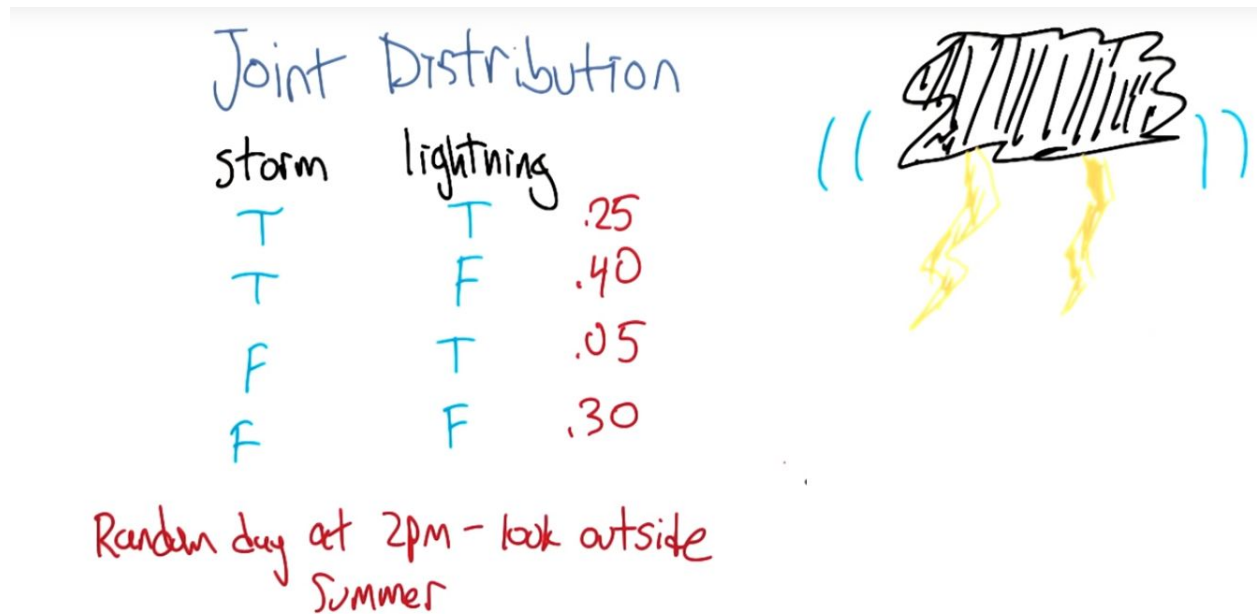


## 1. Introduction



transcript

M: Hey Charles.

C: Hey Michael.

M: So like I get to lecture near you today.

C: Yes you do. I can even see you.

M: This is, this is crazy. I sort of don't have my regular pad. This makes me a little uncomfortable.

C: But you look very dashing in your nice blue suit.

M: Thanks. We're going to record some live action stuff today.

C: Mm.

M: [LAUGH] All right so. Do you remember last time we were talking about Bayesian learning?

C: I do, because I led that.

M: Right. Good point. And so one of the questions that I asked as a follow-up was, these quantities, these probabilistic quantities that we're working with. Is there's anything that we need to know about how to represent and reason with them. And you said that I should look into it. Yeah, because I, I just, I yeah you should look into it.

C: So I did. So, and it's cool. And so I figured it would be fun to tell you about it.

M: Okay, well I look forward to it.

C: Thanks! And also I want to point out, we're using a different color scheme today. Isn't that a nice blue?

M: It is a nice blue, its sort of a relaxing blue. As opposed to that blue blue that we used before.

C: It's like Cerulean... Is it?

M: No.

C: It's more like periwinkle.

M: No, it's definitely not periwinkle.

C: Oh you're right, it's not periwinkle.

M: Navy.

C: No, it's too light to be navy.

M: All right, so so good, so right. It turns out that there's this concept called Bayesian Networks, which is this wonderful representation for representing and manipulating probabilistic quantities over complex spaces. And so it fits in really well with the stuff you were talking about last time.

## 2. Join Distribution

*transcript*

M: Alright, so to make this work, we're going to need to build on this idea of a joint distribution. It's not going to be obvious right away what this has to do with machine

learning, at all. But, I, it's going to connect. So, just bear with me for a little bit. Alright, so to talk about this concept, what we're going to do is look at an example. And the example that I think might work, that would be nice and simple, is the notion of storm and lightning. So, here's a little picture of storm and lightning. And what we're going to do is say, let's say, on a random day, at 2 PM. You look outside. And, what I want you to do is say, what fraction of the time, is, is each of these different possible combination of things happening? So, for example, what's the probability that you look out and there's a storm and there's lightning at the same time? So, what do you think?

C: On a random day?

M: Yeah, random day at 2 PM. And we can be in Atlanta since that's what you're familiar with.

C: Is it summer? Because that happens more often in the summer.

M: Sure, let's say summer.

C: It's fairly high at 2 PM. Let's say it happens a quarter of the time.

M: Wow, that's a rainy summer.

C: Mm-hm.

M: Alright. Now, that's not the only possibility though. It could also be that there's a storm but no lightning.

C: Right. That happens more often at 2 PM in the summer in Atlanta. Let's say it's mm, .4.

M: Wow. Alright. Now what's the probability that you look at the window and there's no storm but there is lightning.

C: Maybe 5%.

M: And what's the probability that you look out and there's, you know, it's nice clear there's no storm no lightning.

C: Coincidentally I picked numbers that made it easier for me to subtract from one. So, it's 0.3.

M: [LAUGH] Right and so these, there's only, these are the only four possibilities. We're saying. And they, so they have to add up to 100%. And so I, yeah it had to be 30 at this

point. So, it's actually more likely that there's a storm than not, according to what you said.

C: It's Atlanta in the summer at 2 PM.

M: There you go. Alright. So, this is a joint distribution. And now we can actually ask various kinds of questions about this. Oh, you know what would be a good form for asking a question.

M: I don't know. I'm looking at you quizzically.

C: Nice. Using the fact that we are in the same place. We are going to do a quiz.

## Joint Distribution

storm	lightning	
T	T	.25
T	F	.40
F	T	.05
F	F	.30

Random day at 2pm - look outside  
Summer



Quiz:

$\Pr(\neg \text{storm}) =$

$\Pr(\text{lightning} | \text{storm}) =$

### 3. Quiz: Joint Distribution

Joint Distribution

storm	lightning	
T	T	.25
T	F	.40
F	T	.05
F	F	.30

Random day at 2pm - look outside  
Summer



Quiz:

$$\Pr(\neg \text{storm}) = \frac{.30 + .05}{.65} = \boxed{.35}$$

$$\Pr(\text{lightning} | \text{storm}) = \frac{.25}{.65} = \boxed{.4015}$$

### 4. Adding Attributes

Joint Distribution

storm	lightning	thunder	
T	T	T	.20
T	T	F	.05
T	F	T	.09
T	F	F	.36
F	T	T	.04
F	T	F	.01
F	F	T	.03
F	F	F	.27

Random day at 2pm - look outside  
Summer



## 5. Conditional Independence

### Conditional Independence

Definition:  $X$  is conditionally independent of  $Y$  given  $Z$  if the probability distribution governing  $X$  is independent of the value of  $Y$  given the value of  $Z$ ; that is, if

$$\forall x, y, z \quad P(X=x | Y=y, Z=z) = P(X=x | Z=z)$$

more compactly we write

$$P(X|Y, Z) = P(X|Z)$$

Independence

$$Pr(x, y) = Pr(x) \cdot Pr(y)$$

chain rule

$$Pr(x, y) = Pr(x|y) \cdot Pr(y)$$

$$\therefore Pr(x|y) = Pr(x) \quad \checkmark$$

## 6. Quiz: Conditional

### Joint Distribution

storm	lightning	thunder
T	T	.25
T	F	.40
F	T	.05
F	F	.30



QUIZ:

$$P(\text{thunder} = \boxed{\phantom{0}} | \text{lightning} = \boxed{\phantom{0}} | \text{storm} = T)$$

Random day at 2pm - look outside = Summer

$$P(\text{thunder} = \boxed{\phantom{0}} | \text{lightning} = \boxed{\phantom{0}} | \text{storm} = F)$$



## Joint Distribution

storm	lightning	thunder
T	T	.25
T	F	.40
F	T	.05
F	F	.30



QUIZ:

$$P(\text{thunder} = T \mid \text{lightning} = F, \text{storm} = T)$$

Random day at 2pm - look outside = Summer  
 $P(\text{thunder} = F \mid \text{lightning} = F, \text{storm} = F)$

$$Pr(Th|L,S) = Pr(Th|L) \text{ OR } \text{CONDITIONALLY INDEPENDENT}$$

## 7. Quiz: Belief Network

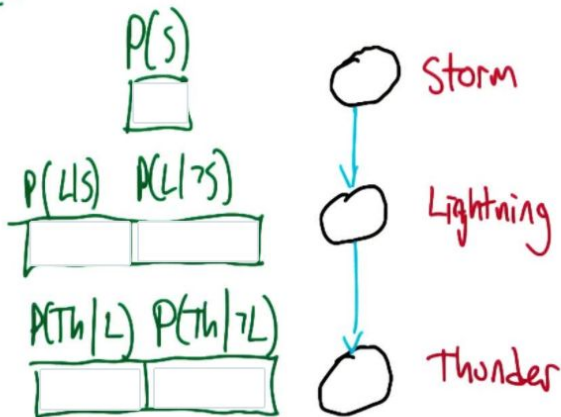
Belief Networks

aka Bayes Nets

aka Bayesian Networks

aka Graphical Models

QUIZ:



storm	lightning	thunder
T	T	.25
T	F	.40
F	T	.05
F	F	.30

QUIZ:

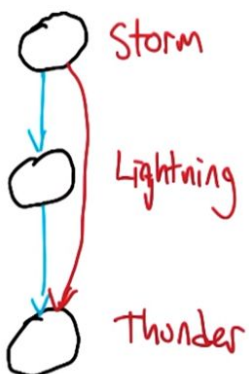
Belief Networks

aka Bayes Nets

aka Bayesian Networks

aka Graphical Models

$P(S)$	
.65	
$P(L S)$	$P(L \neg S)$
.385	.143
$P(Th L)$	$P(Th \neg L)$
.8	.1



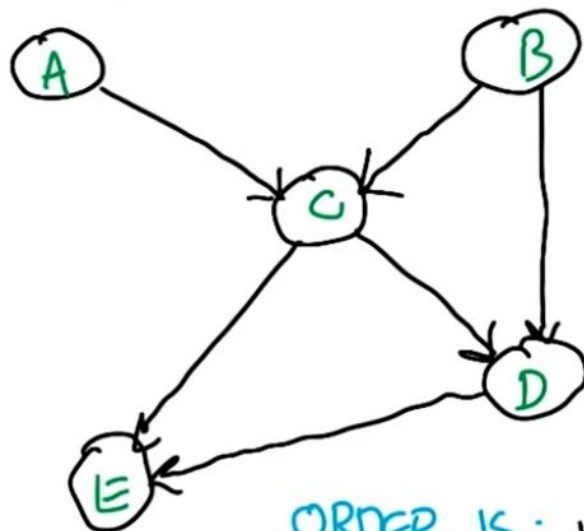
storm	lightning	thunder
T	T	.25
T	F	.40
F	T	.05
F	F	.30

grows exponentially with more variables! (indegree)  
(parents)



## 8. Quiz: Sampling From The Joint Distribution

Sampling From The Joint Distribution



Sample

$$A \sim P(A)$$

$$B \sim P(B)$$

$$C \sim P(C|A,B)$$

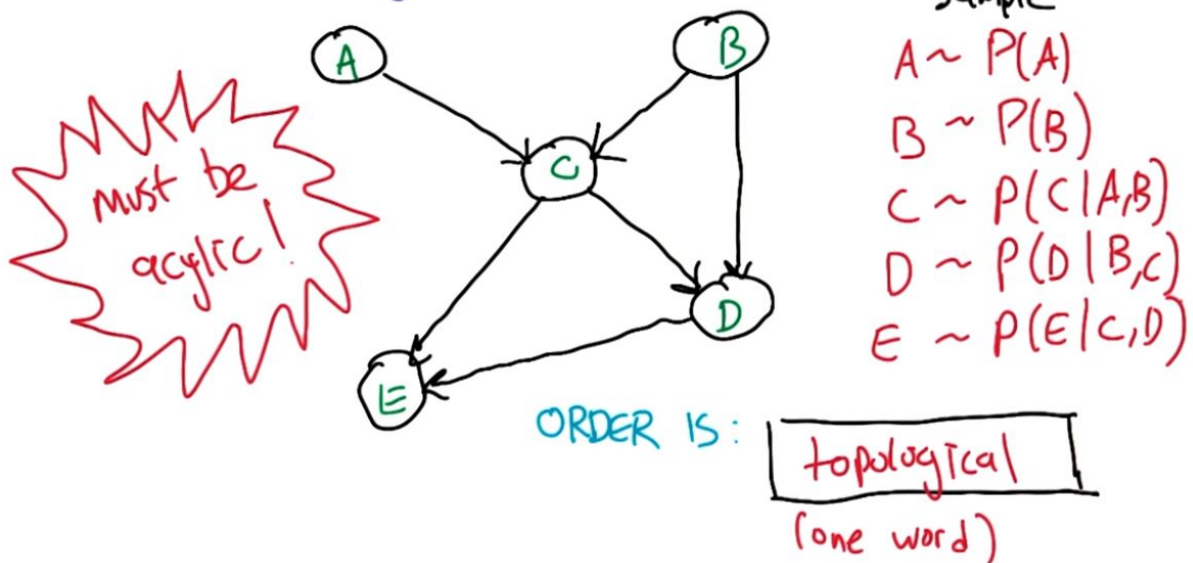
$$D \sim P(D|B,C)$$

$$E \sim P(E|C,D)$$

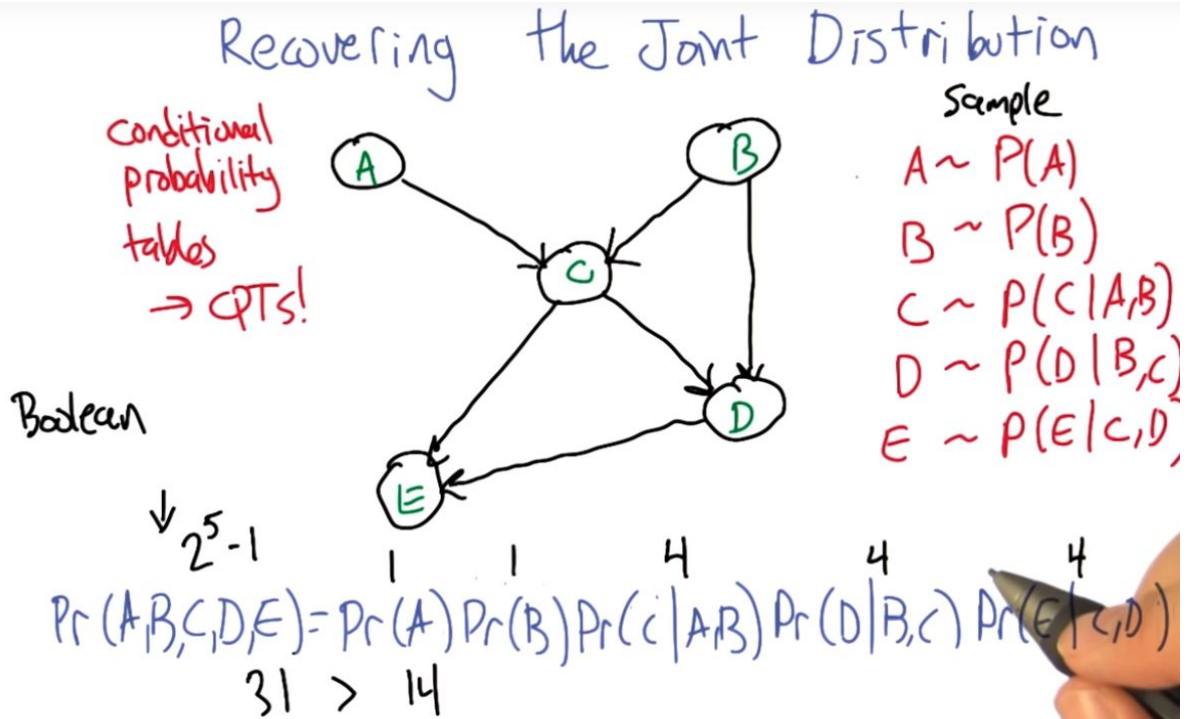
ORDER IS:

(one word)

## Sampling From the Joint Distribution



## 9. Recovering the Joint Distribution



## 10. Sampling

Why sampling?

- two things distributions are for
    - probability of value
    - generate values
  - simulation of a complex process
  - approximate inference
    - machine
  - visualization - get a feel.
    - human
- exact: hard  
approximate: faster

## 11. Inferencing Rules

### Inferencing Rules

Marginalization

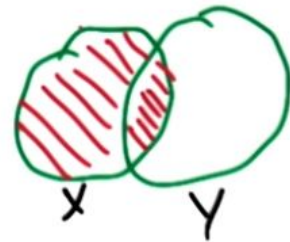
$$P(x) = \sum_y P(x, y)$$

chain rule

$$P(x, y) = P(x) P(y|x)$$

Bayes rule

$$P(y|x) = \frac{P(x|y) P(y)}{P(x)}$$



## 12. Quiz: Inferencing Rules

Inferencing Rules

*Marginalization*  

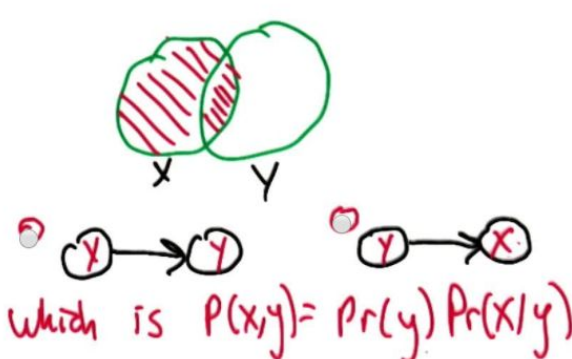
$$P(x) = \sum_y P(x, y)$$

*chain rule*  

$$P(x, y) = P(x) P(y|x)$$

*Bayes rule*  

$$P(y|x) = \frac{P(x|y) P(y)}{P(x)}$$



The Venn diagram shows two overlapping circles, X and Y. The intersection is shaded with red diagonal lines. The two DAGs below illustrate the chain rule. The first DAG has a node Y pointing to a node X, representing the joint probability P(x, y) = P(y)P(x|y). The second DAG has a node X pointing to a node Y, representing the joint probability P(x, y) = P(x)P(y|x).

which is  $P(x, y) = P(y) P(x|y)$

*transcript*

M: All right. So, person who's adept at manipulating Bayes Nets would know that this chain rule idea, this probability of X and Y can be written either as a probability of X times the probability of Y given X. Or as the probability of Y times the probability of X given Y, actually correspond to two different networks. So which of these two networks corresponds to the fact that the probability of x and y, the joint probability of X and can be written as the probability of Y times the probability of X given Y.

C: Go

*Answer*

M: Did you get it?

C: Yeah I did actually. so, so this one I think I understand completely. So we know that from the last discussion we had about how you would recover the joint, that what you're saying on the right of this equation probability y times probability n y means that the probability of y, the variable y doesn't depend on anything. So, between those two

graphs the one on the right is the one where you're saying that. You don't need to know the value of any other variable in order to determine the probability of  $y$ .

M: Good.

C: So it has to be the one on the sec, the second and just to make sure if you look at the second product the probability of  $x$  given  $y$  the second multican? Is it multican?

M: Hm, factor.

C: Factor? Let's say factor. The second factor, this says that while you determine the probability of  $x$  given the value of  $y$  and there is an arrow from  $y$  to  $x$  so, the second one is in fact correct.

M: Yeah. So this is actually just one way you could just read this network is to say what is this node  $x$  with an arrow coming into it? That is the probability of  $x$ . But, the, the things pointing into it are what's exactly being given. What it's being conditioned on. So that's exactly right, the second one.

C: Right. So this, this, so this makes sense to me. This is why when you look at a network, network, it's very hard not to think of them as dependencies. Even though they're not dependencies, they're conditional independencies.

M: Well the arrows are a form of dependence but it's not a causal dependence necessarily, it's it's again it's just the way the probabilities are being decomposed.

C: Hm.

M: And the last of these three equations just Baye's rule, this time written correctly where the denominator has to be the probability of  $x$ , and we've gone over this a couple of times. I don't, I don't need to, to describe it again, but what Would like to, just, bring to your attention to this three together turn out to be kind of our, you know, three musketeers in working out the probability of various kinds of events.

C: Excellent.



# Inferencing Rules

Marginalization

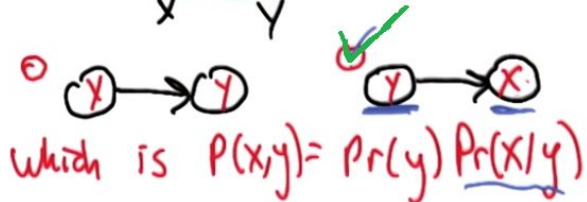
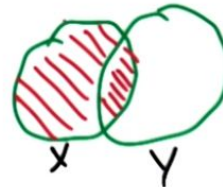
$$P(x) = \sum_y P(x, y)$$

chain rule

$$P(x, y) = P(x) P(y|x)$$

Bayes rule

$$P(y|x) = \frac{P(x|y) P(y)}{P(x)}$$



which is  $P(x, y) = P(y) P(x|y)$

## 13. Quiz: Inferencing by Hand

Inference By Hand

$P(\text{Box}=1) = 1/2$

Ball 1

	G	Y	B
Box 1	3/4	1/4	0
Box 2	2/5	0	3/5

Box 1

Box 2

Ball 2

	G	Y	B
Box 1	2/3	1/3	0
Box 2	1/4	0	3/4

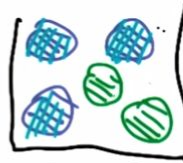
$P(2 = \text{blue} | 1 = \text{green})$

=

## Inference By Hand



Box=1



Box=2

$$P(2=\text{blue} | 1=\text{green}) = \boxed{\phantom{000}}$$

marginalization rule + chain

$$\begin{aligned}
 P(2=\text{blue} | 1=\text{green}) &= \\
 &P(2=\text{blue} | 1=\text{green}, \text{Box}=1) P(\text{Box}=1 | 1=\text{green}) \\
 &+ P(2=\text{blue} | 1=\text{green}, \text{Box}=2) P(\text{Box}=2 | 1=\text{green})
 \end{aligned}$$

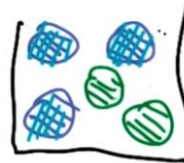
$$\begin{aligned}
 P(\text{Box}=1 | 1=\text{green}) &= P(1=\text{green} | \text{Box}=1) P(\text{Box}=1) / P(1=\text{green}) \\
 P(\text{Box}=2 | 1=\text{green}) &= P(1=\text{green} | \text{Box}=2) P(\text{Box}=2) / P(1=\text{green})
 \end{aligned}$$

$3/4 = 15/40 \rightarrow 15/23$   
 $1/5 = 8/40 \rightarrow 8/23$

## Inference By Hand



Box=1



Box=2

$$P(2=\text{blue} | 1=\text{green}) = \boxed{6/23}$$

marginalization rule + chain

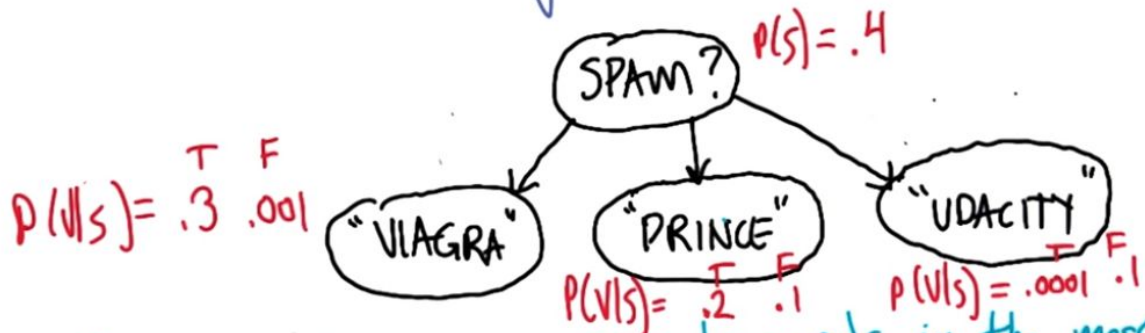
$$\begin{aligned}
 P(2=\text{blue} | 1=\text{green}) &= \\
 &P(2=\text{blue} | 1=\text{green}, \text{Box}=1) P(\text{Box}=1 | 1=\text{green}) \\
 &+ P(2=\text{blue} | 1=\text{green}, \text{Box}=2) P(\text{Box}=2 | 1=\text{green})
 \end{aligned}$$

$$\begin{aligned}
 P(\text{Box}=1 | 1=\text{green}) &= P(1=\text{green} | \text{Box}=1) P(\text{Box}=1) / P(1=\text{green}) \\
 P(\text{Box}=2 | 1=\text{green}) &= P(1=\text{green} | \text{Box}=2) P(\text{Box}=2) / P(1=\text{green})
 \end{aligned}$$

$3/4 = 15/40 \rightarrow 15/23$   
 $1/5 = 8/40 \rightarrow 8/23$

## 14. Naive Bayes

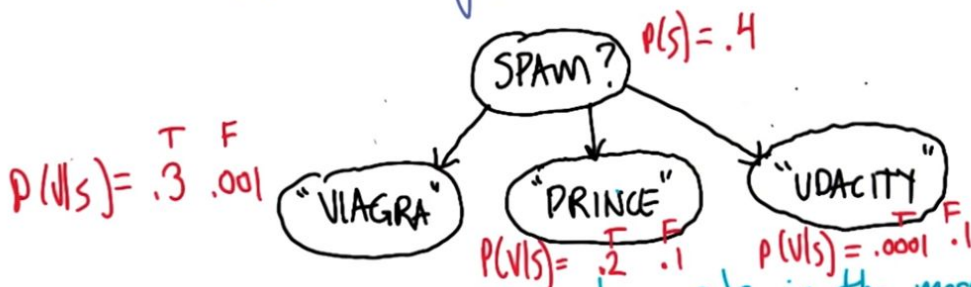
Naïve Bayes : Spectral Case



Know you are in spam  $\rightarrow$  generate words in the message.

$$P(\text{SPAM?} \mid \text{VIAGRA, not PRINCE, not UDACITY}) = \frac{P(\text{VIAGRA, not PRINCE, not UDACITY} \mid \text{SPAM?}) P(\text{SPAM?})}{\dots}$$

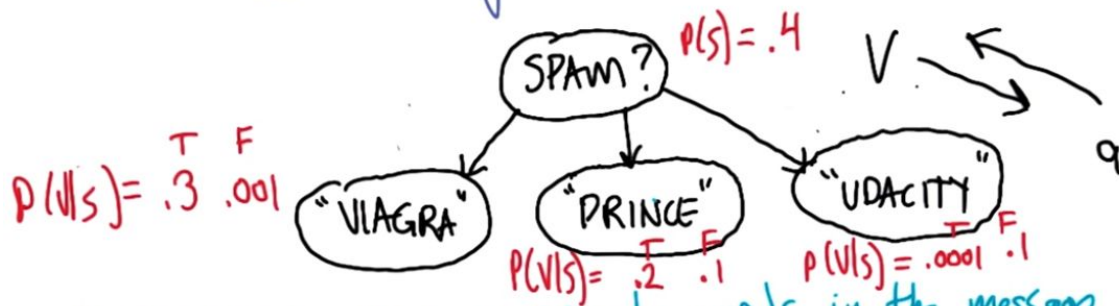
Naïve Bayes : Spectral Case



Know you are in spam  $\rightarrow$  generate words in the message.

$$\begin{aligned} P(\text{SPAM?} \mid \text{VIAGRA, not PRINCE, not UDACITY}) &= \frac{P(\text{VIAGRA, not PRINCE, not UDACITY} \mid \text{SPAM?}) P(\text{SPAM?})}{\dots} \\ &= \frac{P(\text{VIAGRA} \mid \text{SPAM?})^3 P(\text{not PRINCE} \mid \text{SPAM?})^8 P(\text{not UDACITY} \mid \text{SPAM?})^{.9999} P(\text{SPAM?})}{\dots} \end{aligned}$$

## Naïve Bayes : Special Case



Know you are in spam  $\rightarrow$  generate words in the message.

$$P(V | q_1, q_2, \dots, q_n) = \prod_i P(q_i | V) \cdot P(V) / Z$$

$$\text{MAP class} = \arg \max V \prod_i P(q_i | V) \quad \text{MAP SPAM}$$

## 15. Why Naive Bayes is Cool

### Why Naïve Bayes is Cool



- $\rightarrow$  Inference is cheap
- $\rightarrow$  Few parameters
- $\rightarrow$  Estimate parameters with labeled data
- $\rightarrow$  Connects inference and classification
- $\rightarrow$  Empirically Successful

$$P(q_i | V) = \frac{\# q_i, V}{\# V}$$

NO FREE LUNCH!

Doesn't model interrelationships between attributes.



## Why Naive Bayes Is Cool

→ Inference is cheap

→ Few parameters

→ Estimate parameters with labeled data

→ Connects inference and classification

→ Empirically Successful



$$P(q_i|V) = \frac{\# q_i, V}{\# V}$$

one unseen attribute spoils the whole bunch, girl.

"Smooth"  
Inductive bias

NO FREE LUNCH!

Does it model interrelationships between attributes. (ordering preserved)

## 16. Wrapping Up

### Wrapping UP

Bayes Networks - representation of joint distributions

Examples of using networks to compute probabilities

Sampling as a way to do approximate inference

In general, hard to do exact inference

Naive Bayes - link to classification

- tractible

- gold standard

- inference in any direction (missing attributes)

