CS 188: Artificial Intelligence

Probability

Instructors: Stuart Russell and Peyrin Kao

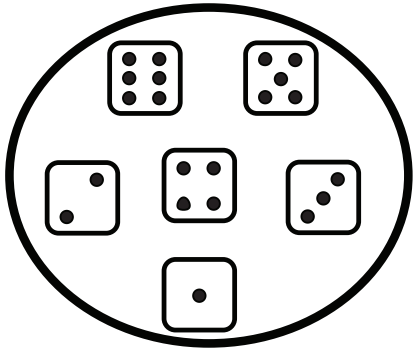
University of California, Berkeley

Uncertainty

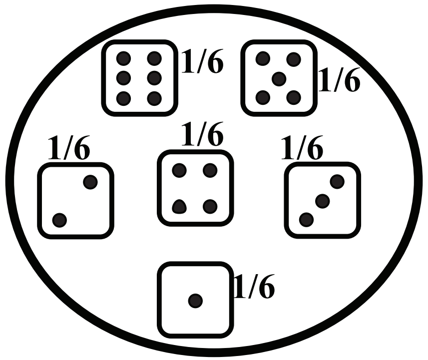
- The real world is rife with uncertainty!
  - E.g., if I leave for SFO 60 minutes before my flight, will I be there in time?
- Problems:
  - partial observability (road state, other drivers' plans, etc.)
  - noisy sensors (radio traffic reports, Google maps)
  - immense complexity of modelling and predicting traffic, security line, etc.
  - lack of knowledge of world dynamics (will tire burst? need COVID test?)
- Probabilistic assertions summarize effects of ***ignorance*** and ***laziness***
- Combine probability theory + utility theory -> decision theory
  - ***Maximize expected utility*** : $a^* = \mathrm{argmax}_a \sum_s P(s \mid a)\, U(s)$

Basic laws of probability (discrete)

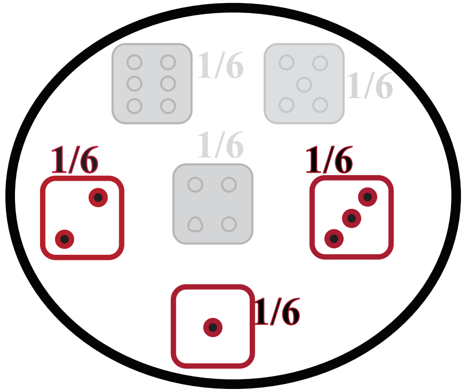
- Begin with a set Ω of possible worlds
  - E.g., 6 possible rolls of a die, $\{1, 2, 3, 4, 5, 6\}$

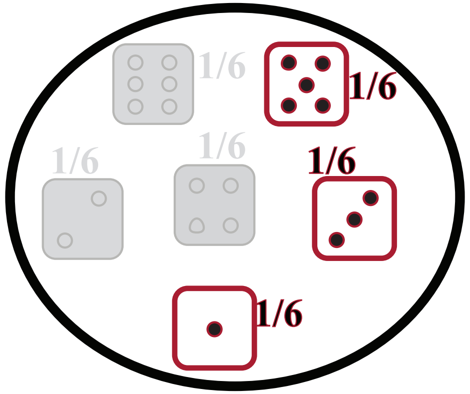


- A ***probability model*** assigns a number $P(\omega)$ to each world ω
  - E.g., $P(1) = P(2) = P(3) = P(5) = P(5) = P(6) = 1/6$.
- These numbers must satisfy
  - $0 \leq P(\omega) \leq 1$
  - $\sum_{\omega \in \Omega} P(\omega) = 1$



Basic laws contd.

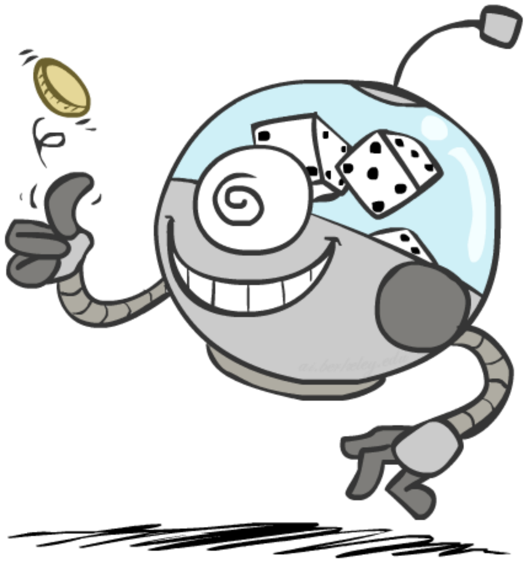




- An ***event*** is any subset of Ω
  - E.g., "roll < 4" is the set {1,2,3}
  - E.g., "roll is odd" is the set {1,3,5}
- The probability of an event is the ***sum*** of probabilities over its worlds
  - $P(A) = \sum_{\omega \in A} P(\omega)$
  - E.g., $P(roll < 4) = P(1) + P(2) + P(3) = 1/2$
- De Finetti (1931): anyone who bets according to probabilities that violate these laws can be forced to lose money on every set of bets

Random Variables

- A random variable (usually denoted by a capital letter) is some aspect of the world about which we may be uncertain
- Formally a ***deterministic function*** of ω
- The ***range*** of a random variable is the set of possible values
  - *Odd* = Is the dice roll an odd number? $\rightarrow$ {true, false}
    - e.g. *Odd*(1)=true, *Odd*(6) = false
    - often write the event *Odd*=true as *odd*, *Odd*=false as $\neg$*odd*
  - *T* = Is it hot or cold? $\rightarrow$ {hot, cold}
  - *D* = How long will it take to get to the airport? $\rightarrow$ $[0, \infty)$
  - L_{Ghost} = Where is the ghost? $\rightarrow$ {(0,0), (0,1), …}
- The ***probability distribution*** of a random variable *X* gives the probability for each value *x* in its range (probability of the event *X*=*x*)
  - $P(X=x) = \sum_{\{\omega:\, X(\omega)=x\}} P(\omega)$
  - $P(x)$ for short (when unambiguous)
  - $P(X)$ refers to the entire distribution (think of it as a vector or table)

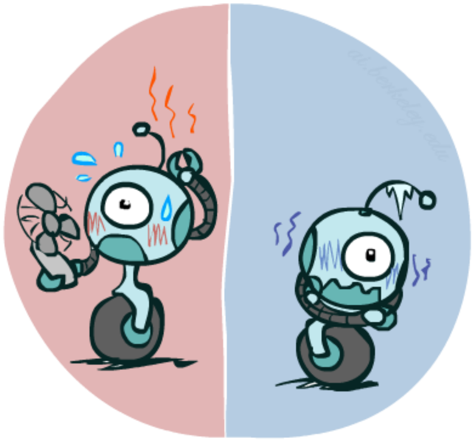


Probability Distributions

- Associate a probability with each value; sums to 1
  - Temperature:

$P(T)$

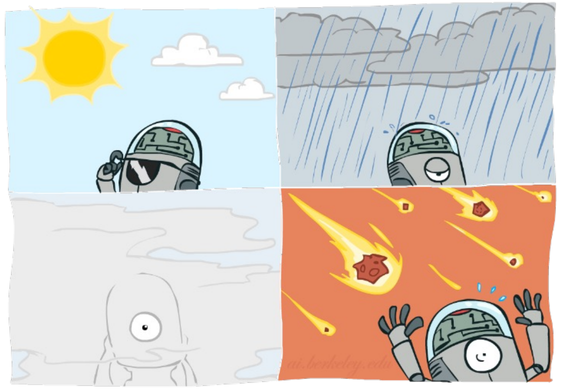
T	P
hot	0.5
cold	0.5



  - Weather:

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0



  - ***Joint distribution***

$P(T,W)$

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

Making possible worlds

- In many cases we
  - begin with random variables and their domains
  - construct possible worlds as assignments of values to all variables
- E.g., two dice rolls $Roll_1$ and $Roll_2$
  - How many possible worlds?
  - What are their probabilities?
- Size of distribution for n variables with range size d?

$$d^n$$

- For all but the smallest distributions, cannot write out by hand!

Probabilities of events

- Recall that the probability of an event is the sum of probabilities of its worlds:
  - $P(A) = \sum_{\omega \in A} P(\omega)$
- So, given a joint distribution over all variables, can compute any event probability!
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR not foggy?

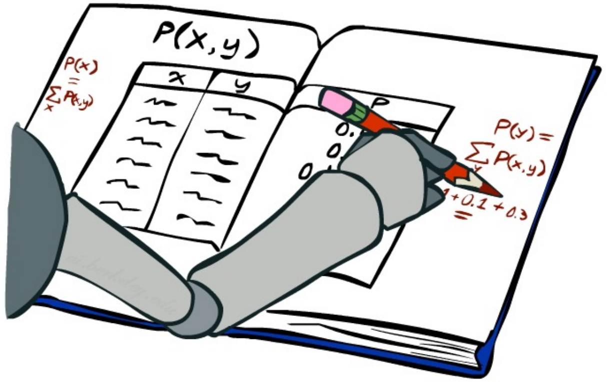
- ***Joint distribution***

$P(T,W)$

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- ***Marginalization*** (***summing out***): Collapse a dimension by adding



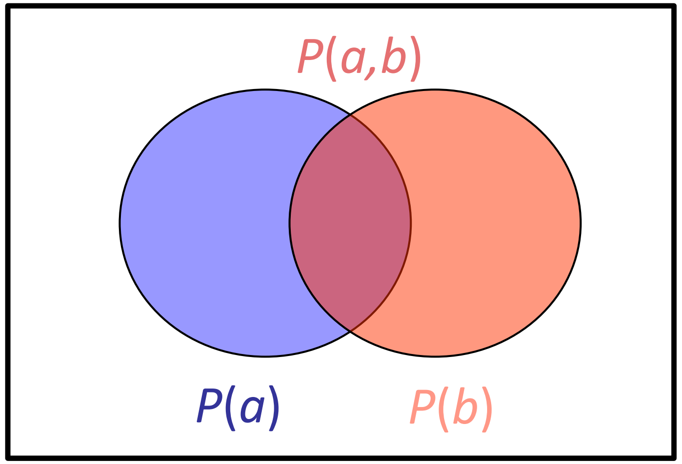
$$P(X{=}x) = \sum_y P(X{=}x, Y{=}y)$$

		Temperature		
		hot	cold	$P(W)$
Weather	sun	0.45	0.15	0.60
	rain	0.02	0.08	0.10
	fog	0.03	0.27	0.30
	meteor	0.00	0.00	0.00
	$P(T)$	0.50	0.50	

Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a \mid b) = \frac{P(a, b)}{P(b)}$$



$P(T,W)$

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

$$P(W=s \mid T=c) = \frac{P(W=s, T=c)}{P(T=c)} = 0.15/0.50 = 0.3$$

$= P(W=s, T=c) + P(W=r, T=c) + P(W=f, T=c) + P(W=m, T=c)$
$= 0.15 + 0.08 + 0.27 + 0.00 = 0.50$

Conditional Distributions

- Distributions for one set of variables given another set

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

$P(W \mid T=h)$

hot
0.90
0.04
0.06
0.00

$P(W \mid T=c)$

cold
0.30
0.16
0.54
0.00

$P(W \mid T)$

hot	cold
0.90	0.30
0.04	0.16
0.06	0.54
0.00	0.00

Normalizing a distribution

- (Dictionary) To bring or restore to a $\boxed{\text{normal condition}}$

All entries sum to ONE

- Procedure:
  - Multiply each entry by α = 1/(sum over all entries)

$P(W,T)$

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

$P(W,T=c)$

0.15
0.08
0.27
0.00

Normalize

$\alpha = 1/0.50 = 2$

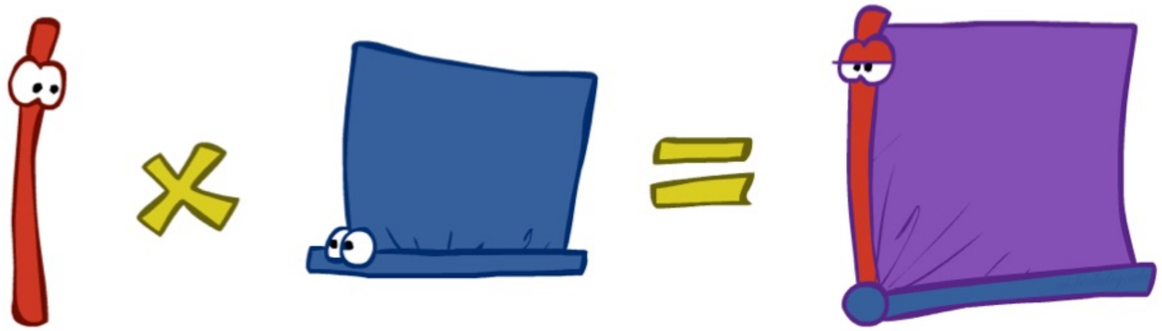
$P(W \mid T=c) = P(W,T=c)/P(T=c)$
$= \alpha\, P(W,T=c)$

0.30
0.16
0.54
0.00

The Product Rule

- Sometimes have conditional distributions but want the joint

$$P(a \mid b)\, P(b) = P(a, b) \quad \Longleftrightarrow \quad P(a \mid b) = \frac{P(a, b)}{P(b)}$$



The Product Rule: Example

$$P(W \mid T)\, P(T) = P(W, T)$$

$P(W \mid T)$

	hot	cold
	0.90	0.30
	0.04	0.16
	0.06	0.54
	0.00	0.00

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W, T)$

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

The Chain Rule

- A joint distribution can be written as a product of conditional distributions by repeated application of the product rule:
- $P(x_1, x_2, x_3) = P(x_3 \mid x_1, x_2)\, P(x_1, x_2) = P(x_3 \mid x_1, x_2)\, P(x_2 \mid x_1)\, P(x_1)$

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

Probabilistic Inference

- Probabilistic inference: compute a desired probability from a probability model
  - Typically for a ***query variable*** given ***evidence***
  - E.g., P(airport on time | no accidents) = 0.90
  - These represent the agent's *beliefs* given the evidence

- Probabilities change with new evidence:
  - P(airport on time | no accidents, 5 a.m.) = 0.95
  - P(airport on time | no accidents, 5 a.m., raining) = 0.80
  - Observing new evidence causes ***<u>beliefs to be updated</u>***

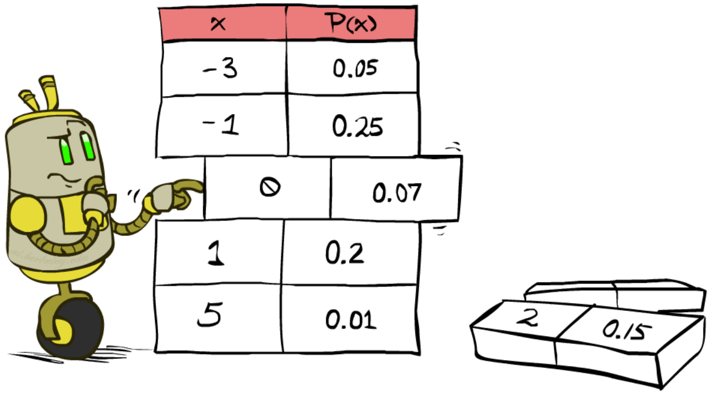
Inference by Enumeration

- Probability model $P(X_1, \ldots, X_n)$ is given
- Partition the variables $X_1, \ldots, X_n$ into sets as follows:
  - Evidence variables: $\boldsymbol{E} = \boldsymbol{e}$
  - Query variables: $\boldsymbol{Q}$
  - Hidden variables: $\boldsymbol{H}$

- We want:

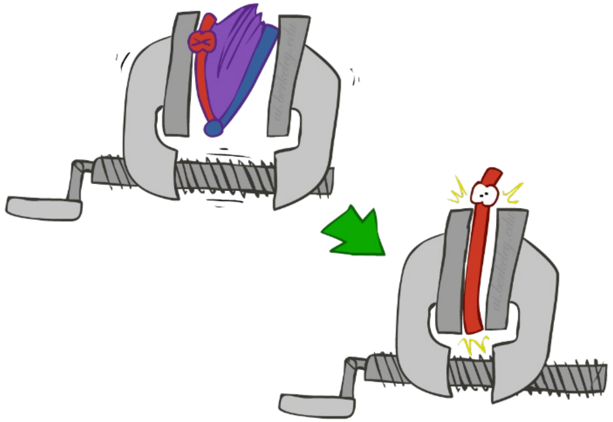
$$P(\boldsymbol{Q} \mid \boldsymbol{e})$$

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out $\boldsymbol{H}$ from model to get joint of query and evidence

$$P(\boldsymbol{Q}, \boldsymbol{e}) = \sum_h \underbrace{P(\boldsymbol{Q}, \boldsymbol{h}, \boldsymbol{e})}_{X_1, \ldots, X_n}$$



- Step 3: Normalize

$$P(\boldsymbol{Q} \mid \boldsymbol{e}) = \alpha\, P(\boldsymbol{Q}, \boldsymbol{e})$$

Inference by Enumeration

- P(W)?

Season	Temp	Weather	P
summer	hot	sun	0.35
summer	hot	rain	0.01
summer	hot	fog	0.01
summer	hot	meteor	0.00
summer	cold	sun	0.10
summer	cold	rain	0.05
summer	cold	fog	0.09
summer	cold	meteor	0.00
winter	hot	sun	0.10
winter	hot	rain	0.01
winter	hot	fog	0.02
winter	hot	meteor	0.00
winter	cold	sun	0.15
winter	cold	rain	0.20
winter	cold	fog	0.18
winter	cold	meteor	0.00

Inference by Enumeration

- P(W)?

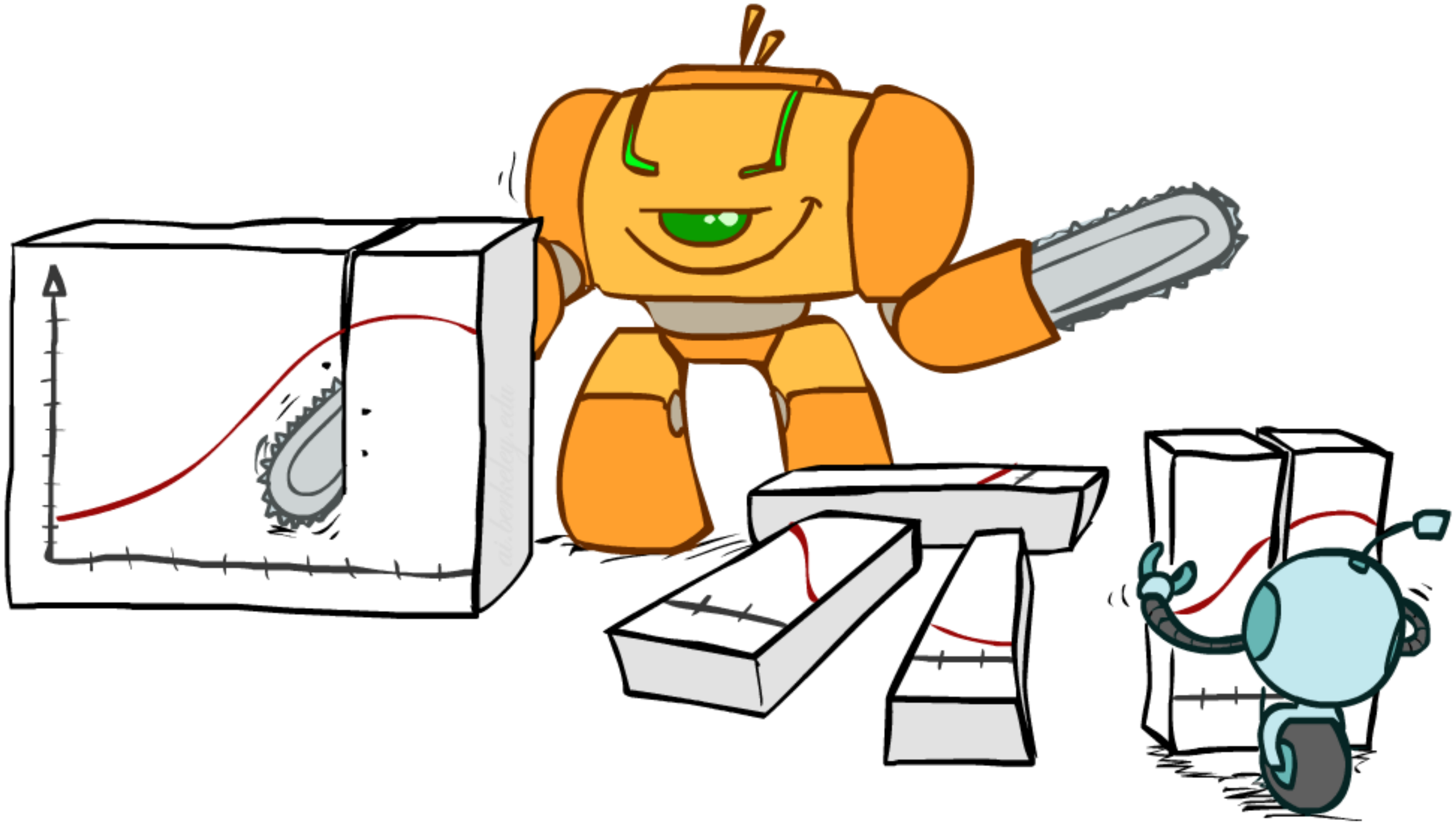
- P(W | winter)?

Season	Temp	Weather	P
summer	hot	sun	0.35
summer	hot	rain	0.01
summer	hot	fog	0.01
summer	hot	meteor	0.00
summer	cold	sun	0.10
summer	cold	rain	0.05
summer	cold	fog	0.09
summer	cold	meteor	0.00
winter	hot	sun	0.10
winter	hot	rain	0.01
winter	hot	fog	0.02
winter	hot	meteor	0.00
winter	cold	sun	0.15
winter	cold	rain	0.20
winter	cold	fog	0.18
winter	cold	meteor	0.00

Inference by Enumeration

- Obvious problems:
  - Worst-case time complexity $O(d^n)$ (exponential in #hidden variables)
  - Space complexity $O(d^n)$ to store the joint distribution
  - $O(d^n)$ data points to estimate the entries in the joint distribution

Bayes Rule



Bayes' Rule

- Write the product rule both ways:

$$P(a \mid b)\, P(b) = P(a, b) = P(b \mid a)\, P(a)$$

- Dividing left and right expressions, we get:

$$P(a \mid b) = \frac{P(b \mid a)\, P(a)}{P(b)}$$

- 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
  - Describes an "update" step from prior $P(a)$ to posterior $P(a \mid b)$
    - Hence provides a simple, formal theory of learning

That's my rule!

Independence

- Two variables X and Y are (absolutely) ***independent*** if

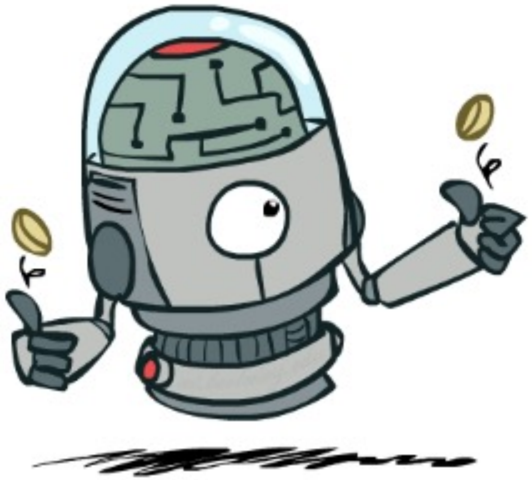
  $$\forall x,y \qquad P(x, y) = P(x)\, P(y)$$

  - I.e., the joint distribution ***factors*** into a product of two simpler distributions

- Equivalently, via the product rule $P(x,y) = P(x|y)P(y)$,

  $$P(x \mid y) = P(x) \quad \text{or} \quad P(y \mid x) = P(y)$$

- Example: two dice rolls $Roll_1$ and $Roll_2$
  - $P(Roll_1{=}5, Roll_2{=}3) \; = \; P(Roll_1{=}5)\, P(Roll_2{=}3) \; = \; 1/6 \times 1/6 \; = \; 1/36$
  - $P(Roll_2{=}3 \mid Roll_1{=}5) \; = \; P(Roll_2{=}3)$



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

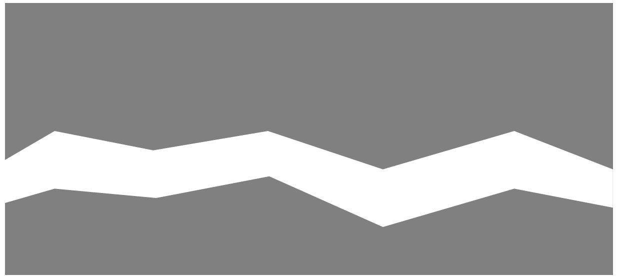
. . .

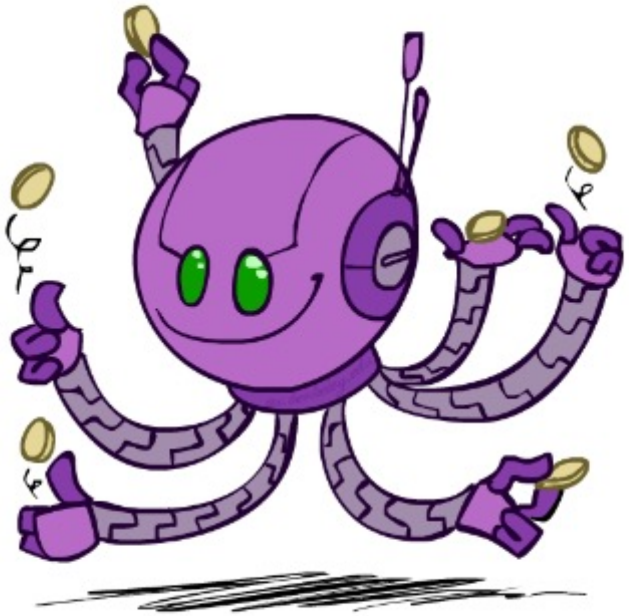
$P(X_n)$

H	0.5
T	0.5

$P(X_1,X_2,\ldots,X_n)$

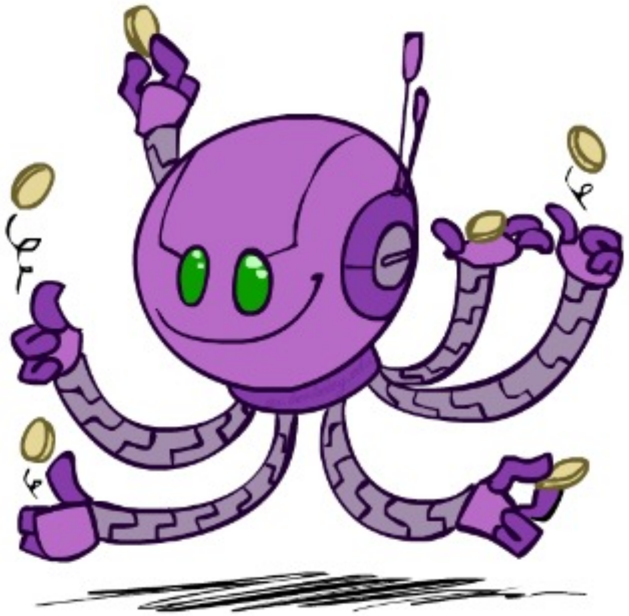
2^n



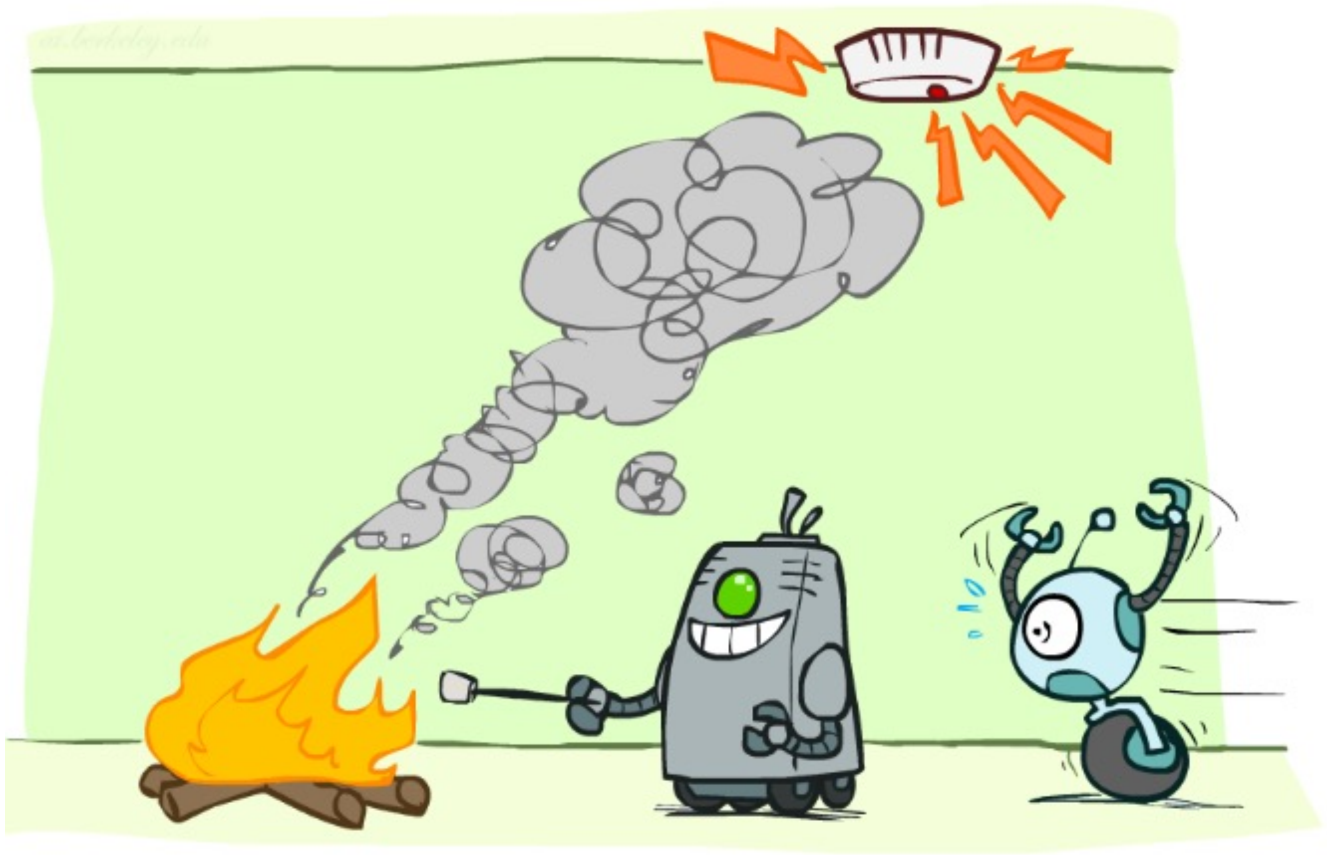


Independence, contd.

- Independence is incredibly powerful
  - Exponential reduction in representation size
- Independence is extremely rare!
- ***Conditional*** independence is ubiquitous!!

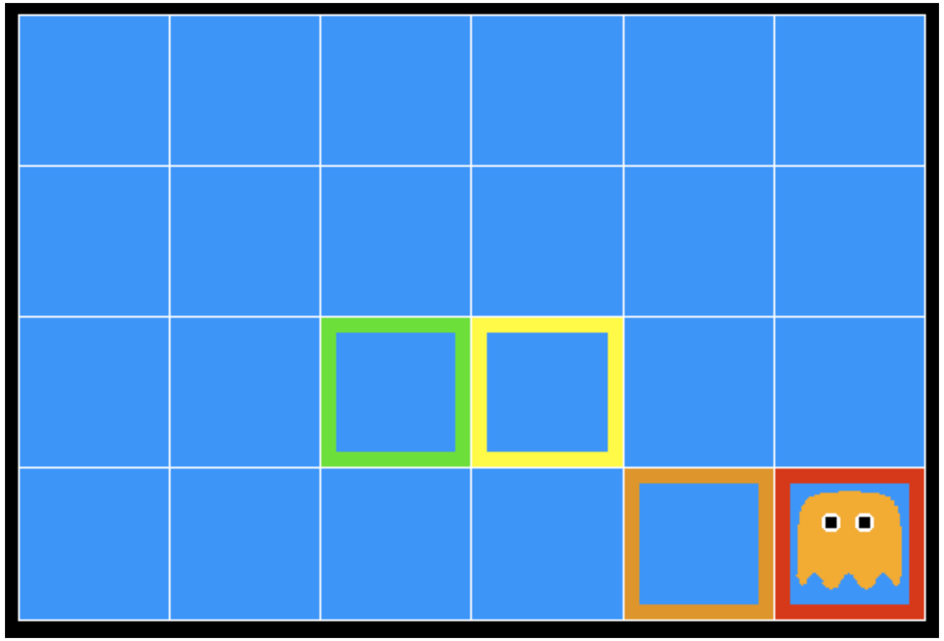


Conditional Independence



Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
  - On the ghost: usually red
  - 1 or 2 away: mostly orange
  - 3 or 4 away: typically yellow
  - 5+ away: often green
- Click on squares until confident of location, then "***bust***"



Video of Demo Ghostbusters with Probability

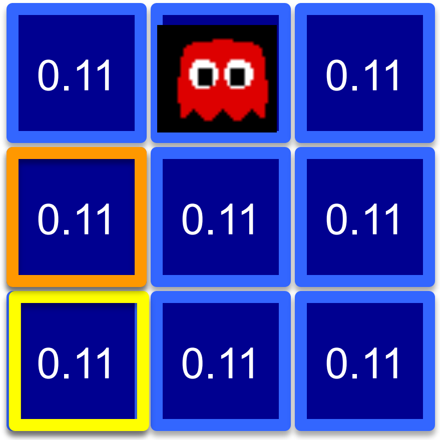
Ghostbusters model

- Variables and ranges:
  - G (ghost location) in $\{(1,1),\ldots,(3,3)\}$
  - $C_{x,y}$ (color measured at square x,y) in {red,orange,yellow,green}

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

- Ghostbuster physics:
  - ***Uniform prior distribution*** over ghost location: $P(G)$
  - ***Sensor model***: $P(C_{x,y} \mid G)$ (depends only on distance to G)
    - E.g. $P(C_{1,1} = \text{yellow} \mid G = (1,1)) = 0.1$

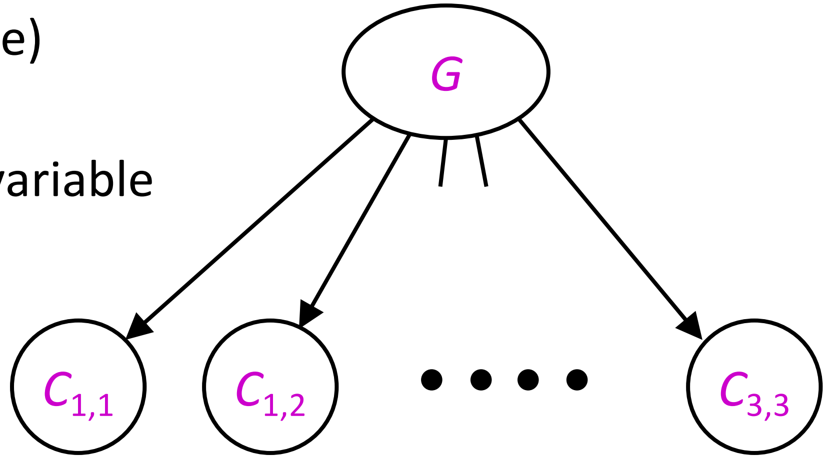
Ghostbusters model, contd.



- $P(G, C_{1,1}, \ldots C_{3,3})$ has $9 \times 4^9 = 2{,}359{,}296$ entries!!!
- Ghostbuster independence:
  - Are $C_{1,1}$ and $C_{1,2}$ independent?
    - E.g., does $P(C_{1,1} = \text{yellow}) = P(C_{1,1} = \text{yellow} \mid C_{1,2} = \text{orange})$?
- Ghostbuster physics again:
  - $P(C_{x,y} \mid G)$ ***depends only on distance to G***
    - So $P(C_{1,1} = \text{yellow} \mid \underline{G = (2,3)}) = P(C_{1,1} = \text{yellow} \mid \underline{G = (2,3)}, C_{1,2} = \text{orange})$
    - I.e., $C_{1,1}$ is ***conditionally independent*** of $C_{1,2}$ ***given*** G

Ghostbusters model, contd.

- Apply the chain rule to decompose the joint probability model:
- $P(G, C_{1,1}, \ldots C_{3,3}) = P(G)\, P(C_{1,1} \mid G)\, P(C_{1,2} \mid G, C_{1,1})\, P(C_{1,3} \mid G, C_{1,1}, C_{1,2}) \ldots P(C_{3,3} \mid G, C_{1,1}, \ldots, C_{3,2})$
- Now simplify using conditional independence:
- $P(G, C_{1,1}, \ldots C_{3,3}) = P(G)\, P(C_{1,1} \mid G)\, P(C_{1,2} \mid G)\, P(C_{1,3} \mid G) \ldots P(C_{3,3} \mid G)$
- I.e., conditional independence properties of ghostbuster physics simplify the probability model from ***exponential*** to ***quadratic*** in the number of squares
- This is called a ***Naïve Bayes*** model:
  - One discrete query variable (often called the ***class*** or ***category*** variable)
  - All other variables are (potentially) evidence variables
  - Evidence variables are all conditionally independent given the query variable



Conditional Independence

- ***Conditional independence*** is our most basic and robust form of knowledge about uncertain environments.

- X is conditionally independent of Y given Z if and only if:

$$\forall x,y,z \qquad P(x \mid y, z) = P(x \mid z)$$

or, equivalently, if and only if

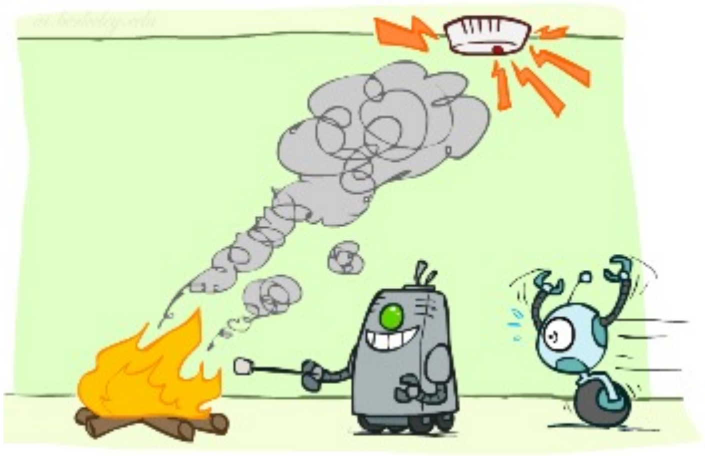
$$\forall x,y,z \qquad P(x, y \mid z) = P(x \mid z)\, P(y \mid z)$$

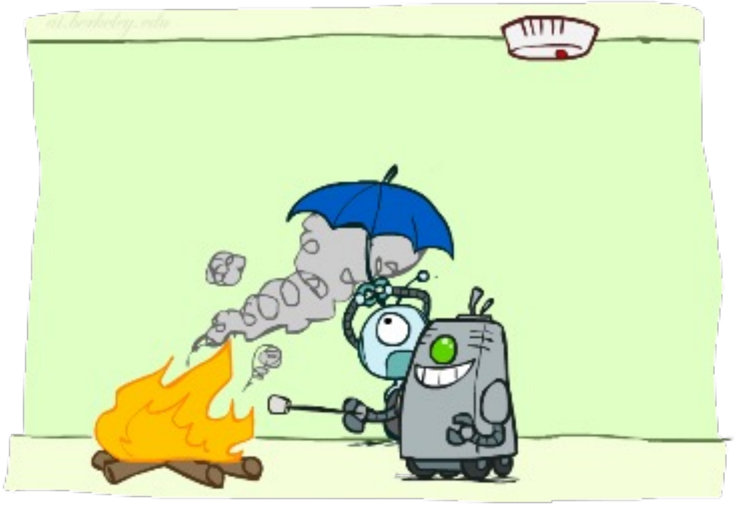
Conditional Independence

- What about this domain:
  - Traffic
  - Umbrella
  - Raining

Conditional Independence

- What about this domain:
  - Fire
  - Smoke
  - Alarm





Next time

- Bayes nets
- Elementary inference in Bayes nets