CS 687
Jana Kosecka

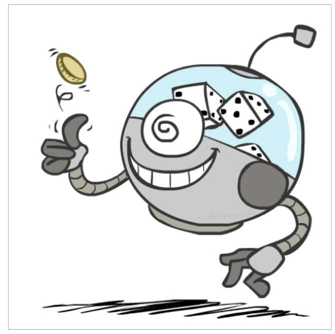
Uncertainty, Bayesian Networks
Chapter 13, Russell and Norvig
Chapter 14, 14.1-14.3

Outline

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

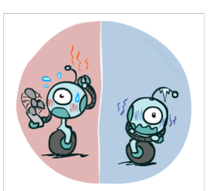
Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - T = Is it hot or cold?
  - D = How long will it take to drive to work?
  - L = Where is the ghost?
- We denote random variables with capital letters
- Random variables have domains
  - R in {true, false} (often write as {+r, -r})
  - T in {hot, cold}
  - D in [0, ∞)
  - L in possible locations, maybe {(0,0), (0,1), …}



Probability Distributions

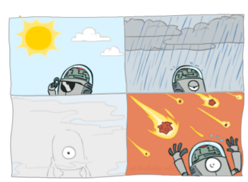
- Associate a probability with each value
  - Temperature:



$P(T)$

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

Probability Distributions

- Unobserved random variables have distributions

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

- A distribution is a TABLE of probabilities of values

- A probability (lower case value) is a single number

$$P(W = rain) = 0.1$$

- Must have: $\forall x \;\; P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Shorthand notation:

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$
$$\dots$$

OK *if* all domain entries are unique

Joint Distributions

- A *joint distribution* over a set of random variables $X_1, X_2, \dots X_n$ specifies a real number 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: $P(x_1, x_2, \dots x_n) \geq 0$

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

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

$P(T, W)$

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

Probabilistic Models

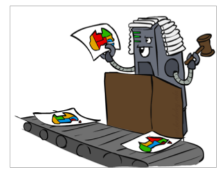
- A probabilistic model is a joint distribution over a set of random variables
- Probabilistic models:
  - (Random) variables with domains
  - Assignments are called *outcomes*
  - Joint distributions: say whether assignments (outcomes) are likely
  - *Normalized:* sum to 1.0
  - Ideally: only certain variables directly interact
- Constraint satisfaction problems:
  - Variables with domains
  - Constraints: state whether assignments are possible
  - Ideally: only certain variables directly interact

Distribution over T,W

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

Constraint over T,W

T	W	P
hot	sun	T
hot	rain	F
cold	sun	F
cold	rain	T



Events

- An *event* is a set E of outcomes

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?
- Typically, the events we care about are *partial assignments*, like P(T=hot)

$P(T, W)$

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

Quiz: Events

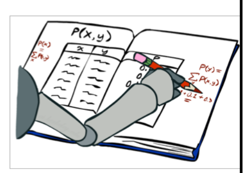
- P(+x, +y) ?
- P(+x) ?
- P(-y OR +x) ?

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

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

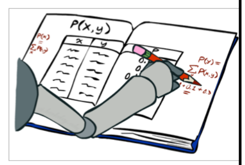
$$P(s) = \sum_t P(t, s)$$

$P(W)$

W	P
sun	0.6
rain	0.4

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

Quiz: Marginal Distributions



$P(X,Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

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

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

$P(X)$

X	P
+x	
-x	

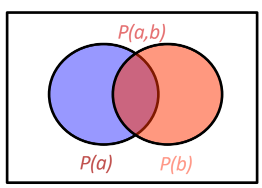
$P(Y)$

Y	P
+y	
-y	

Conditional Probabilities

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

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



$P(T,W)$

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

$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W=s,T=c) + P(W=r,T=c)$$
$$= 0.2 + 0.3 \ = 0.5$$

Quiz: Conditional Probabilities

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

- P(+x | +y) ?
- P(-x | +y) ?
- P(-y | +x) ?

Conditional Distributions

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

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

Joint Distribution

$P(T, W)$

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

Normalization Trick

$P(T, W)$

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

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)} = \frac{0.2}{0.2 + 0.3} = 0.4$$

$$P(W = r|T = c) = \frac{P(W = r, T = c)}{P(T = c)} = \frac{P(W = r, T = c)}{P(W = s, T = c) + P(W = r, T = c)} = \frac{0.3}{0.2 + 0.3} = 0.6$$

$P(W|T = c)$

W	P
sun	0.4
rain	0.6

Normalization Trick

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)} = \frac{0.2}{0.2 + 0.3} = 0.4$$

$P(T, W)$

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

$$P(W = r|T = c) = \frac{P(W = r, T = c)}{P(T = c)} = \frac{P(W = r, T = c)}{P(W = s, T = c) + P(W = r, T = c)} = \frac{0.3}{0.2 + 0.3} = 0.6$$

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)

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

Quiz: Normalization Trick

- P(X | Y=-y) ?

$P(X,Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

SELECT the joint probabilities matching the evidence →

NORMALIZE the selection (make it sum to one) →

To Normalize

- (Dictionary) To bring or restore to a normal condition

All entries sum to ONE

- Procedure:
  - Step 1: Compute Z = sum over all entries
  - Step 2: Divide every entry by Z

- Example 1

W	P
sun	0.2
rain	0.3

Normalize

Z = 0.5

W	P
sun	0.4
rain	0.6

- Example 2

T	W	P
hot	sun	20
hot	rain	5
cold	sun	10
cold	rain	15

Normalize

Z = 50

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

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

- We generally compute conditional probabilities
  - P(on time | no reported accidents) = 0.90
  - These represent the agent's *beliefs* given the evidence

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

Inference by Enumeration

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

$\left.\right\} X_1, X_2, \ldots X_n$ *All variables*

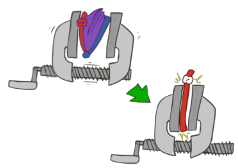
- We want:

** Works fine with multiple query variables, too*

$$P(Q|e_1 \ldots e_k)$$

- Step 1: Select the entries consistent with the evidence

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



$$P(Q, e_1 \ldots e_k) = \sum_{h_1 \ldots h_r} \underbrace{P(Q, h_1 \ldots h_r, e_1 \ldots e_k)}_{X_1, X_2, \ldots X_n}$$

- Step 3: Normalize

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

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

$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

Inference by Enumeration

- P(W)?
- P(W | winter)?
- P(W | winter, hot)?

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

Inference by Enumeration

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

The Product Rule

- Sometimes have conditional distributions but want the joint

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

The Product Rule

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

- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$P(D, W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	. . .

The Chain Rule

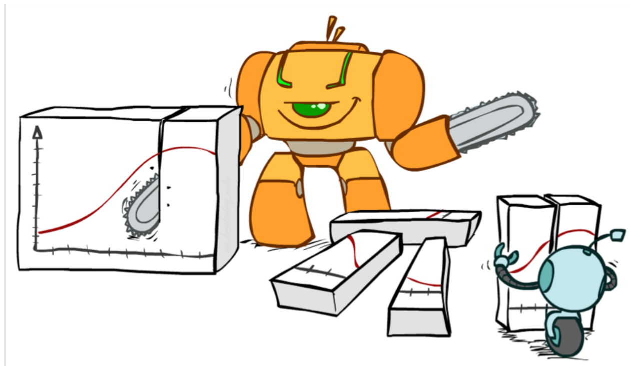
- More generally, can always write any joint distribution as an incremental product of conditional distributions

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

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

- Why is this always true?

Bayes Rule



Bayes' Rule

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

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

- Dividing, we get:

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

- 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!

That's my rule!

Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Example:
  - M: meningitis, S: stiff neck

$$\left.\begin{array}{r} P(+m) = 0.0001 \\ P(+s|+m) = 0.8 \\ P(+s|-m) = 0.01 \end{array}\right\} \text{Example givens}$$

  - Note: posterior probability of meningitis still very small
  - Note: you should still get stiff necks checked out! Why?

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m)+P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$

Quiz: Bayes' Rule

- Given:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

- What is P(W | dry) ?

Probabilistic Models

- Models describe how (a portion of) the world works
- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - "All models are wrong; but some are useful."
    – George E. P. Box

- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)
  - Example: value of information

Independence

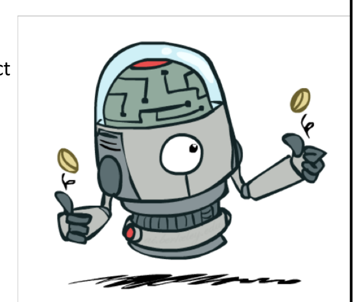
- Two variables are *independent* if:

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

  - This says that their joint distribution *factors* into a product two simpler distributions
  - Another form:

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

  - We write: $X \perp\!\!\!\perp Y$



- Independence is a simplifying *modeling assumption*
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity, Toothache}?

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

$P(W)$

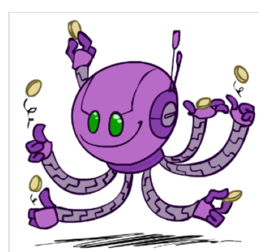
W	P
sun	0.6
rain	0.4

$P_2(T, W)$

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

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

$\cdots$

$P(X_n)$

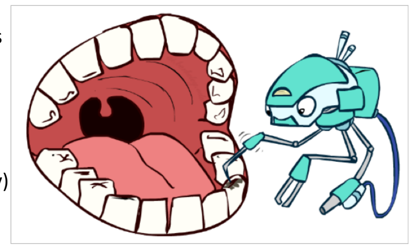
H	0.5
T	0.5

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

2^n

Conditional Independence

- P(Toothache, Cavity, Catch)
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)
- The same independence holds if I don't have a cavity:
  - P(+catch | +toothache, -cavity) = P(+catch| -cavity)
- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)



- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily

Conditional Independence

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

  if and only if:

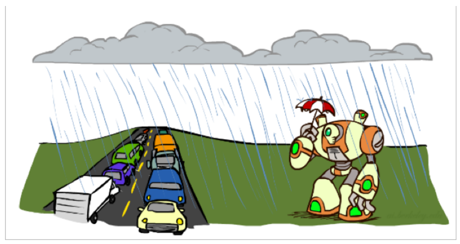
  $$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

  or, equivalently, if and only if

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

Conditional Independence

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



Conditional Independence and the Chain Rule

- Chain rule: $P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots$

- Trivial decomposition:

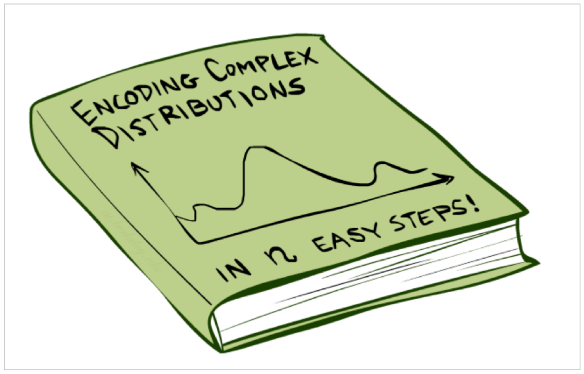
$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) = P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain}, \text{Traffic})$$

- With assumption of conditional independence:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) = P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$

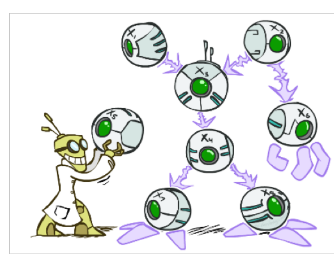
- Bayes' nets / graphical models help us express conditional independence assumptions

Bayes' Nets: Big Picture

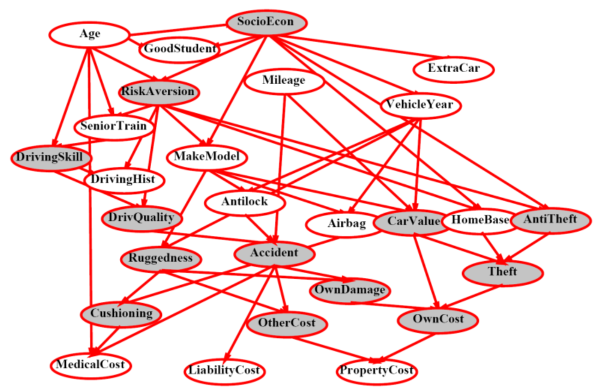


Bayes' Nets: Big Picture

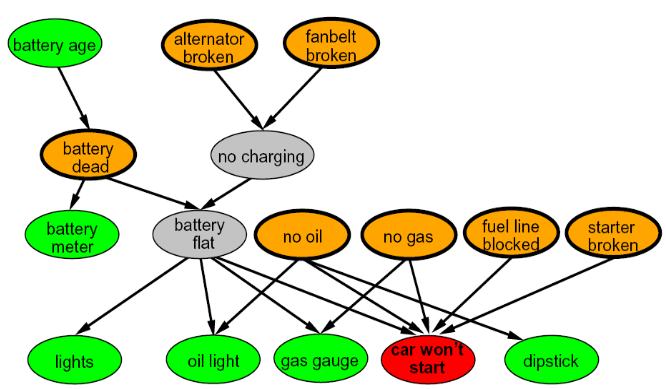
- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time
- Bayes' nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions
  - For about 10 min, we'll be vague about how these interactions are specified



Example Bayes' Net: Insurance

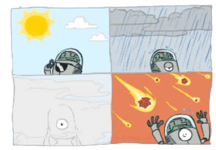


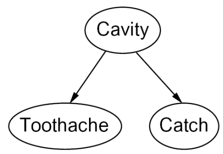
Example Bayes' Net: Car

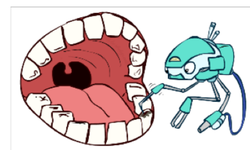


Graphical Model Notation

- Nodes: variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)
- Arcs: interactions
  - Similar to CSP constraints
  - Indicate "direct influence" between variables
  - Formally: encode conditional independence (more later)
- For now: imagine that arrows mean direct causation (in general, they don't!)

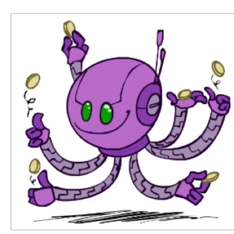






Example: Coin Flips

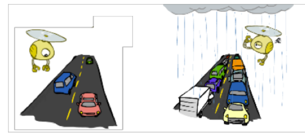
- N independent coin flips



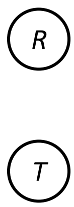
- No interactions between variables: absolute independence

Example: Traffic

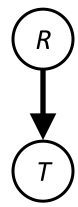
- Variables:
  - R: It rains
  - T: There is traffic



- Model 1: independence



- Model 2: rain causes traffic



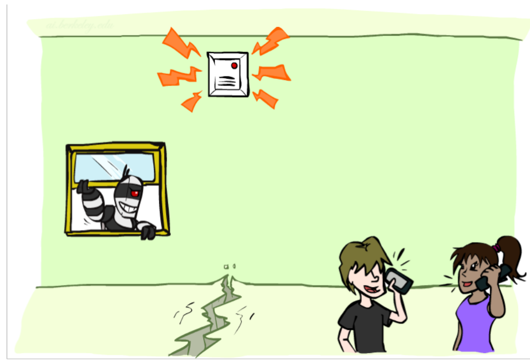
- Why is an agent using model 2 better?

Example: Traffic II

- Let's build a causal graphical model!
- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame
  - C: Cavity

Example: Alarm Network

- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!

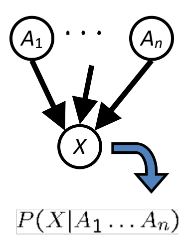


Bayes' Net Semantics

- 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

    $P(X|a_1 \ldots a_n)$

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



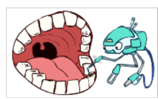
A Bayes net = Topology (graph) + Local Conditional Probabilities

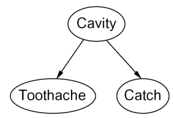
Probabilities in BNs

- Bayes' nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

  - Example:





$P(+cavity,\ +catch,\ -toothache)$

Probabilities in BNs

- Why are we guaranteed that setting

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

results in a proper joint distribution?

- Chain rule (valid for all distributions): $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | x_1 \ldots x_{i-1})$
- Assume conditional independences: $P(x_i | x_1, \ldots x_{i-1}) = P(x_i | parents(X_i))$

  → Consequence: $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$

- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies

Example: Coin Flips

X_1 X_2 $\cdots$ X_n

$P(X_1)$

h	0.5
t	0.5

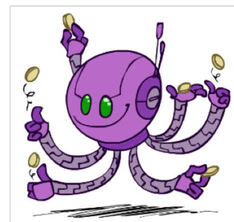
$P(X_2)$

h	0.5
t	0.5

$\cdots$

$P(X_n)$

h	0.5
t	0.5



$P(h, h, t, h) =$

Only distributions whose variables are absolutely independent can be represented by a Bayes' net with no arcs.

Example: Traffic

$R \rightarrow T$

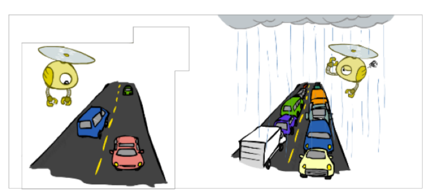
$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

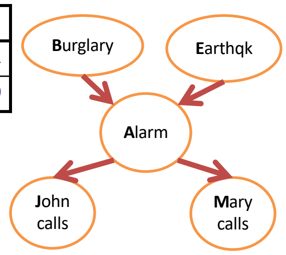
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(+r, -t) =$

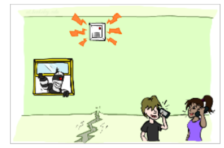


Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



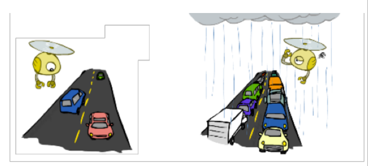
B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

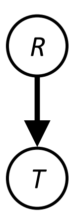
A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

Example: Traffic

- Causal direction





$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

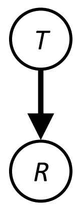
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Example: Reverse Traffic

- Reverse causality?



$P(T)$

+t	9/16
-t	7/16

$P(R|T)$

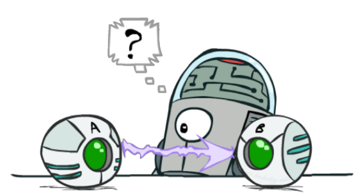
+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7

$P(T,R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Causality?

- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation
- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence



$$P(x_i|x_1,\ldots x_{i-1}) = P(x_i|\text{parents}(X_i))$$

Bayes' Nets

- So far: how a Bayes' net encodes a joint distribution
- Next: how to answer queries about that distribution
  - Today:
    - First assembled BNs using an intuitive notion of conditional independence as causality
    - Then saw that key property is conditional independence
  - Main goal: answer queries about conditional independence and influence
- After that: how to answer numerical queries (inference)