Bayes Nets

CS 480

Intro to Artificial Intelligence

Representing independence

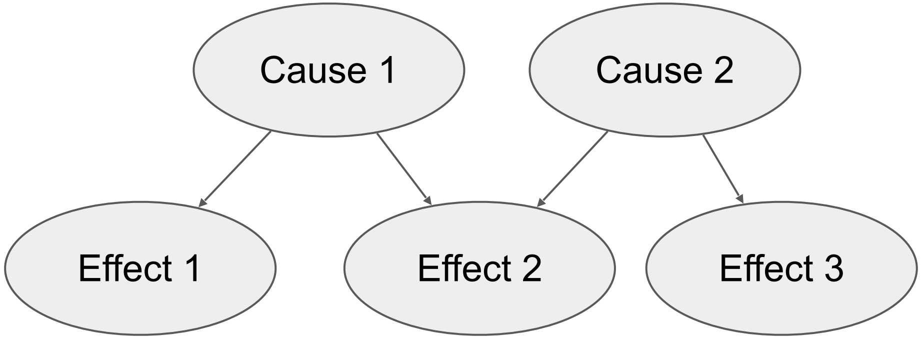
It's difficult to tell what the independence relationships are just by looking at the joint probability distribution

Intuitively, independence is related to **cause** and **effect** relationships: the *reason* you have a toothache, or the dentist tool catches is because you have a cavity

One natural way to represent these relationships is with a **directed acyclic graph**

	Tooth		¬Tooth	
	Cat	¬Cat	Cat	¬Cat
Cav	0.108	0.012	0.072	0.008
¬Cav	0.016	0.064	0.144	0.576

$$Toothache \perp Catch \mid Cavity$$



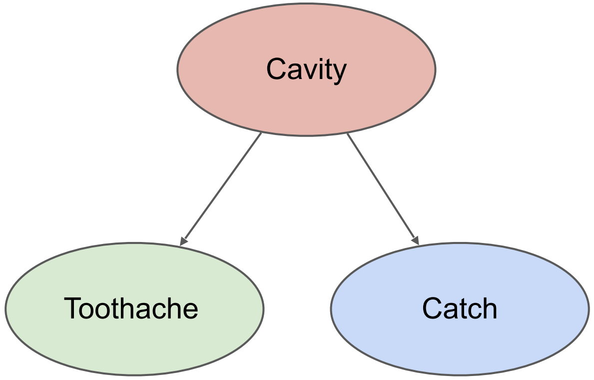
Bayes Nets (1)

IF the graph accurately represents the independence structure, nodes are **independent** of their siblings given their immediate parents: $$Toothache \perp Catch \mid Cavity$$

Nodes: random variables

Directed edges: from cause to effect

Conditional Probability Table (CPT)
For discrete RV's we can represent the conditional probability as a table



p(Cav)	p(¬Cav)
0.2	0.8

	p(Cat|Cav)	p(¬Cat|Cav)
Cav=T	0.9	0.1
Cav=F	0.6	0.4

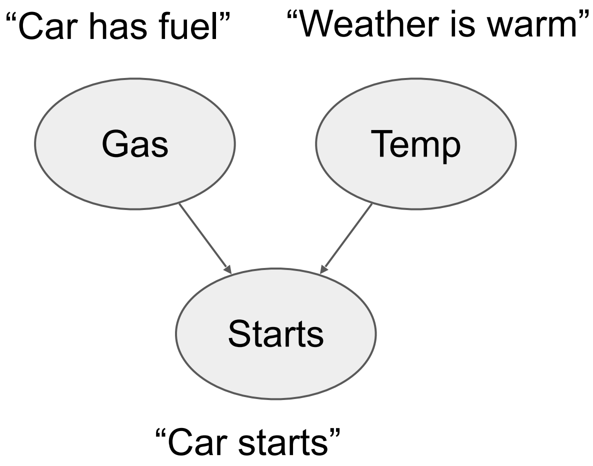
	p(Tth|Cav)	p(¬Tth|Cav)
Cav=T	0.6	0.4
Cav=F	0.1	0.9

Bayes Nets (2)

IF the graph accurately represents the independence structure, parents are **independent** if **not** conditioned on common children $Gas \perp Temp$

Nodes can have more than one parent, or none

- CPT includes all parents
- Nodes without parents: marginals



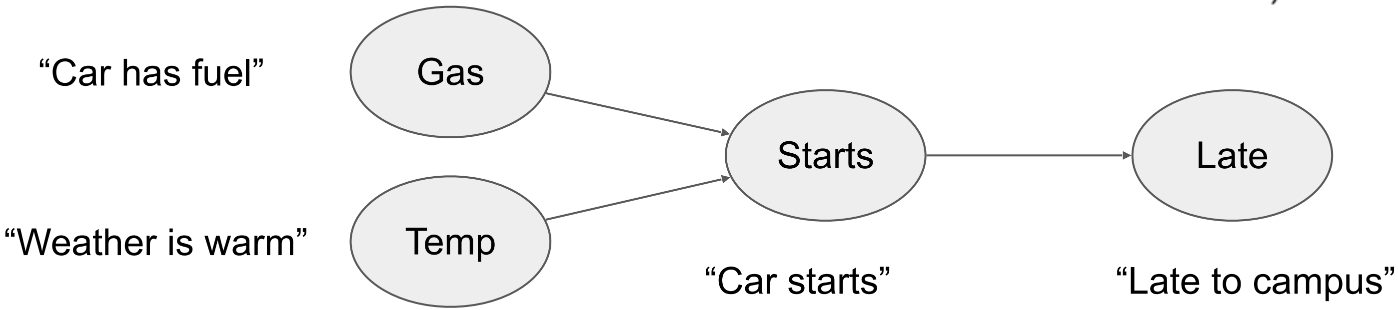
	p(Starts|Gas,Temp)	p(¬Starts|Gas,Temp)
Gas=T, Temp=T	0.9	0.1
Gas=T, Temp=F	0.8	0.2
Gas=F, Temp=T	0.5	0.5
Gas=F,Temp=F	0.3	0.7

BUT if conditioned on a common child, parents are **no longer** independent (knowing effect influences the probability of both causes) $Gas \not\perp Temp \mid Starts$

Bayes Nets (3)

Nodes can have “grandparents” (chains)

$$Late \perp Gas, Temp \mid Starts$$



IF the graph accurately represents the independence structure, nodes are **independent** of their grandparents given their immediate parents

Independence in Bayes Nets

IF the graph accurately represents the independence structure, we can **factor** the joint probability into a convenient form

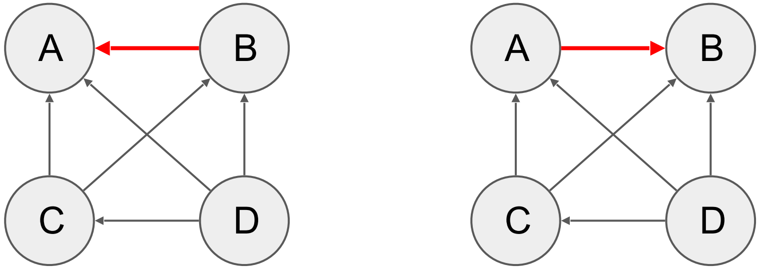
$$p(X_1, X_2, \ldots, X_D) = \prod_{i=1}^{D} p(X_i \mid \textsc{Parents}(X_i))$$

In words: nodes are **conditionally independent** of their **ancestors** and **siblings** (non-descendents) given their **parents**

Cause and Effect

If you know the cause and effect relationships for your problem, it's easy to build a Bayes Net. Can you also **infer** cause and effect relationship **from** a graph? No!

$$
\begin{aligned}
p(A, B, C, D) &= p(A \mid B, C, D) p(B \mid C, D) p(C \mid D) p(D) \\
&= p(B \mid A, C, D) p(A \mid C, D) p(C \mid D) p(D)
\end{aligned}
$$



This also shows that a graph for a given set of RVs is **not necessarily unique**

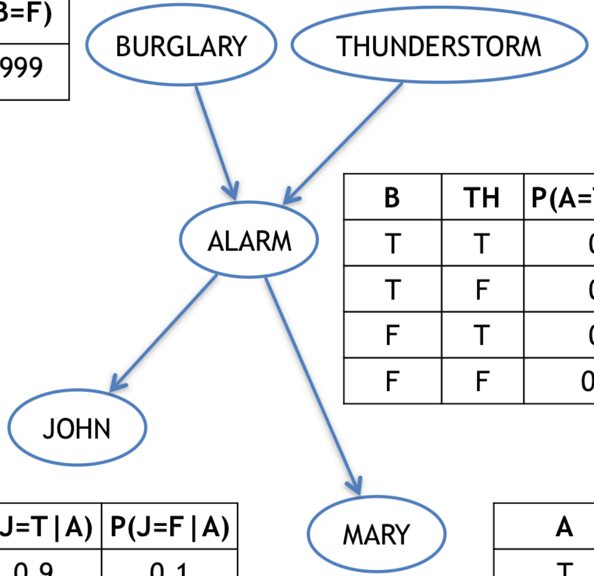
Using Bayes Nets - Example (1)

Setup

You are on vacation, and you've asked your neighbors to keep an eye on your house while you are away. They'll call you if your house alarm goes off.

Your alarm system could be triggered because of an actual burglar, or because a thunderstorm sets it off.

P(B=T)	P(B=F)
0.001	0.999



P(TH=T)	P(TH=F)
0.002	0.998

B	TH	P(A=T|B,TH)	P(A=F|B,TH)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(J=T|A)	P(J=F|A)
T	0.9	0.1
F	0.05	0.95

A	P(M=T|A)	P(M=F|A)
T	0.7	0.3
F	0.01	0.99

Using Bayes Nets - Example (2)

$$p(X_1, X_2, \ldots, X_D) = \prod_{i=1}^{D} p(X_i \mid \textsc{Parents}(X_i))$$

What's the probability that

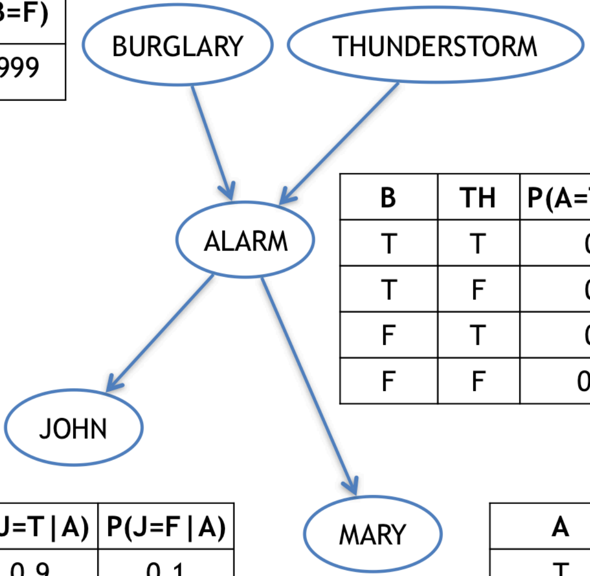
- Both neighbors call
- The alarm goes off
- There is no burglar
- There is no storm

p(j,m,a,¬b,¬t) =
p(j|a) p(m|a) p(a|¬b,¬t) p(¬b) p(¬t) =
(.9) (.7) (.001) (.999) (.998) = 0.00062

Joint probability table: 2^5=32 cells

CPT factorization: 20 cells

P(B=T)	P(B=F)
0.001	0.999



P(TH=T)	P(TH=F)
0.002	0.998

B	TH	P(A=T|B,TH)	P(A=F|B,TH)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(J=T|A)	P(J=F|A)
T	0.9	0.1
F	0.05	0.95

A	P(M=T|A)	P(M=F|A)
T	0.7	0.3
F	0.01	0.99

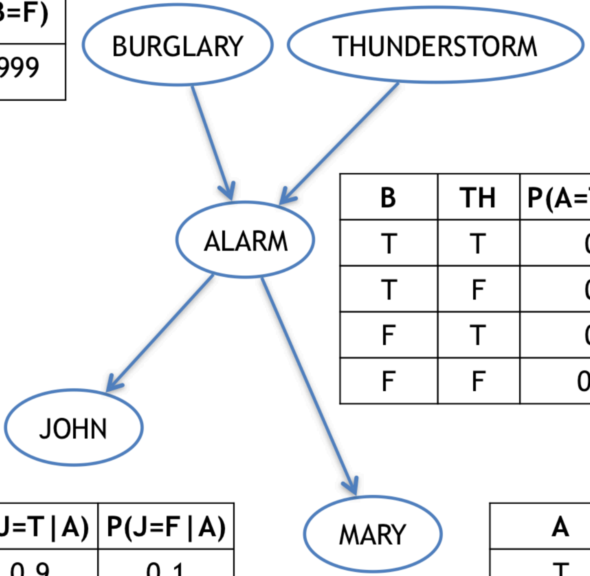
Using Bayes Nets - Example (3)

What's the probability that there is a burglar if both John and Mary call?

In general, there's a 4 step process to solve **any** query about a Bayes Net:

1. Write the query as a statement about probabilities
2. Rewrite statement in terms of the joint probability distribution
3. Factor the joint probability using Bayes Net independencies
4. Simplify, and plug in numbers from CPTs

P(B=T)	P(B=F)
0.001	0.999



P(TH=T)	P(TH=F)
0.002	0.998

B	TH	P(A=T|B,TH)	P(A=F|B,TH)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(J=T|A)	P(J=F|A)
T	0.9	0.1
F	0.05	0.95

A	P(M=T|A)	P(M=F|A)
T	0.7	0.3
F	0.01	0.99

Using Bayes Nets - Example (4)

1. Write the query as a statement about probabilities

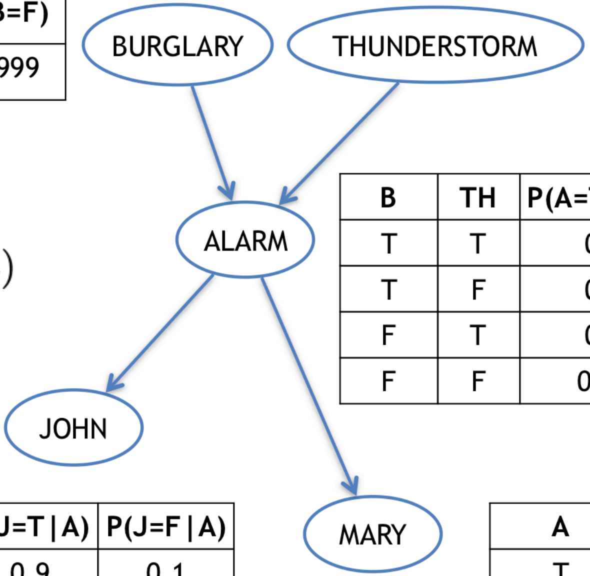
"What's the probability that there is a burglar if both John and Mary call?"

$$p(Burglar = \textsc{True} \mid John = \textsc{True}, Mary = \textsc{True})$$

Abbreviated

$$p(b \mid j, m)$$

P(B=T)	P(B=F)
0.001	0.999



P(TH=T)	P(TH=F)
0.002	0.998

B	TH	P(A=T|B,TH)	P(A=F|B,TH)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(J=T|A)	P(J=F|A)
T	0.9	0.1
F	0.05	0.95

A	P(M=T|A)	P(M=F|A)
T	0.7	0.3
F	0.01	0.99

Using Bayes Nets - Example (5)

2. Rewrite in joint probability form

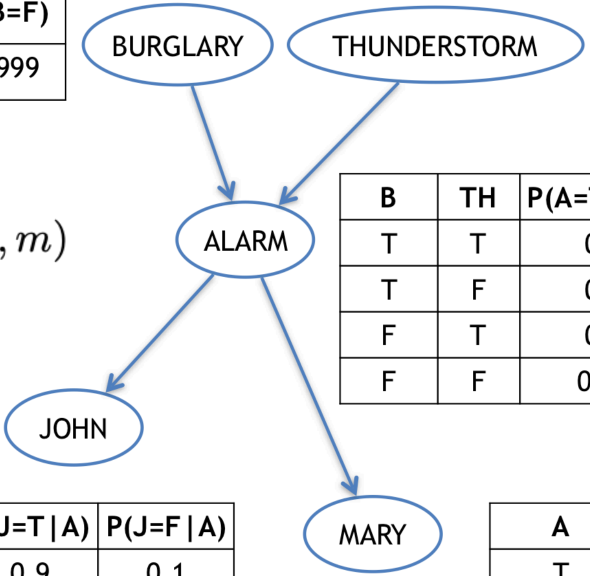
$$p(b \mid j, m) = \frac{p(b, j, m)}{p(j, m)} = \alpha \cdot p(b, j, m)$$

$$= \alpha \sum_{h_1} \sum_{h_2} p(b, Th = h_1, Al = h_2, j, m)$$

Using

- definition of conditional probability and normalization
- marginalization over Thunderstorm and Alarm

P(B=T)	P(B=F)
0.001	0.999



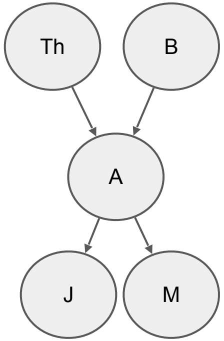
P(TH=T)	P(TH=F)
0.002	0.998

B	TH	P(A=T|B,TH)	P(A=F|B,TH)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(J=T|A)	P(J=F|A)
T	0.9	0.1
F	0.05	0.95

A	P(M=T|A)	P(M=F|A)
T	0.7	0.3
F	0.01	0.99

Using Bayes Nets - Example (6)



3. Factor joint probability using Bayes Net

$$p(b \mid j, m) = \frac{p(b, j, m)}{p(j, m)} = \alpha \cdot p(b, j, m)$$

$$= \alpha \sum_{h_1} \sum_{h_2} p(b, Th = h_1, Al = h_2, j, m)$$

$$= \alpha \sum_{h_1} \sum_{h_2} p(j, m \mid Al = h_2, b, Th = h_1) p(Al = h_2, b, Th = h_1)$$

$$= \alpha \sum_{h_1} \sum_{h_2} p(j \mid Al = h_2) p(m \mid Al = h_2) p(Al = h_2, b, Th = h_1)$$

$$= \alpha \sum_{h_1} \sum_{h_2} p(j \mid Al = h_2) p(m \mid Al = h_2) p(Al = h_2 \mid b, Th = h_1) p(b, Th = h_1)$$

$$= \alpha \sum_{h_1} \sum_{h_2} p(j \mid Al = h_2) p(m \mid Al = h_2) p(Al = h_2 \mid b, Th = h_1) p(b) p(Th = h_1)$$

Using Bayes Nets - Example (7)

$$p(b \mid j, m) = \alpha \sum_{h_1} \sum_{h_2} p(j \mid Al = h_2) p(m \mid Al = h_2) p(Al = h_2 \mid b, Th = h_1) p(b) p(Th = h_1)$$

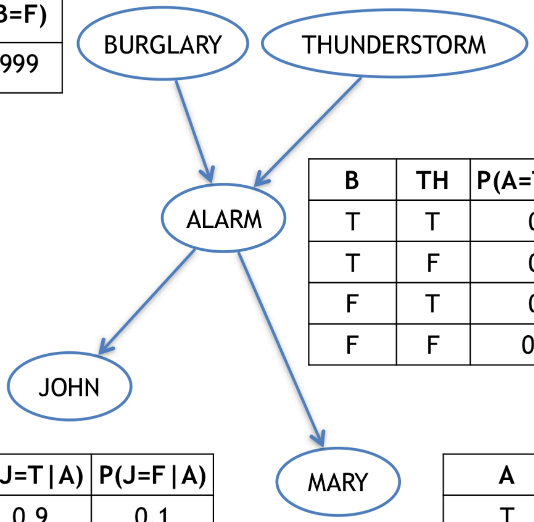
4. Simplify and plug in CPTs

$$= \alpha \cdot p(b) \sum_{h_1} p(Th = h_1) \sum_{h_2} p(j \mid Al = h_2) p(m \mid Al = h_2) p(Al = h_2 \mid b, Th = h_1)$$

p(b|j,m) = α (0.001)*[
(0.002)*
[(0.9)(0.7)(0.95) + (0.05)(0.01)(0.05)]+
(0.998)*
[(0.9)(0.7)(0.94) + (0.05)(0.01)(0.06)]]
= α (0.00059224259)

To find alpha, repeat for p(¬b|j,m)

P(B=T)	P(B=F)
0.001	0.999



P(TH=T)	P(TH=F)
0.002	0.998

B	TH	P(A=T|B,TH)	P(A=F|B,TH)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(J=T|A)	P(J=F|A)
T	0.9	0.1
F	0.05	0.95

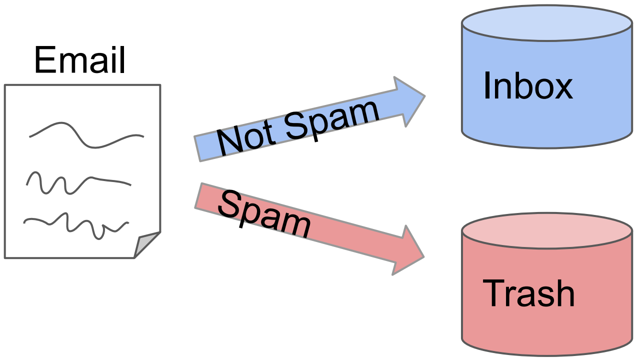
A	P(M=T|A)	P(M=F|A)
T	0.7	0.3
F	0.01	0.99

Naive Bayes Classifier (1)

We can use the Bayes Net framework to introduce a simple and effective technique from Machine Learning known as the **Naive Bayes Classifier**

Task: label an email as **spam** or **notspam**

How can we train an agent to do this for us by giving it lots and lots of examples of spam and notspam emails?



Naive Bayes Classifier (2)

For any email, let's associate a set of binary RVs that correspond with "features" of the email that we can easily measure and we think correspond with whether an email is spam or not

Example Features

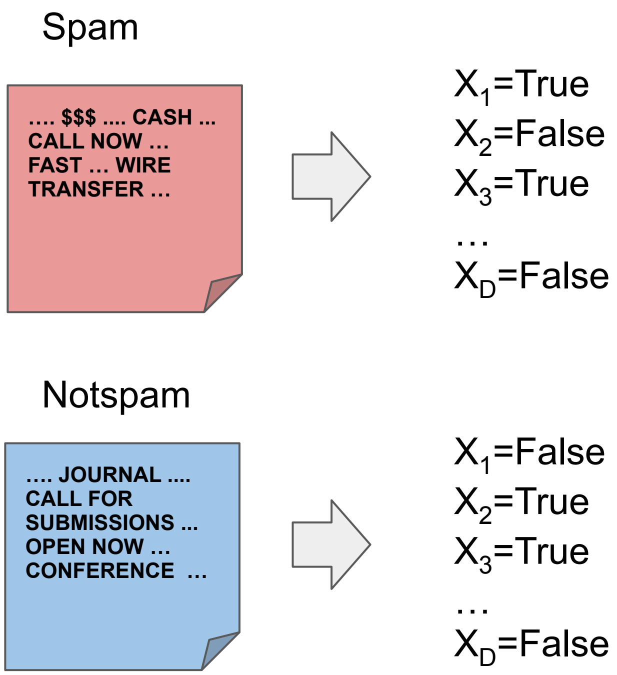
X_1 = Contains "CASH!"

X_2 = From an address in my contact list

X_3 = Contains no text (only images)

…

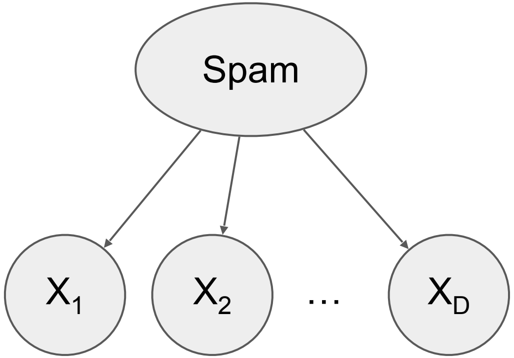
We'll manually label some training data as spam and notspam



Naive Bayes Classifier (3)

Now let's make a **big** assumption: Each X_i is independent of the other X_j **given** whether the email is spam or not.

If this were true, we could **factor** the probability that a given email was spam or not!



$$
\begin{aligned}
p(Spam \mid X_1, X_2, \dots, X_D) &= p(X_1, \dots X_D \mid Spam) \frac{p(Spam)}{p(X_1, \dots, X_D)} \\
&= \alpha \cdot p(Spam) p(X_1, \dots, X_D \mid Spam) \\
&= \alpha \cdot p(Spam) \prod_{j=1}^{D} p(X_j \mid Spam)
\end{aligned}
$$

Estimating probabilities from data

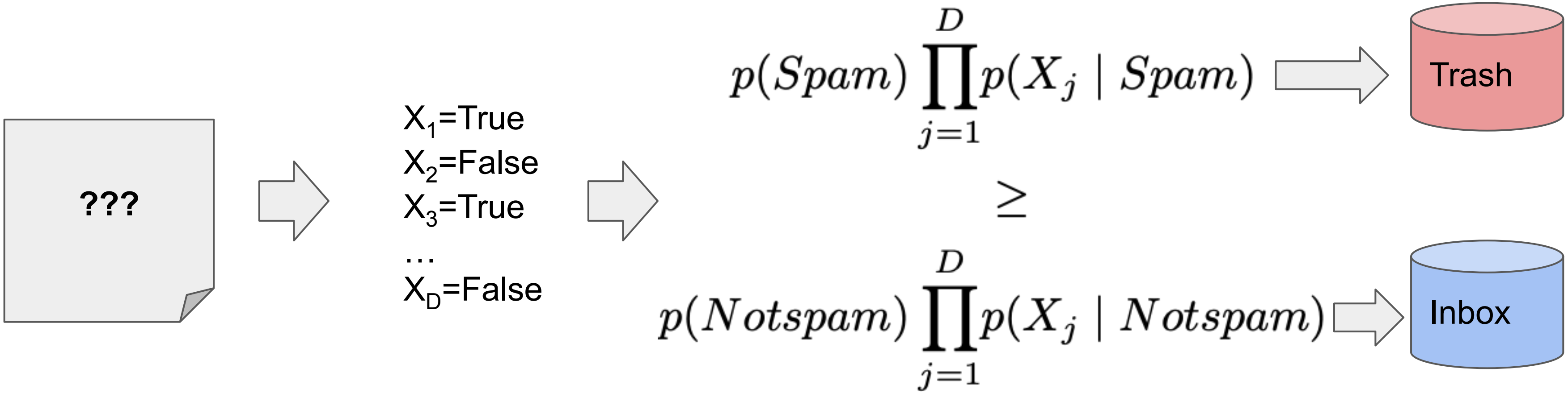
Class label probability

$$\hat{p}(y = \text{Spam}) = \frac{\text{Number of training samples labeled `Spam'}}{\text{Number of training samples total}}$$

"Feature" conditional probability

$$\hat{p}(x_i = b \mid y = \text{Spam}) = \frac{\text{Number of training samples where } x_i = b, y = \text{Spam}}{\text{Number of training samples where } y = \text{Spam}}$$

Classifying spam



Naive Bayes notes

Properties

- Works incredibly (as a **classifier**) well in practice, even though the Naive Bayes assumption is often completely wrong
- Is a **generative** model: can actually “produce” spam emails by sampling according to the distribution
- Generalizes to non-binary features and classes

Implementation

- Need to be careful about picking the features (zero counts) and computing product with many terms (log space)

Summary and preview

Wrapping up

- We can use **directed graphs** to capture our intuitive notion of independence
- This also allows us to break a single large joint probability table into several smaller **c**onditional **p**robability **t**ables (CPTs)
- This **factored** representation lets us answer any question we would need to use the joint probability table for, potentially saving on computation too

Next time

- Probabilities in time (Filtering and Smoothing)