CSE 573: Artificial Intelligence

Hanna Hajishirzi Bayes Nets

slides adapted from Dan Klein, Pieter Abbeel ai.berkeley.edu And Dan Weld, Luke Zettlemoyer



Our Status in CSE573

- We're done with Search and planning
- We are done with learning to make decisions
- Probabilistic Reasoning
 - Diagnosis
 - Speech recognition
 - Tracking objects
 - Robot mapping
 - Genetics
 - Error correcting codes
 - ... lots more!



Inference in Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
 - On the ghost: red
 - 1 or 2 away: orange
 - 3 or 4 away: yellow
 - 5+ away: green



Sensors are noisy, but we know P(Color | Distance)

P(red 3)	P(orange 3)	P(yellow 3)	P(green 3)
0.05	0.15	0.5	0.3

3

Video of Demo Ghostbuster

7% ghostbusters		Inter	ence	e in G	nost	Duste		
						_	GHOSTS REMAINING: 1 BUSTS REMAINING: 1 SCODE	
							MESSAGES:	
	_			_				
							BUST	
				_			TIME+1	
						6		
								7, 1
				1 Januara	anu indey to a			

Uncertainty

General situation:

- Observed variables (evidence): Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
- Unobserved variables: Agent needs to reason about other aspects (e.g. where an object is or what disease is present)
- Model: Agent knows something about how the known variables relate to the unknown variables
- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge







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 outcome
 - Temperature:

Weather:







P	(W	·)
V		

W	Р
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Probability Distributions

Unobserved random variables have distributions





• A distribution is a TABLE of probabilities of values

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

Shorthand notation:

$$P(hot) = P(T = hot),$$

$$P(cold) = P(T = cold),$$

$$P(rain) = P(W = rain),$$

OK *if* all domain entries are unique

. . .

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

$$\forall x \ P(X = x) \ge 0$$

 $\sum_{x} P(X = x) = 1$

Joint Distributions

 A joint distribution over a set of random variables: 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:

$$P(x_1, x_2, \dots x_n) \ge 0$$

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

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

$$X_1, X_2, \ldots X_n$$

Т	W	Р
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Events

An event is a set E of outcomes

$$P(E) = \sum_{\substack{(x_1...x_n) \in E \\ \text{probability of any event}}} P(x_1...x_n)$$

- 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)



Т	W	Р
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

TWPhotsun0.4hotrain0.1coldsun0.2coldrain0.3

Distribution over T,W



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)

Independence



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?



sun

rain

0.4

$D_{\mathbf{r}}$	(T)	W	
12	\ <i>L</i> ,		

Т	W	Ρ
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

Example: 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 Probabilities

P(a,b)

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

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

$$P(T,W)$$

$$\frac{T \ W \ P}{hot \ sun \ 0.4} \\ hot \ rain \ 0.1 \\ \hline cold \ sun \ 0.2 \\ \hline cold \ rain \ 0.3 \end{bmatrix}$$

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

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

$$= 0.2 + 0.3 = 0.5$$

19

0.4

Quiz: Conditional Probabilities

■ P(+x | +y) ?



Х	Y	Р
+x	+у	0.2
+x	-у	0.3
-X	+у	0.4
-X	-у	0.1

■ P(-x | +y) ?

■ P(-y | +x) ?

Quiz: Conditional Probabilities

■ P(+x | +y) ?

P(X,	Y)
------	----

Х	Y	Р
+x	+у	0.2
+x	-у	0.3
-X	+у	0.4
-X	-y	0.1

.2/.6=1/3

P(-x | +y) ?
.4/.6=2/3

■ P(-y | +x) ?

.3/.5=.6

Conditional Distributions

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

Conditional Distributions



P(W|T)

Joint Distribution

P(T, W)

Т	W	Р
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

The Product Rule

Sometimes have conditional distributions but want the joint

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



m /

The Product Rule

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

• Example:

P(W)

Ρ

0.8

0.2

R

sun

rain

P(D W)		
D	W	Ρ
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

P(D,W)

D	W	Р
wet	sun	
dry	sun	
wet	rain	
dry	rain	

- 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

 $X \bot\!\!\!\!\perp Y | Z$

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

or, equivalently, if and only if $\forall x, y, z : P(x|z, y) = P(x|z)$

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



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





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



Example: 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

R



R

Example: Traffic II

- 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!



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 narents' 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

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(*+cavity*, *+catch*, *-toothache*)

=P(-toothache|+cavity)P(+catch|+cavity)P(+cavity)

Bayes' Net Representation

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
 - A collection of distributions over X, one for each combination of parents' values $P(X|a_1 \dots a_n)$
- 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 and the relevant $P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | parents(X_i))$





Probabilities in BNs



• Why are we guaranteed that setting

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

results in a proper joint distribution?

• Chain rule (valid for all distributions):

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

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

→ Consequence:

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

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

Example: Coin Flips



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

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

Example: Traffic



$$P(+r, -t) = P(+r)P(-t|+r) = \frac{1}{4} \cdot \frac{1}{4}$$





Example: Alarm Network



E	P(E)	
+e	0.002	
-е	0.998	



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

P(M|A)P(J|A)P(A|B,E)

Example: Traffic

Causal direction







P(T,R)

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

Example: Reverse Traffic

Reverse causality?







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



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!



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

$$\begin{array}{c} P(+m) = 0.0001 \\ P(+s|+m) = 0.8 \\ P(+s|-m) = 0.01 \end{array} \end{array} \ \ \begin{array}{c} \mbox{Example} \\ \mbox{givens} \end{array} \ \ \end{array}$$

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$
• Note: posterior probability of meningitis still very small

Note: you should still get stiff necks checked out! Why?

Quiz: Bayes' Rule



P(D W)
5	147

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

What is P(W | dry) ?

Quiz: Bayes' Rule



P(D W)		
D	W	Р
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

What is P(W | dry) ?

 $P(sun|dry) \sim P(dry|sun)P(sun) = .9^*.8 = .72$ $P(rain|dry) \sim P(dry|rain)P(rain) = .3^*.2 = .06$ P(sun|dry)=12/13P(rain|dry)=1/13

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 = yellow | G=(1,1)) = 0.1
- We can calculate the posterior distribution P(G|r) over ghost locations given a reading using Bayes' rule:

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

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11



54

Video of Demo Ghostbusters with Probability

Ghostbusters, Revisited



Uncertainty Summary

Conditional probability

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

- Product rule P(x,y) = P(x|y)P(y)
- Chain rule $P(X_1, X_2, ..., X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)...$ $= \prod_{i=1}^n P(X_i|X_1, ..., X_{i-1})$
- 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: $X \perp \!\!\!\perp Y | Z$ $\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$

BN lecture

Bayes' Net Representation

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
 - A collection of distributions over X, one for each combination of parents' values $P(X|a_1 \dots a_n)$
- 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 and the relevant and the relevant n $P(x_1, x_2, \dots x_n) = \prod_{i=1}^n P(x_i | parents(X_i))$



