Probability 101++

Hal Daumé III

Computer Science
University of Maryland

me@hal3.name

CS 421: Introduction to Artificial Intelligence

27 Mar 2012

Many slides courtesy of
Dan Klein, Stuart Russell,
or Andrew Moore

Announcements

- Today is a brand new day!

Today

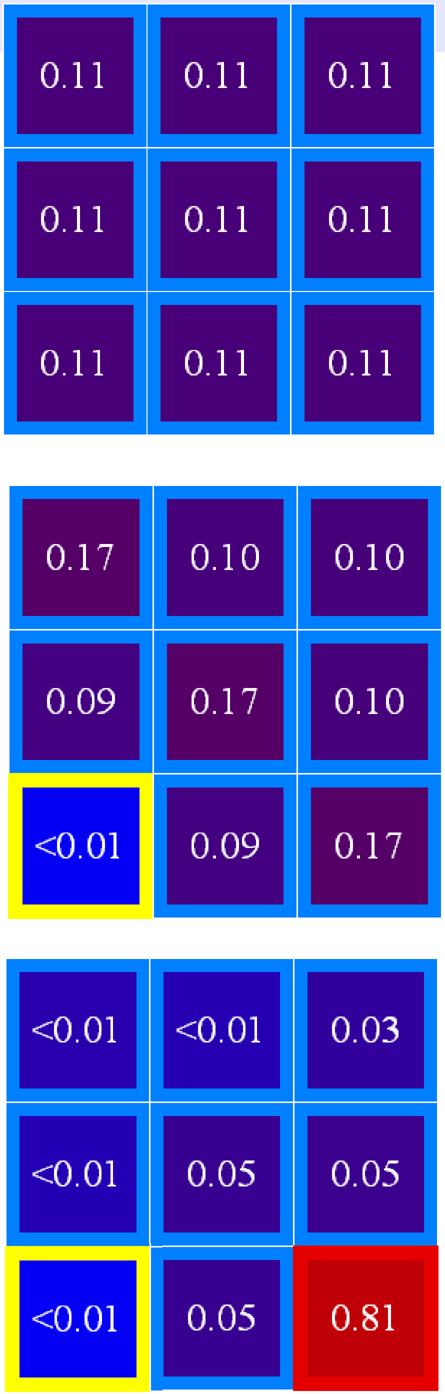
- Probability
  - Random Variables
  - Joint and Conditional Distributions
  - Inference, Bayes' Rule
  - Independence

- You'll need all this stuff for the next few weeks, so make sure you go over it!

Uncertainty

- General situation:
  - Evidence: Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
  - Hidden 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?
  - D = How long will it take to drive to work?
  - L = Where am I?

- We denote random variables with capital letters

- Like in a CSP, each random variable has a domain
  - R in {true, false}   (often write as $\{r, \neg r\}$)
  - D in $[0, \infty)$
  - L in possible locations

Probabilities

- We generally calculate 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 reported accidents, 5 a.m.) = 0.95
  - P(on time | no reported accidents, 5 a.m., raining) = 0.80
  - Observing new evidence causes *beliefs to be updated*

Probabilistic Models

- CSPs:
  - Variables with domains
  - Constraints: state whether assignments are possible
  - Ideally: only certain variables directly interact

- 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

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

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

Joint Distributions

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

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

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

  - Size of distribution if n variables with domain sizes d?
  - Must obey: $0 \leq P(x_1, x_2, \ldots x_n) \leq 1$

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

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

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

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=h)

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

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

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

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

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

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

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

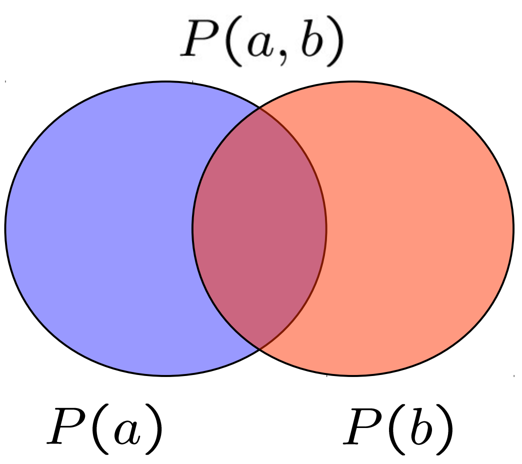
$$P(T, W)$$

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

Conditional Distributions

- 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

$$P(W = r|T = c) = ???$$

Conditional Probabilities

- *Conditional* or *posterior probabilities:*
  - E.g., P(*cavity* | *toothache*) = 0.8
  - Given that *toothache* is all I know…
- Notation for conditional distributions:
  - P(*cavity* | *toothache*) = a single number
  - P(Cavity, Toothache) = 2x2 table summing to 1
  - P(Cavity | Toothache) = Two 2-element vectors, each summing to 1
- If we know more:
  - P(*cavity* | *toothache*, *catch*) = 0.9
  - P(*cavity* | *toothache*, *cavity*) = 1
- Note: the less specific belief remains *valid* after more evidence arrives, but is not always *useful*
- New evidence may be irrelevant, allowing simplification:
  - P(*cavity* | *toothache, traffic*) = P(*cavity* | *toothache*) = 0.8
- This kind of inference, guided by domain knowledge, is crucial

Normalization Trick

- A trick to get a whole conditional distribution at once:
  - Select the joint probabilities matching the evidence
  - Normalize the selection (make it sum to one)

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

Select →

$P(T, r)$

T	P
hot	0.1
cold	0.3

Normalize →

$P(T|r)$

T	P
hot	0.25
cold	0.75

- Why does this work? Because sum of selection is P(evidence)!

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

The Product Rule

- Sometimes have a joint distribution but want a conditional
- Sometimes the reverse

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

- Example:

$P(S)$

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

$\Longleftrightarrow$

$P(D,W)$

D	W	P
wet	sun	0.08
dry	sun	0.72
wet	rain	0.14
dry	rain	0.06

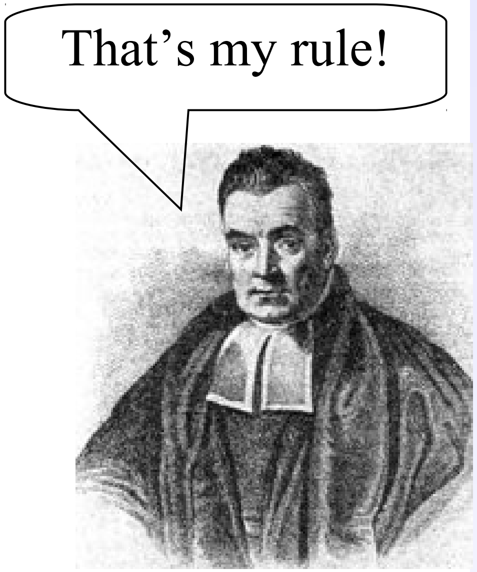
Bayes' Rule

- 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 is meningitis, s is stiff neck

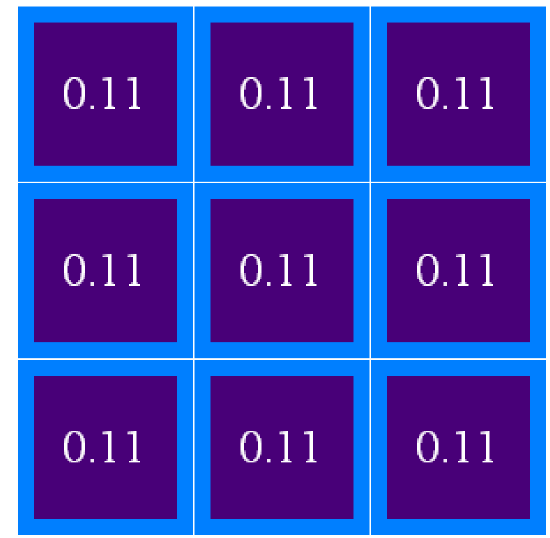
$$\left.\begin{aligned} &P(s|m) = 0.8 \\ &P(m) = 0.0001 \\ &P(s) = 0.1 \end{aligned}\right\} \text{Example givens}$$

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

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

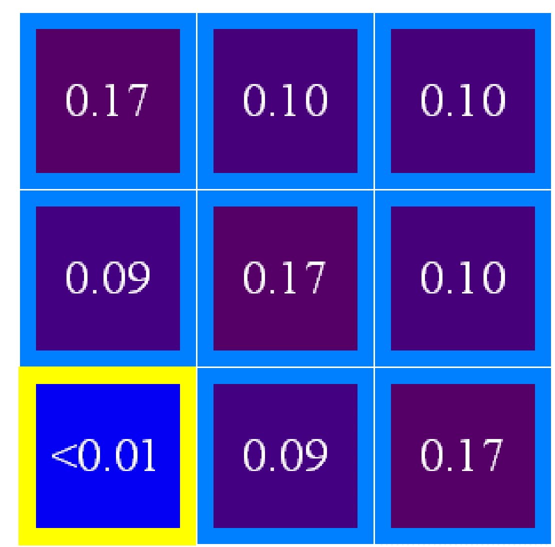
Ghostbusters

- Let’s say we have two distributions:
  - Prior distribution over ghost locations: P(L)
    - Say this is uniform (for now)
  - Sensor reading model: P(R | L)
    - Given by some known black box process
    - E.g. P(R = yellow | L=(1,1)) = 0.1
    - For now, assume the reading is always for the lower left corner



- We can calculate the posterior distribution over ghost locations using Bayes’ rule:

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



Example Problems

- Suppose a murder occurs in a town of population 10,000 (10,001 before the murder). A suspect is brought in and DNA tested. The probability that there is a DNA match give that a person is innocent is 1/100,000; the probability of a match on a guilty person is 1. What is the probability he is guilty given a DNA match?

- Doctors have found that people with Kreuzfeld-Jacob disease (KJ) are almost invariably ate lots of hamburgers, thus p(HamburgerEater|KJ) = 0.9. KJ is a rare disease: about 1 in 100,000 people get it. Eating hamburgers is widespread: p(HamburgerEater) = 0.5. What is the probability that a regular hamburger eater will have KJ disease?

Inference by Enumeration

- P(sun)?
- P(sun | winter)?
- P(sun | winter, warm)?

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

- General case:
  - Evidence variables: $(E_1 \dots E_k) = (e_1 \dots e_k)$
  - Query variables: $Y_1 \dots Y_m$
  - Hidden variables: $H_1 \dots H_r$

  $\Big\}\ X_1, X_2, \dots X_n$ — *All variables*

- We want: $P(Y_1 \dots Y_m | e_1 \dots e_k)$

- First, select the entries consistent with the evidence
- Second, sum out H:

$$P(Y_1 \dots Y_m, e_1 \dots e_k) = \sum_{h_1 \dots h_r} \underbrace{P(Y_1 \dots Y_m, h_1 \dots h_r, e_1 \dots e_k)}_{X_1, X_2, \dots X_n}$$

- Finally, normalize the remaining entries to conditionalize

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

21

Independence

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

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

  - This says that their joint distribution *factors* into a product two simpler distributions
  - Usually variable 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?

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

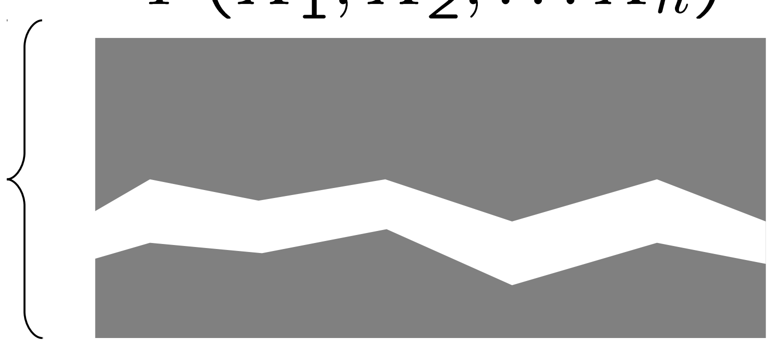
$\dots$

$P(X_n)$

H	0.5
T	0.5

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

2^n



Example: Independence?

- Arbitrary joint distributions can be poorly modeled by independent factors

$P(T)$

T	P
warm	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

$P(T,W)$

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

$P(T)P(W)$

T	S	P
warm	sun	0.3
warm	rain	0.2
cold	sun	0.3
cold	rain	0.2

Conditional Independence

- Warning: we're going to use domain knowledge, not laws of probability, here to simplify a model!
- 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)

Conditional Independence

- Unconditional (absolute) independence is very rare (why?)
- Conditional independence is our most basic and robust form of knowledge about uncertain environments:

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

- What about this domain:
  - Traffic
  - Umbrella
  - Raining
- What about fire, smoke, alarm?

The Chain Rule II

- Can *always* write any joint distribution as an incremental product of conditional distributions

$$P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)\ldots$$

$$P(X_1, X_2, \ldots X_n) = \prod_i P(X_i|X_1 \ldots X_{i-1})$$

- Why?
- This actually claims nothing…
- What are the sizes of the tables we supply?

The Chain Rule III

- 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 conditional independence:

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

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

- Graphical models (next class) will help us work with and think about conditional independence

Birthday Paradox

- What's the probability that no two people in this room have the same birthday?