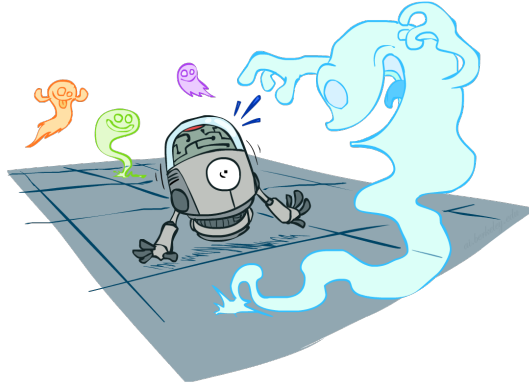


CSE 473: Artificial Intelligence

Probability Review... → Markov Models



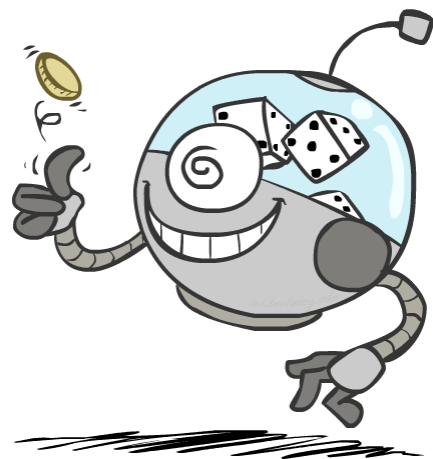
Daniel Weld

University of Washington

[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley. All CS188 materials are available at <http://ai.berkeley.edu>.]

Outline

- **Probability**
 - Random Variables
 - Joint and Marginal Distributions
 - Conditional Distribution
 - Product Rule, Chain Rule, Bayes' Rule
 - Inference
 - Independence & Conditional Independence
 - ... Markov Models
- You'll need all this stuff A LOT for the next few weeks, so make sure you go over it now!



Joint Distributions

- A *joint distribution* over a set of random variables: X_1, X_2, \dots, X_n specifies a probability for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

$$P(x_1, x_2, \dots, x_n)$$

- Must obey: $P(x_1, x_2, \dots, x_n) \geq 0$

$$\sum_{(x_1, x_2, \dots, x_n)} P(x_1, x_2, \dots, x_n) = 1$$

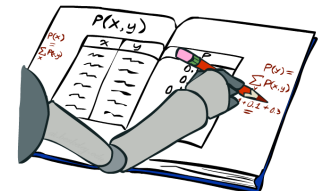
$P(T, W)$

| T | W | P |
|------|------|-----|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |

- Size of joint distribution if n variables with domain sizes d?
 - For all but the smallest distributions, impractical to write out!

Marginal Distributions

- Marginal distributions are **sub-tables** which eliminate variables
- Marginalization* (summing out): Combine collapsed rows by adding



$P(T, W)$

| T | W | P |
|------|------|-----|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |

$$P(t) = \sum_s P(t, s)$$

$P(T)$

| T | P |
|------|-----|
| hot | 0.5 |
| cold | 0.5 |

$$P(s) = \sum_t P(t, s)$$

$P(W)$

| W | P |
|------|-----|
| sun | 0.6 |
| rain | 0.4 |

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Joint Distribution
 $P(T, W)$

| T | W | P |
|------|------|-----|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

| W | P |
|------|-----|
| sun | 0.8 |
| rain | 0.2 |

$P(W|T = cold)$

| W | P |
|------|-----|
| sun | 0.4 |
| rain | 0.6 |

Normalization Trick

$P(T, W)$

| T | W | P |
|------|------|-----|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |

SELECT the joint probabilities matching the evidence

$P(c, W)$

| T | W | P |
|------|------|-----|
| cold | sun | 0.2 |
| cold | rain | 0.3 |

NORMALIZE the selection (make it sum to one)

$P(W|T = c)$

| W | P |
|------|-----|
| sun | 0.4 |
| rain | 0.6 |

- Why does this work? Sum of selection is P(evidence)! (P(T=c), here)

$$P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{P(x_1, x_2)}{\sum_{x_1} P(x_1, x_2)}$$

Probabilistic Inference

- Probabilistic inference = *“compute a desired probability from other known probabilities (e.g. conditional from joint)”*
- We generally compute conditional probabilities
 - $P(\text{on time} \mid \text{no reported accidents}) = 0.90$
 - These represent the agent’s *beliefs* given the evidence
- Probabilities change with new evidence:
 - $P(\text{on time} \mid \text{no accidents, 5 a.m.}) = 0.95$
 - $P(\text{on time} \mid \text{no accidents, 5 a.m., raining}) = 0.80$
 - Observing new evidence causes *beliefs to be updated*



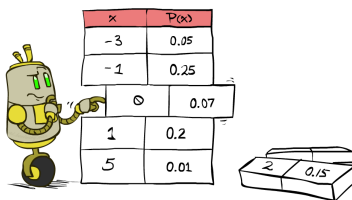
Inference by Enumeration

- General case:
 - Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
 - Query* variable: Q
 - Hidden variables: $H_1 \dots H_r$

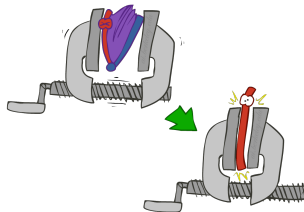
- We want: $P(Q \mid e_1 \dots e_k)$

** Works fine with multiple query variables, too*

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \dots e_k)$$

$$P(Q \mid e_1 \dots e_k) = \frac{1}{Z} P(Q, e_1 \dots e_k)$$

$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} P(Q, h_1 \dots h_r, e_1 \dots e_k)$$

X_1, X_2, \dots, X_n

Example: Inference by Enumeration

$P(W=\text{sun} \mid S=\text{winter})?$

1. Select data consistent with evidence

| S | T | W | P |
|--------|------|------|------|
| summer | hot | sun | 0.30 |
| summer | hot | rain | 0.05 |
| summer | cold | sun | 0.10 |
| summer | cold | rain | 0.05 |
| winter | hot | sun | 0.10 |
| winter | hot | rain | 0.05 |
| winter | cold | sun | 0.15 |
| winter | cold | rain | 0.20 |

Example: Inference by Enumeration

$P(W=\text{sun} \mid S=\text{winter})?$

1. Select data consistent with evidence
2. Marginalize away hidden variables (sum out temperature)




| S | T | W | P |
|--------|------|------|------|
| summer | hot | sun | 0.30 |
| summer | hot | rain | 0.05 |
| summer | cold | sun | 0.10 |
| summer | cold | rain | 0.05 |
| winter | hot | sun | 0.10 |
| winter | hot | rain | 0.05 |
| winter | cold | sun | 0.15 |
| winter | cold | rain | 0.20 |

Example: Inference by Enumeration

$P(W=\text{sun} \mid S=\text{winter})?$

1. Select data consistent with evidence
2. Marginalize away hidden variables
(sum out temperature)
3. Normalize

| S | T | W | P |
|--------|------|------|------|
| summer | hot | sun | 0.30 |
| summer | hot | rain | 0.05 |
| summer | cold | sun | 0.10 |
| summer | cold | rain | 0.05 |
| winter | hot | sun | 0.10 |
| winter | hot | rain | 0.05 |
| winter | cold | sun | 0.15 |
| winter | cold | rain | 0.20 |




| S | W | P |
|--------|------|------|
| winter | sun | 0.25 |
| winter | rain | 0.25 |

Example: Inference by Enumeration

$P(W=\text{sun} \mid S=\text{winter})?$

1. Select data consistent with evidence
2. Marginalize away hidden variables
(sum out temperature)
3. Normalize

| S | T | W | P |
|--------|------|------|------|
| summer | hot | sun | 0.30 |
| summer | hot | rain | 0.05 |
| summer | cold | sun | 0.10 |
| summer | cold | rain | 0.05 |
| winter | hot | sun | 0.10 |
| winter | hot | rain | 0.05 |
| winter | cold | sun | 0.15 |
| winter | cold | rain | 0.20 |



| S | W | P |
|--------|------|------|
| winter | sun | 0.50 |
| winter | rain | 0.50 |

Inference by Enumeration

- Computational problems?
 - Worst-case time complexity $O(d^n)$
 - Space complexity $O(d^n)$ to store the joint distribution

Don't be Fooled

- It may look cute...



https://fc08.deviantart.net/fs71/2010/258/4/4/baby_dragon_charles_by_imsorrybuti-d2yl11.png

The Sword of Conditional Independence!



Slay
the
Basilisk!

I am a BIG joint
distribution!



$X \perp\!\!\!\perp Y | Z$ Means: $\forall x, y, z : P(x, y | z) = P(x | z)P(y | z)$

Or, equivalently: $\forall x, y, z : P(x | z, y) = P(x | z)$

40

A Brief Trip Forward in Time...



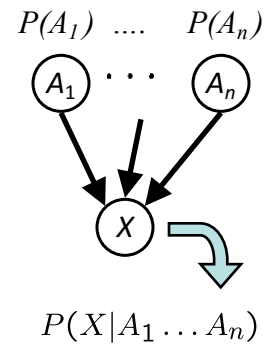
41

Preview: Bayes Nets Encode Joint Distributions

- A set of nodes, one per variable X
- A directed, acyclic graph
- A **conditional distribution** for each node
 - A collection of distributions over X , one for each combination of parents' values

$$P(X|a_1 \dots a_n)$$

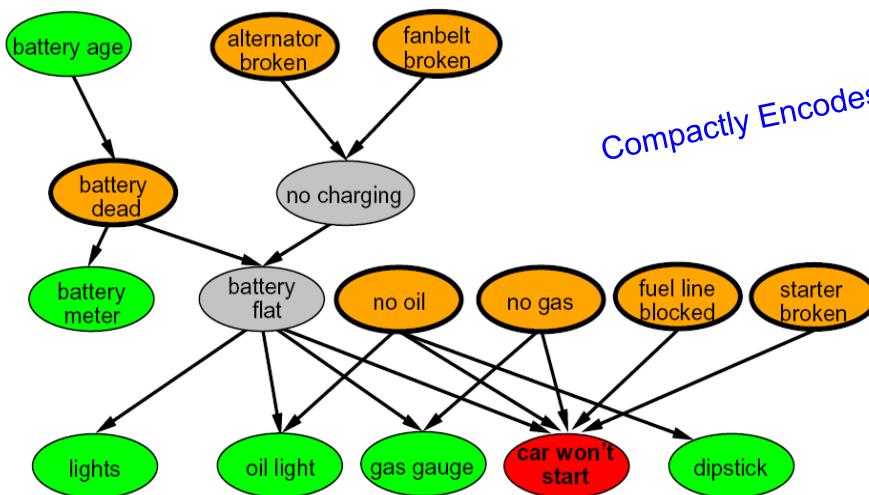
- CPT: conditional probability table
- Description of a noisy "causal" process



A Bayes net = Topology (graph) + Local Conditional Probabilities

Benefits: Smaller, Allows Fast Inference, Learnable!

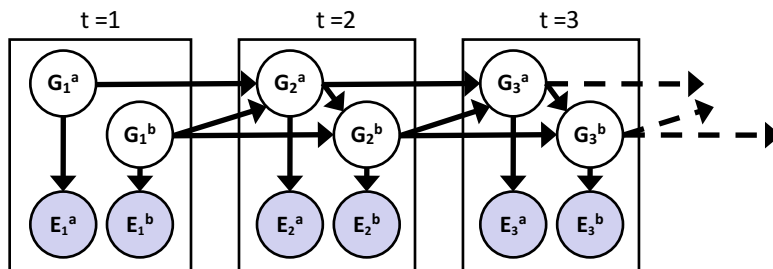
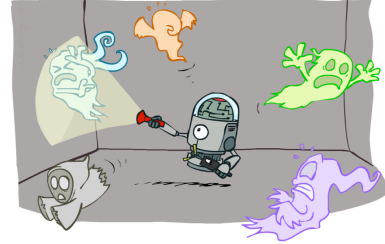
Preview: Example Bayes Net - Car



Compactly Encodes 2^{16} Parameters!

Preview: Dynamic Bayes Nets (DBNs) - Ghosts

- We want to track multiple variables over time, using multiple sources of evidence
- Idea: Repeat a fixed Bayes net structure at each time
 - Generalization of Hidden Markov Models (HMMs)
 - Itself a generalization of Markov Models
- Variables from time t may condition on those from $t-1$



Back to Our Own Universe... (for now)



Ghostbusters, Revisited

- Let's say we have two distributions:

- Prior distribution** over ghost location: $P(G)$
 - Let's say this is uniform
 - Sensor reading model: $P(R | G)$
 - Given: we know what our sensors do
 - R = reading color measured at $(1,1)$
 - E.g. $P(R = \text{yellow} | G=(1,1)) = 0.1$

| | | |
|------|------|------|
| 0.11 | 0.11 | 0.11 |
| 0.11 | 0.11 | 0.11 |
| 0.11 | 0.11 | 0.11 |

| | | |
|-------|------|------|
| 0.17 | 0.10 | 0.10 |
| 0.09 | 0.17 | 0.10 |
| <0.01 | 0.09 | 0.17 |

- We can calculate the **posterior distribution** $P(G | r)$ over ghost locations given a reading using Bayes' rule:

$$P(g|r) \propto P(r|g)P(g)$$

[Demo: Ghostbuster – with probability (L12D2)]

What's Our Probabilistic Model

- Random Variables

- Location of Ghost. Values = $\{L_{1,1}, L_{1,2}, \dots, L_{6,10}\}$
 - Sensor value at locations $S_{1,1}, \dots, S_{6,10}$. Values = $\{R, O, Y, G\}$

- Joint Distribution

- Too big to write down $60 * 4^{60} = 7.98 * 10^{37}$
 - Here's a **schema** for a **conditional distribution** specifying part of it:

| | | | |
|------------|---------------|---------------|--------------|
| P(red 3) | P(orange 3) | P(yellow 3) | P(green 3) |
| 0.05 | 0.15 | 0.50 | 0.30 |
| . . . | | | |
| P(red 0) | P(orange 0) | P(yellow 0) | P(green 0) |
| 0.70 | 0.15 | 0.10 | 0.05 |

Model for a Tiny Ghostbuster

- Random Variables

- Location of Ghost, G. Values = {L1, L2}
- Sensor value at locations S1, S2 with values {R, O, Y, G}



- Joint Distribution

| | | G=L1 | | | | G=L2 | | | |
|----|---|-------------|---|---|---|------|---|---|---|
| | | S2 | | | | S2 | | | |
| | | R | O | Y | G | R | O | Y | G |
| S1 | R | Select G=L1 | | | | | | | |
| | O | | | | | | | | |
| | Y | | | | | | | | |
| | G | | | | | | | | |

$$\sum_{S2}$$

Can marginalize to get $P(S1 \mid \text{distance} = 0)$

48

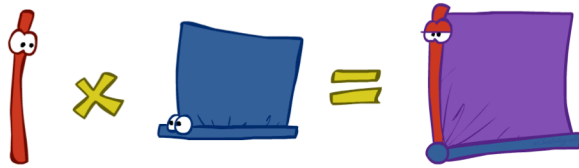
Video of Demo Ghostbusters with Probability



The Product Rule

- Sometimes have conditional distributions but want the joint

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



The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i|x_1 \dots x_{i-1})$$

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

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

That's my rule!

- Dividing, we get:

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

- Why is this at all helpful?
 - Lets us build one conditional from its reverse
 - Often one conditional is tricky but the other one is simple
 - Foundation of many systems we'll see later (e.g. ASR, MT)
- In the running for most important AI equation!



Independence

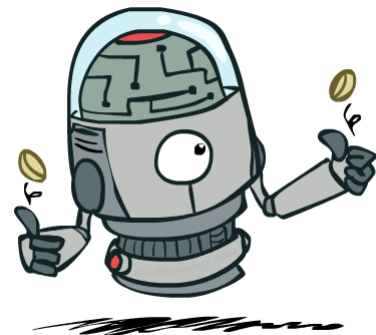
- Two variables are *independent* in a joint distribution if:

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

$$X \perp\!\!\!\perp Y$$

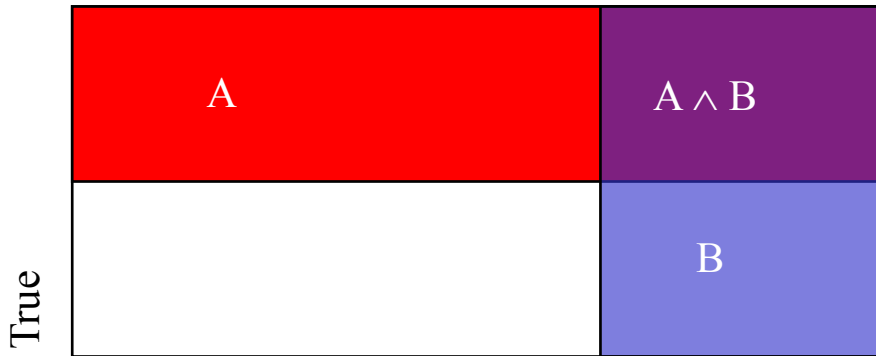
$$\forall x, y P(x, y) = P(x)P(y)$$

- Says the joint distribution *factors* into a product of two simple ones
- Usually variables aren't independent!
- Can use independence as a *modeling assumption*
 - Independence can be a simplifying assumption
 - Empirical* joint distributions: at best "close" to independent
 - What could we assume for {Weather, Traffic, Cavity}?
- Independence is like something from CSPs: what?



Independence

$$P(A \wedge B) = P(A)P(B)$$



© Daniel S. Weld

58

Example: Independence

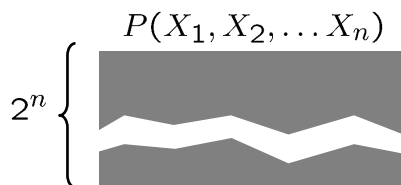
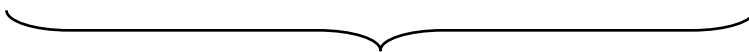
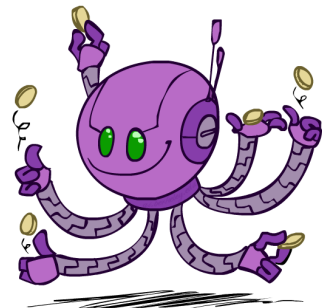
- N fair, independent coin flips:

| $P(X_1)$ | |
|----------|-----|
| H | 0.5 |
| T | 0.5 |

| $P(X_2)$ | |
|----------|-----|
| H | 0.5 |
| T | 0.5 |

...

| $P(X_n)$ | |
|----------|-----|
| H | 0.5 |
| T | 0.5 |



Example: Independence?

$P_1(T, W)$

| T | W | P |
|------|------|-----|
| hot | sun | 0.4 |
| hot | rain | 0.1 |
| cold | sun | 0.2 |
| cold | rain | 0.3 |

$P(T)$

| T | P |
|------|-----|
| hot | 0.5 |
| cold | 0.5 |

$P(W)$

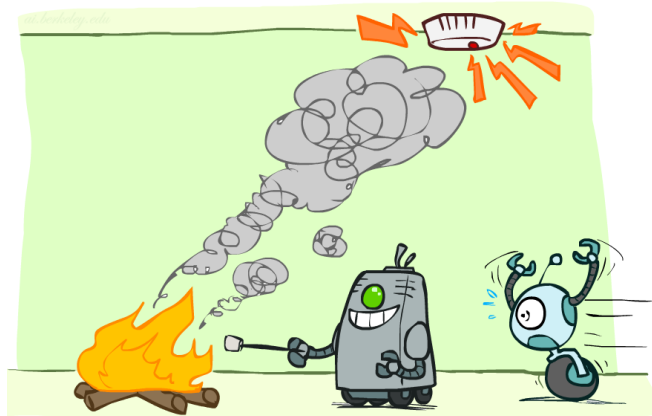
| W | P |
|------|-----|
| sun | 0.6 |
| rain | 0.4 |

$P_2(T, W) = P(T)P(W)$

| T | W | P |
|------|------|-----|
| hot | sun | 0.3 |
| hot | rain | 0.2 |
| cold | sun | 0.3 |
| cold | rain | 0.2 |

\neq

Conditional Independence



Conditional Independence

- Unconditional (absolute) independence very rare
- **Conditional independence** is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z (written $X \perp\!\!\!\perp Y|Z$)

if and only if:

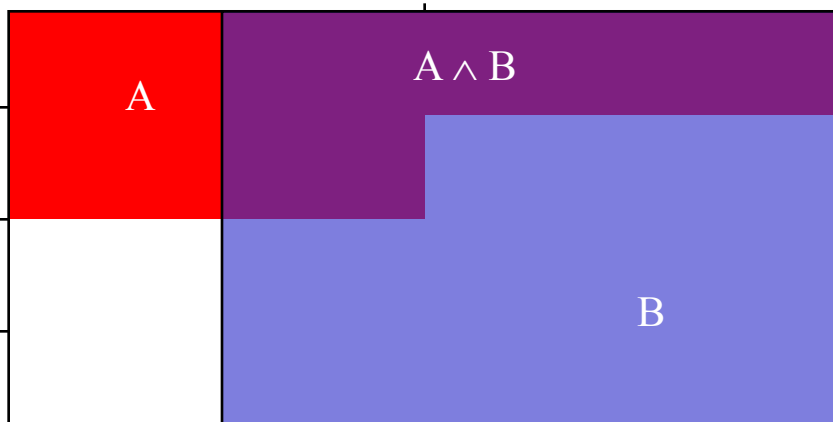
$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

Conditional Independence

Are A & B independent? $P(A|B) < P(A)$



$$P(A) = (.25 + .5) / 2 = .375$$

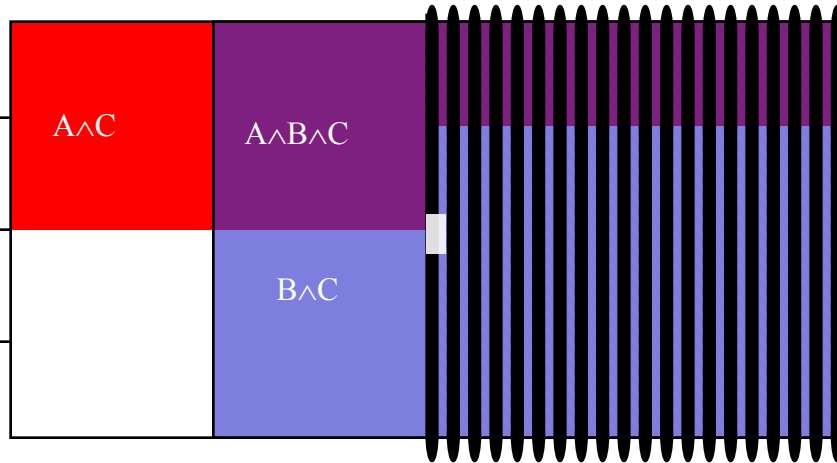
$$P(B) = .75$$

$$P(A|B) = (.25 + .25 + .5) / 3 = .3333$$

A, B Conditionally Independent Given C

$$P(A|B,C) = P(A|C)$$

C = striped



$$P(A|\neg C) = .5$$

$$P(A|B,\neg C) = .5$$

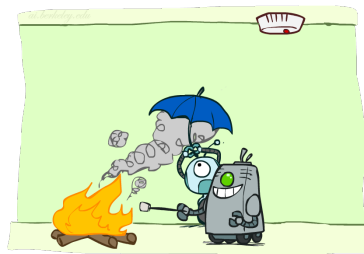
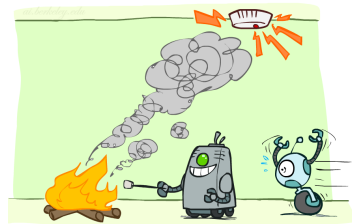
© Daniel S. Weld

64

Conditional Independence

What about this domain:

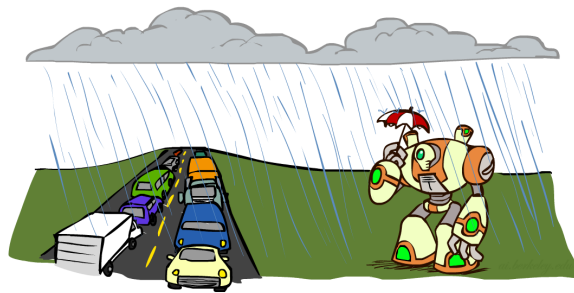
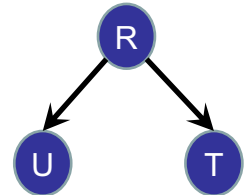
- Fire
- Smoke
- Alarm



Conditional Independence

- What about this domain:

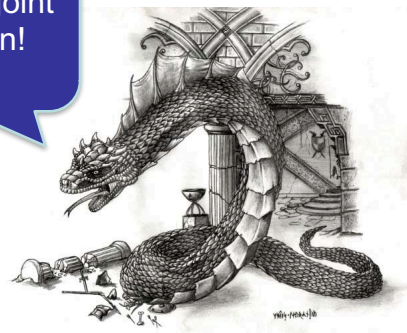
- Traffic
- Umbrella
- Raining



What is Conditional Independence?



I am a BIG joint distribution!



Slay the Basilisk!

Probability Recap

- **Conditional probability** $P(x|y) = \frac{P(x,y)}{P(y)}$
- **Product rule** $P(x,y) = P(x|y)P(y)$
- **Chain rule** $P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \dots$
 $= \prod_{i=1}^n P(X_i|X_1, \dots, X_{i-1})$
- **Bayes rule** $P(x|y) = \frac{P(y|x)}{P(y)}P(x)$
- **X, Y independent if and only if:** $\forall x, y : P(x, y) = P(x)P(y)$
- **X and Y are conditionally independent given Z:** $X \perp\!\!\!\perp Y | Z$
if and only if: $\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$

Markov Models

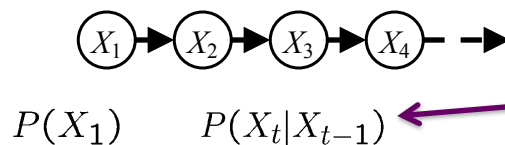


Reasoning over Time or Space

- Often, we want to reason about a sequence of observations
 - Speech recognition
 - Robot localization
 - User attention
 - Medical monitoring
- Need to introduce time (or space) into our models

Markov Models

- Value of X at a given time is called the **state**

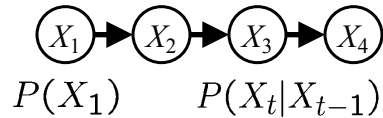


Just a random variable

A conditional probability table
(or schema for said tables)

- Parameters: called **transition probabilities** or dynamics, specify how the state evolves over time (also, initial state probabilities)
- Stationarity assumption: transition **probabilities** the same at all **times**
 - Means $P(X_5 | X_4) = P(X_{12} | X_{11})$ etc.
- Same as MDP transition model, but no choice of action

Joint Distribution of a Markov Model



- Joint distribution:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

- More generally:

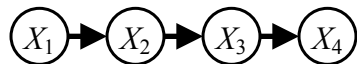
$$P(X_1, X_2, \dots, X_T) = P(X_1)P(X_2|X_1)P(X_3|X_2) \dots P(X_T|X_{T-1})$$

$$= P(X_1) \prod_{t=2}^T P(X_t|X_{t-1})$$

- Questions to be resolved:

- Does this indeed define a joint distribution?
- Can every joint distribution be factored this way, or are we making some assumptions about the joint distribution by using this factorization?

Chain Rule and Markov Models



- From the chain rule, **every** joint distribution over X_1, X_2, X_3, X_4 can be written as:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, \underline{X_2})P(X_4|X_1, \underline{X_2}, \underline{X_3})$$

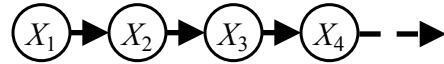
- And, if we assume that

$$X_3 \perp\!\!\!\perp X_1 \mid X_2 \quad \text{and} \quad X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$$

This formula simplifies to

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

Chain Rule and Markov Models



- From the chain rule, every joint distribution over X_1, X_2, \dots, X_T can be written as:

$$P(X_1, X_2, \dots, X_T) = P(X_1) \prod_{t=2}^T P(X_t | X_1, X_2, \dots, X_{t-1})$$

- So, if we assume that for all t :

$$X_t \perp\!\!\!\perp X_1, \dots, X_{t-2} \mid X_{t-1}$$

We get

$$P(X_1, X_2, \dots, X_T) = P(X_1) \prod_{t=2}^T P(X_t | X_{t-1})$$