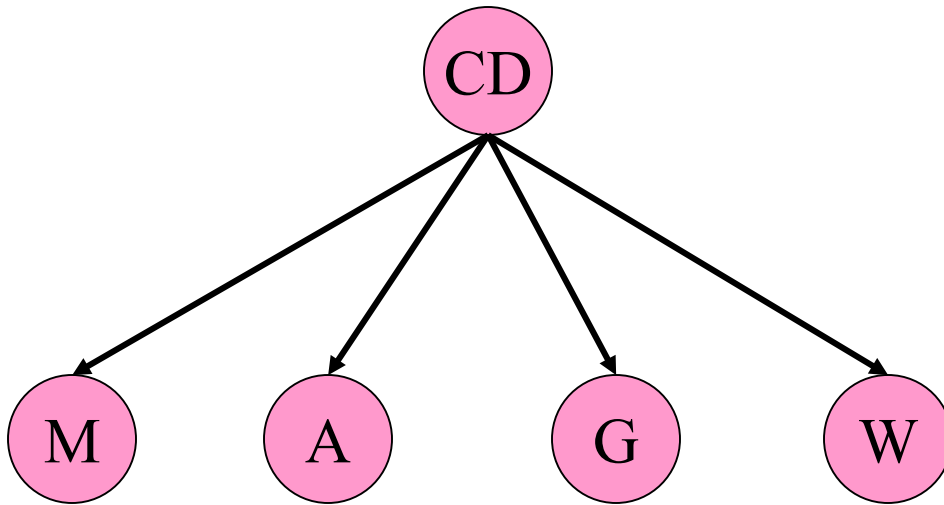


Notice the conditional independence assumption:

The features are conditionally independent given the class variable.

Naïve Bayes Parameters

$$P(CD) = ?$$



$$P(M | CD) = ?$$

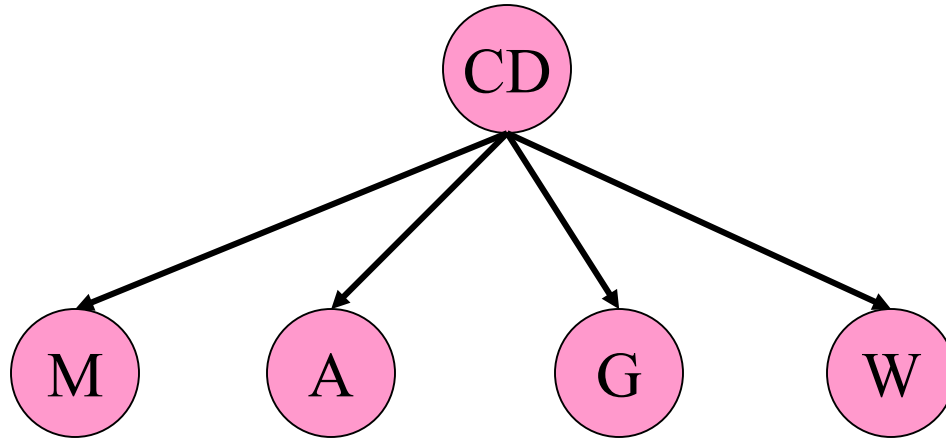
$$P(A | CD) = ?$$

$$P(G | CD) = ?$$

$$P(W | CD) = ?$$

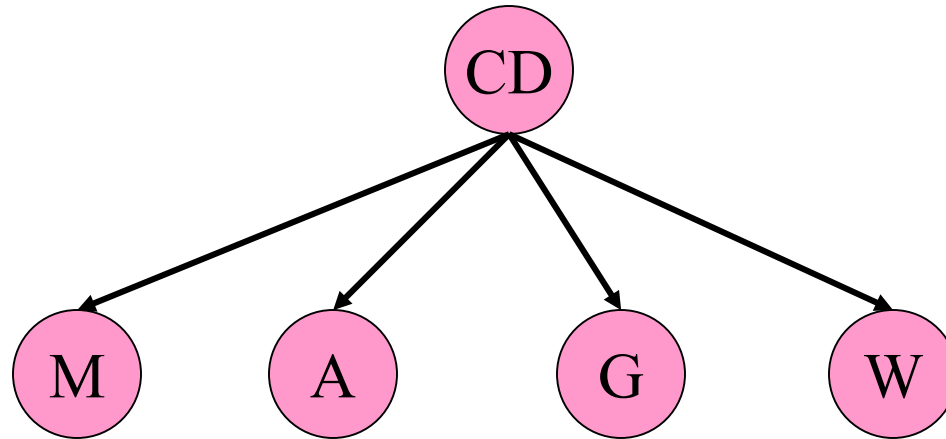
How do you get these parameters from the training data?

Naïve Bayes Parameters



CD	$P(CD)$
false	$(\# \text{ of records in training data with } CD = \text{false}) / (\# \text{ of records in training data})$
true	$(\# \text{ of records in training data with } CD = \text{true}) / (\# \text{ of records in training data})$

Naïve Bayes Parameters



M	CD	$P(M CD)$
false	false	$(\# \text{ of records with } M = \text{false and } CD = \text{false}) / (\# \text{ of records with } CD = \text{false})$
false	true	$(\# \text{ of records with } M = \text{false and } CD = \text{true}) / (\# \text{ of records with } CD = \text{true})$
true	false	$(\# \text{ of records with } M = \text{true and } CD = \text{false}) / (\# \text{ of records with } CD = \text{false})$
true	true	$(\# \text{ of records with } M = \text{true and } CD = \text{true}) / (\# \text{ of records with } CD = \text{true})$

Inference in Naïve Bayes

$$P(CD | M, A, G, W)$$

$$= \frac{P(M, A, G, W | CD)P(CD)}{P(M, A, G, W)}$$

By Bayes Rule

$$= \alpha P(M, A, G, W | CD)P(CD)$$

Treat denominator
as constant

$$= \alpha P(CD)P(M | CD)P(A | CD)P(G | CD)P(W | CD)$$

From conditional
independence

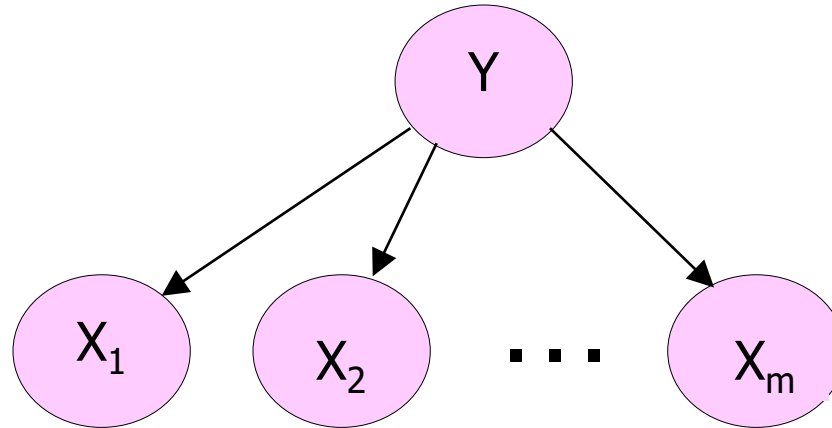
Prediction

- Suppose you are now in a day when $M=\text{true}$, $A=\text{true}$, $G=\text{true}$, $W=\text{true}$.
- You need to predict if $CD=\text{true}$ or $CD=\text{false}$.
- We will use the notation that $CD=\text{true}$ is equivalent to cd and $CD=\text{false}$ is equivalent to $\neg cd$.

Prediction

- You need to compare:
 - $P(\text{cd} \mid m, a, g, w) = \alpha P(\text{cd}) P(m \mid \text{cd}) P(a \mid \text{cd}) P(g \mid \text{cd}) P(w \mid \text{cd})$
 - $P(\neg \text{cd} \mid m, a, g, w) = \alpha P(\neg \text{cd}) P(m \mid \neg \text{cd}) P(a \mid \neg \text{cd}) P(g \mid \neg \text{cd}) P(w \mid \neg \text{cd})$
- Whichever probability is the bigger of the two above, that is your prediction for CD
- Because you take the max of the two probabilities above, you can ignore α (since it is the same in both)

The General Case



1. Estimate $P(Y=v)$ as fraction of records with $Y=v$
2. Estimate $P(X_i=u \mid Y=v)$ as fraction of “ $Y=v$ ” records that also have $X=u$.
3. To predict the Y value given observations of all the X_i values, compute

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

Naïve Bayes Classifier

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

Naïve Bayes Classifier

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \frac{P(Y = v, X_1 = u_1 \cdots X_m = u_m)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

Naïve Bayes Classifier

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \frac{P(Y = v, X_1 = u_1 \cdots X_m = u_m)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \frac{P(X_1 = u_1 \cdots X_m = u_m \mid Y = v)P(Y = v)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

Naïve Bayes Classifier

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \frac{P(Y = v, X_1 = u_1 \cdots X_m = u_m)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \frac{P(X_1 = u_1 \cdots X_m = u_m \mid Y = v)P(Y = v)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(X_1 = u_1 \cdots X_m = u_m \mid Y = v)P(Y = v)$$

Naïve Bayes Classifier

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \frac{P(Y = v, X_1 = u_1 \cdots X_m = u_m)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \frac{P(X_1 = u_1 \cdots X_m = u_m \mid Y = v)P(Y = v)}{P(X_1 = u_1 \cdots X_m = u_m)}$$

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(X_1 = u_1 \cdots X_m = u_m \mid Y = v)P(Y = v)$$

Because of the structure of
the Bayes Net

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(Y = v) \prod_{j=1}^m P(X_j = u_j \mid Y = v)$$

Technical Point #1

- The probabilities $P(X_j = u_j \mid Y = v)$ can sometimes be really small
- This can result in numerical instability since floating point numbers are not represented exactly on any computer architecture
- To get around this, use the log of the last line in the previous slide i.e.

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \left[\log(P(Y = v)) + \sum_{j=1}^m \log(P(X_j = u_j \mid Y = v)) \right]$$

Technical Point #2

- When estimating parameters, what happens if you don't have any records that match a certain combination of features?
- For example, in our training data, we didn't have M=false, A=false, G=false, W=false
- This means that $P(X_j = u_j \mid Y = v)$ in the formula below will be 0 and the entire expression will be 0.

$$P(Y = v) \prod_{j=1}^m P(X_j = u_j \mid Y = v)$$

Even more horrible things happen if you had this expression in log space

Uniform Dirichlet Priors

Let N_j be the number of values that X_j can take on.

$$P(X_j = u_j \mid Y = v) = \frac{(\text{\#records with } X_j = u_j \text{ and } Y = v) + 1}{(\text{\#records with } Y = v) + N_j}$$

What happens when you have no records with $Y = v$?

$$P(X_j = u_j \mid Y = v) = \frac{1}{N_j}$$

This means that each value of X_j is equally likely in the absence of data. If you have a lot of data, it dominates the $1/N_j$ value. We call this trick a “uniform Dirichlet prior”.

Technical Point #3

- You often want to learn incrementally and predict as needed
- Instead of $P(Y = v)$, $P(X_j = U_j | Y = v)$
- Maintain counts $N(Y = v)$, $N(Y = v, X_j = U_j)$
- Convert them to probabilities when you need them for prediction
- Dirichlet priors: To avoid division by zero, initialize
 - $N(Y = v, X_j = U_j) = 1$
 - $N(Y = v) = \sum_j N(Y = v, X_j = U_j) = N_j$

Programming Assignment #3

You will classify text into two classes.

There are two files:

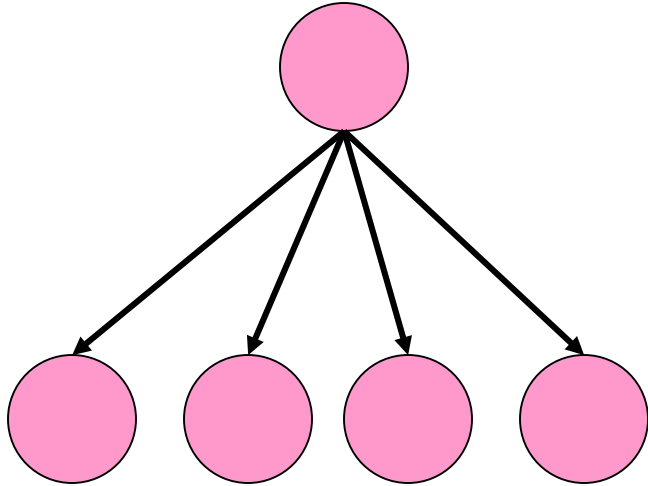
1. Training data: trainingSet.txt
2. Testing data: testSet.txt

Programming Assignment #3

Two parts to this assignment:

1. Pre-processing step
2. Classification step

1. Preprocessing Step



- Recall that naïve Bayes has the structure shown to the right
- The nodes correspond to random variables, which are the features or attributes in the data
- What are the features in the documents?
- **Note: a “document” in our assignment is a Yelp review to be classified as positive or negative**

The Vocabulary

- The features of the documents will be the presence/absence of words in the vocabulary
- The **vocabulary** is the list of words that are known to the classifier
- Ideally, the vocabulary would be all the words in the English language
- For this assignment, you will form the vocabulary using all the words in the training data

Bag of Words

Suppose you have the following documents:

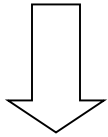
<u>Training Data</u>	<u>Class Label</u>
This is an excellent laptop	Class 1
No, this is not sarcasm!	Class 0
<u>Test Data</u>	
Excellent Laptop =P	Class 1

You will ignore
punctuation for this
assignment

The vocabulary will be:
this, is, an, excellent, laptop, no, not, sarcasm

Bag of Words

Vocab: this, is, an, excellent, laptop, no, not, sarcasm



Keep this in alphabetical order to help with debugging

Vocab: an, excellent, is, laptop, no, not, sarcasm, this

Training data

Next, convert your training and test data into features

Training Data

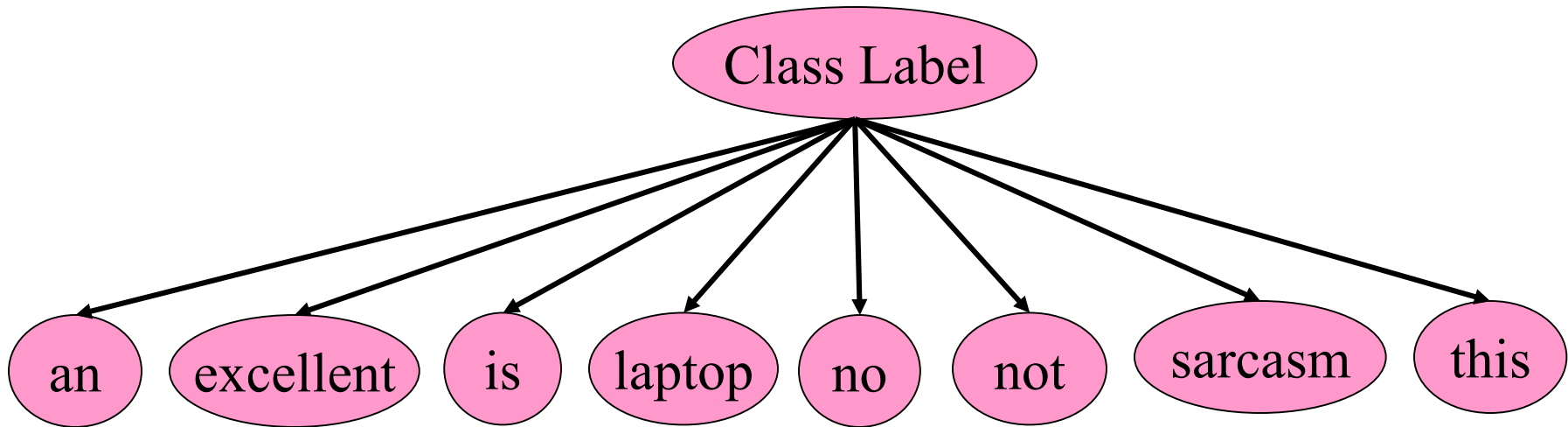
an	excellent	is	laptop	no	not	sarcasm	this	Class Label
1	1	1	1	0	0	0	1	1
0	0	1	0	1	1	1	1	0

Test Data

an	excellent	is	laptop	no	not	sarcasm	this	Class Label
0	1	0	1	0	0	0	0	1

You will output the training data in feature form, with the features alphabetized (we will grade you on this output).

2. Classification Step (Training Phase)



- Your naïve Bayes classifier now looks something like the above
- You still need to fill in the conditional probability tables in each node
- This is done in the **training phase** (as described on slides 9 and 10)
- Remember to use the uniform Dirichlet prior trick (see slide 21)
- Do incremental training on 4 parts of the training set using indices of records

2. Classification Step (Testing Phase)

Testing phase

- Load the featurized test data
- For each document in the test data, predict its class label
- This requires computing:
 $P(\text{Class label} \mid \text{Words in document})$

2. Classification Step (Testing Phase)

Suppose you have the following test instance:

an	excellent	is	laptop	no	not	sarcasm	this	Class Label
0	1	0	1	0	0	0	0	(to be predicted)

$$\begin{aligned} &P(\text{Class} = 1 \mid \text{an} = 0, \text{excellent} = 1, \text{is} = 0, \text{laptop} = 1, \text{no} = 0, \text{not} \\ &= 0, \text{sarcasm} = 0, \text{this} = 0) \\ &= \alpha P(\text{Class} = 1) * P(\text{an} = 0 \mid \text{Class} = 1) * P(\text{excellent} = 1 \mid \text{Class} = 1) * \\ &\quad P(\text{is} = 0 \mid \text{Class} = 1) * P(\text{laptop} = 1 \mid \text{Class} = 1) * P(\text{no} = 0 \mid \text{Class} = 1) * \\ &\quad P(\text{not} = 0 \mid \text{Class} = 1) * P(\text{sarcasm} = 0 \mid \text{Class} = 1) * \\ &\quad P(\text{this} = 0 \mid \text{Class} = 1) \end{aligned}$$

Note: Use $P(\text{Word} = 1 \mid \text{Class})$ if you have a 1 for the word. Otherwise use $P(\text{Word} = 0 \mid \text{Class})$

2. Classification Step (Testing Phase)

an	excellent	is	laptop	no	not	sarcasm	this	Class Label
0	1	0	1	0	0	0	0	(to be predicted)

Then compute the following:

$$\begin{aligned} &P(\text{Class} = 0 \mid \text{an} = 0, \text{excellent} = 1, \text{is} = 0, \text{laptop} = 1, \text{no} = 0, \text{not} \\ &= 0, \text{sarcasm} = 0, \text{this} = 0) \\ &= \alpha P(\text{Class} = 0) * P(\text{an} = 0 \mid \text{Class} = 0) * P(\text{excellent} = 1 \mid \text{Class} = 0) * \\ &\quad P(\text{is} = 0 \mid \text{Class} = 0) * P(\text{laptop} = 1 \mid \text{Class} = 0) * P(\text{no} = 0 \mid \text{Class} = 0) * \\ &\quad P(\text{not} = 0 \mid \text{Class} = 0) * P(\text{sarcasm} = 0 \mid \text{Class} = 0) * \\ &\quad P(\text{this} = 0 \mid \text{Class} = 0) \end{aligned}$$

2. Classification Step (Testing Phase)

an	excellent	is	laptop	no	not	sarcasm	this	Class Label
0	1	0	1	0	0	0	0	(to be predicted)

If

$$\alpha P(\text{Class} = 1 \mid \text{an} = 0, \text{excellent} = 1, \text{is} = 0, \text{laptop} = 1, \text{no} = 0, \text{not} = 0, \text{sarcasm} = 0, \text{this} = 0)$$

>

$$\alpha P(\text{Class} = 0 \mid \text{an} = 0, \text{excellent} = 1, \text{is} = 0, \text{laptop} = 1, \text{no} = 0, \text{not} = 0, \text{sarcasm} = 0, \text{this} = 0)$$

Predict **Class = 1** otherwise predict Class = 0

2. Classification Step (Testing Phase)

- For each document in the testing data set, predict its class label
- Compare the predicted class label to the actual class label
- Output the accuracy for each class:

$$\frac{\text{\# correctly predicted class labels}}{\text{total \# of predictions}}$$

Results

There are two sets of results we require:

1. Results #1:

- Use trainingSet.txt for the training phase
- Use trainingSet.txt for the testing phase
- Report accuracies after incrementally training on 4 parts

2. Results #2:

- Use trainingSet.txt for the training phase
- Use testSet.txt for the testing phase
- Report accuracies after incrementally training on 4 parts

What You Should Know

- How to learn the parameters for a Naïve Bayes model
- How to make predictions with a Naïve Bayes model
- How to implement a Naïve Bayes Model

Example

Monday	Assn	Grades	Win	CD
true	true	true	false	true
false	true	true	true	false
true	false	false	false	false
false	true	false	false	true
true	true	true	false	true
false	false	true	false	true
true	true	false	true	false

Compute $P(M|CD)$ using uniform Dirichlet priors

Monday	Assn	Grades	Win	CD
true	true	true	false	true
false	true	true	true	false
true	false	false	false	false
false	true	false	false	true
true	true	true	false	true
false	false	true	false	true
true	true	false	true	false

CW: Practice

Monday	Assn	Grades	Win	CD
true	true	true	false	true
false	true	true	true	false
true	false	false	false	false
false	true	false	false	true
true	true	true	false	true
false	false	true	false	true
true	true	false	true	false

Compute $P(W=\text{true}|CD=\text{true})$ using uniform Dirichlet priors