# CS188: Artificial Intelligence.

Probability

Inference.

Begin: Bayes Networks

# Probability Distributions

Unobserved random variables have distributions:

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

| W | P |
|---|---|
| sun | 0.6 |
| rain | 0.1 |
| fog | 0.3 |
| meteor | 0.0 |

Shorthand notation:

$P(hot) = P(T = hot)$
$P(cold) = P(T = cold)$
$P(rain) = P(W = rain)$
...

If domains don't overlap.

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

$P(W = rain) = 0.1.$

A distribution is a TABLE of probabilities of values:

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

# Joint Distributions

Set of random variables: $X_1, \ldots, X_n$
Joint Distribtuion:

$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$
or $P(x_1, x_2, \ldots, x_n)$.

$P(T, W)$

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

Same table:

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

- Must obey:

$P(x_1, x_2, \ldots, x_n) \geq 0$

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

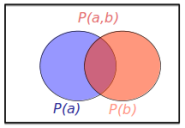
Size of distribution if n variables with domain sizes d? $d^n$

- For all but the smallest distributions, impractical to write out!

# Conditional Probabilities

A simple relation between joint and conditional probabilities

- In fact, this is taken as the definition of a conditional probability




The probability of event *a* given event *b*.

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

Probability of a given b.

Natural? Yes!

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

$P(W = s|T = c) = \frac{P(w=s, T=c)}{P(T=c)}$.

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

$P(W = s|T = c) = \frac{P(w=s, T=c)}{P(T=c)} = \frac{.2}{.5} = 2/5.$

# Conditional Distributions

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

Joint Distribution

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

Conditional Distributions

$P(W|T = hot)$

| W | P |
|---|---|
| sun | 0.8 |
| cold | 0.2 |

$P(W|T = cold)$

| W | P |
|---|---|
| sun | 0.4 |
| cold | 0.6 |

# Normalization Trick

| 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

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

- NORMALIZE the selection (make it sum to one)

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

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

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

## Quiz: Normalization Trick

| 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

| X | Y | P |
|---|---|---|
| -x | +y | 0.4 |
| -x | -y | 0.1 |

- NORMALIZE the selection (make it sum to one)

| X | Y | P |
|---|---|---|
| -x | +y | 0.8 |
| -x | -y | 0.2 |

## To Normalize

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

| 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

| 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
- $P(on\ time|no\ reported\ accidents) = 0.90$
- 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

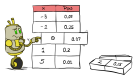
General case:
- Evidence variables: $\bar{E}_1, \ldots, \bar{E}_k = e_1, \ldots, e_k$
- Query* variable: $Q$
- Hidden variables: $H_1, \ldots, H_r$.

We Want:

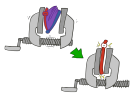
$P(Q|e_1, \ldots, e_k)$.

* Works fine with multiple query variables, too

Step 1:
Entries consistent with evidence



Step 2: Sum out H



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

Step 3: Normalize

$$Z = \sum_q P(q, e_1, \ldots, e_k)$$

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

## Inference by Enumeration

| 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 |

$P(W = sun)$?
Sum out S, T.
Get 0.65

$P(W = sun|winter)$?
Sum out T over S=winter.
Get .25 for W=sun.
Get .25 for W=rain.
Normalize: 1/2

$P(W = sun|winter, hot)$?
Get .10 for W=sun.
Get .05 for W=rain.
Normalize: 2/3.

## 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

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

## The Product Rule

Example:

$P(W)$

| W | 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 | 0.08 |
| dry | sun | 0.72 |
| wet | rain | 0.14 |
| dry | rain | 0.06 |

## The Chain Rule

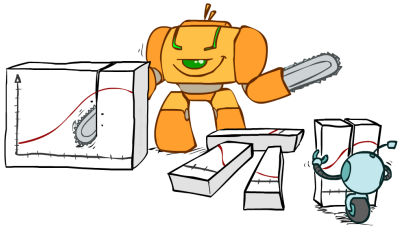
More generally, can always write any joint distribution as an incremental product of conditional distributions

- $P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$
- $P(x_1, \ldots, x_n) = \Pi_i P(x_i|x_1, \ldots, x_{i-1})$

Why is this always true?

Induction.

## 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:

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

Example:

- M: meningitis, S: stiff neck
- Givens.

$P(+m) = 0.0001, P(+s|+m) = 0.8, P(+s|-m) = 0.01$

- Bayes Rule

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m)+P(+s|-m)Pr(-m)}$$
$$= \frac{(0.8)(0.0001)}{(0.8)(0.0001)+(0.01)(0.9999)} = small$$

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

## Quiz: Bayes' Rule

Given:

| R | P |
|---|---|
| sun | 0.8 |
| rain | 0.2 |

Pr(D—W)

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

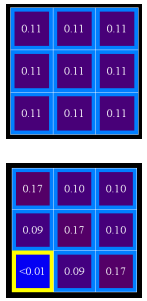
What is $P(sun|dry)$ ?

$= P(dry|sun) \times P(sun)/P(dry) = \frac{.72}{.78} = 12/13$

$P(dry) = Pr(dry, sun) + Pr(dry, rain)$

$= Pr(dry|sun) \times Pr(sun) + Pr(dry|rain)Pr(rain) = 0.72 + .06 = 0.78.$

## Ghostbusters, Revisited




Let's say we have two distributions:

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

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

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

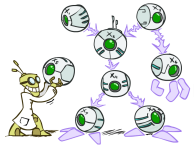
[Demo: Ghostbuster – with probability (L12D2) ]

## Video of Demo Ghostbusters with Probability

## Next up: bayes nets.

## CS 188: Artificial Intelligence

http://bit.ly/cs188bn



Bayes' Nets

## Probabilistic Models

http://bit.ly/cs188bn

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

## Probability Recap

http://bit.ly/cs188bn

Conditional probability: $P(x|y) = \frac{P(x,y)}{P(y)}$.

Product rule: $P(x,y) = P(x|y)P(y)$.

Bayes Rule: $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$.

Chain rule:

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)\ldots$$
$$= \Pi_{i=1}^{n} P(x_i|x_1, \ldots, x_{i-1}).$$

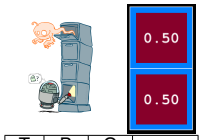
$X, Y$ independent if and only if: $\forall x, y : P(x,y) = P(x)P(y)$.

$X$ and $Y$ are conditionally independent given $Z$ if and only if:

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

## Ghostbusters Chain Rule

http://bit.ly/cs188bn




| T | B | G | P(T,B,G) |
|---|---|---|---|
| +t | +b | +g | 0.16 |
| +t | +b | -g | 0.16 |
| +t | -b | +g | 0.24 |
| +t | -b | -g | 0.04 |
| +t | +b | +g | 0.04 |
| -t | +b | -g | 0.24 |
| -t | -b | +g | 0.06 |
| -t | -b | -g | 0.06 |

Each sensor depends only on where the ghost is.

That means, the two sensors are conditionally independent, given the ghost position.

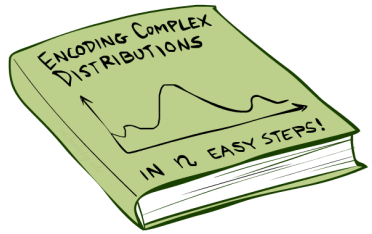
T: Top square is red.
B: Bottom square is red.
G: Ghost is in the top.

Givens:

P( +g ) = 0.5
P( -g ) = 0.5
P( +t | +g ) = 0.8
P( +t | -g ) = 0.4
P( +b | +g ) = 0.4
P( +b | -g ) = 0.8

$P(T,B,G) = P(G)P(T|G)P(B|G)$

## Bayes'Nets: Big Picture




## Bayes' Nets: Big Picture

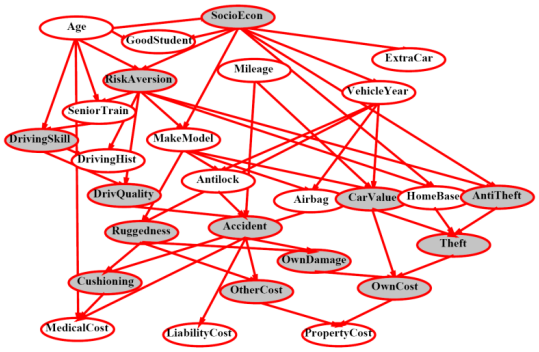
Two problems with using full joint distribution tables as our probabilistic models:

- 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

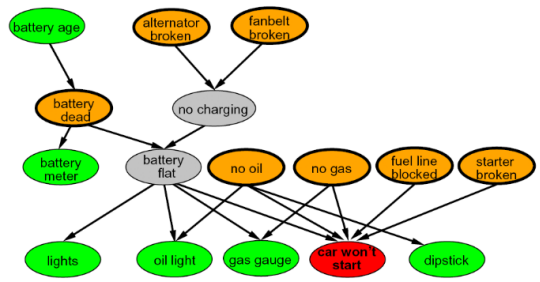
Bayes' nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)

- More properly called graphical models
- We describe how variables locally interact
- Local interactions chain together to give global, indirect interactions
- For now, we'll be vague about how these interactions are specified

## Example Bayes' Net: Insurance




## Example Bayes' Net: Car

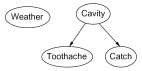


## Graphical Model Notation

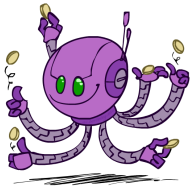
Nodes: variables (with domains)
- Can be assigned (observed) or unassigned (unobserved)

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

$X_1$ $X_2$ $\cdots$ $X_n$

## Example: Traffic

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

Model 1: independence

$R$

$T$

Model 2: rain causes traffic

$R \rightarrow T$

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!

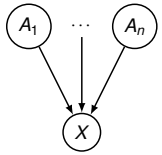
## Bayes' Net Semantics

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

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

## Probabilities in Bayesnets

Bayes' nets implicitly encode joint distributions

- As a product of local conditional distributions

$$P(x_1,\ldots,x_n) = \prod_i Pr(x_i|Parents(X_i))$$

- To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

$$P(x_1,\ldots,x_n) = \prod_i Pr(x_i|Parents(X_i))$$

Example:

$$Pr(G=b|T=r,B=r) = Pr(G=r|T=r)P(S=s|B=r)$$

$$Pr(G=b|T=r,B=r) = Pr(G=r|T=r,B=r)P(S=s|B=r)$$

## Probabilities in BNs

Why are we guaranteed that setting results is joint distribution?

Chain rule (valid for all distributions):

$$P(x_1,\ldots,x_n) = \prod_i P(x_i|x_1,\ldots,x_i)$$

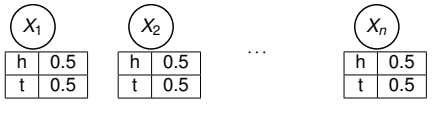
Assume conditional independences:

$$P(x_1,\ldots,x_n) = \prod_i P(x_i|parents(X_i))$$

**Consequence:** Not every BN can represent every joint distribution.

- The topology enforces certain conditional independencies!

## Example: Coin Flips




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

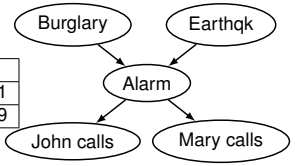
## Example: Traffic


| R | T |
|---|---|
| +r | 1/4 |
| -r | 3/4 |

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

P(+r,-t) = P(+r) × P(-t | +r)
= (1/4) × (1/4) = 1/16

## 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


| R | T |
|---|---|
| +r | 1/4 |
| -r | 3/4 |

P(T|R)

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

P(R,T)

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

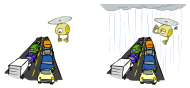
## Example: Reverse Traffic

Reverse causality?


| | |
|---|---|
| +t | 9/16 |
| -t | 7/16 |

P(R|T)

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



P(R,T)

| | | |
|---|---|---|
| +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

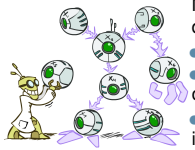
- Sometimes no causal net exists over the domain (especially if variables are missing)
- E.g. consider the variables Traffic and Drips
- End up with arrows that reflect correlation, not causation

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

What do the arrows really mean?

- Topology may happen to encode causal structure
- Topology really encodes conditional independence

## Bayes' Nets



So far: how a Bayes' net encodes a joint distribution

Next: how to answer queries about that distribution

- So far:
- 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)