CS 188: Artificial Intelligence
Spring 2006

Lecture 8: Probability
2/9/2006

Dan Klein – UC Berkeley
Many slides from either Stuart Russell or Andrew Moore

Today

- Uncertainty
- Probability Basics
  - Joint and Condition Distributions
  - Models and Independence
  - Bayes Rule
  - Estimation
- Utility Basics
  - Value Functions
  - Expectations

Uncertainty

- Let action A_t = leave for airport t minutes before flight
- Will A_t get me there on time?
- Problems:
  - partial observability (road state, other drivers' plans, etc.)
  - noisy sensors (KCBS traffic reports)
  - uncertainty in action outcomes (flat tire, etc.)
  - immense complexity of modeling and predicting traffic
- A purely logical approach either
  - Risks falsehood: "A_{25} will get me there on time" or
  - Leads to conclusions that are too weak for decision making:
    - "A_{25} will get me there on time if there's no accident on the bridge, and it doesn't rain, and my tires remain intact, etc., etc."
- A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport…

Probabilities

- Probabilistic approach
  - Given the available evidence, A_{25} will get me there on time with probability 0.04
  - $P(A_{25}$ | no reported accidents) = 0.04
- Probabilities change with new evidence:
  - $P(A_{25}$ | no reported accidents, 5 a.m.) = 0.15
  - $P(A_{25}$ | no reported accidents, 5 a.m., raining) = 0.08
  - i.e., observing evidence causes *beliefs to be updated*

Probabilistic Models

- CSPs:
  - Variables with domains
  - Constraints: map from assignments to true/false
  - Ideally: only certain variables directly interact

A	B	P
warm	sun	T
warm	rain	F
cold	sun	F
cold	rain	T

- Probabilistic models:
  - (Random) variables with domains
  - Joint distributions: map from assignments (or outcomes) to positive numbers
  - *Normalized:* sum to 1.0
  - Ideally: only certain variables are directly correlated

A	B	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

What Are Probabilities?

- Objectivist / frequentist answer:
  - Averages over repeated *experiments*
  - E.g. empirically estimating P(rain) from historical observation
  - Assertion about how future experiments will go (in the limit)
  - New evidence changes the *reference class*
  - Makes one think of *inherently random* events, like rolling dice
- Subjectivist / Bayesian answer:
  - Degrees of belief about unobserved variables
  - E.g. an agent's belief that it's raining, given the temperature
  - Often *estimate* probabilities from past experience
  - New evidence *updates beliefs*
- Unobserved variables still have fixed assignments (we just don't know what they are)

Probabilities Everywhere?

- Not just for games of chance!
  - I'm snuffling: am I sick?
  - Email contains "FREE!": is it spam?
  - Tooth hurts: have cavity?
  - Safe to cross street?
  - 60 min enough to get to the airport?
  - Robot rotated wheel three times, how far did it advance?
- Why can a random variable have uncertainty?
  - Inherently random process (dice, etc)
  - Insufficient or weak evidence
  - Unmodeled variables
  - Ignorance of underlying processes
  - The world's just noisy!
- Compare to *fuzzy logic*, which has *degrees of truth*, or soft assignments

Distributions on Random Vars

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ is a map from assignments (or *outcome*, or *atomic event*) to reals:

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

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

  - Size of distribution if n variables with domain sizes d?
  - Must obey:

$$0 \leq P(x_1, x_2, \ldots x_n) \leq 1$$

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

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

Examples

- An event is a set E of assignments (or 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 warm AND sunny?
- Probability that it's warm?
- Probability that it's warm OR sunny?

T	S	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

Marginalization

- Marginalization (or summing out) is *projecting* a joint distribution to a sub-distribution over subset of variables

$$P(X_1, X_3) = \sum_{x_2} P(X_1, x_2, X_3)$$

$P(T, S)$

T	S	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

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

$P(T)$

T	P
warm	0.5
cold	0.5

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

$P(S)$

S	P
sun	0.6
rain	0.4

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) = 4-element vector 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, sanctioned by domain knowledge, is crucial

Conditioning

- Conditioning is fixing some variables and renormalizing over the rest:

$$P(X_1, X_3 | x_2) = \frac{P(X_1, x_2, X_3)}{\sum_{x_1, x_3} P(x_1, x_2, x_3)}$$

$$P(X_1, X_3 | x_2) = \frac{P(X_1, x_2, X_3)}{P(x_2)}$$

T	S	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

Select

$P(T, r)$

T	P
warm	0.1
cold	0.3

Normalize

$P(T|r)$

T	P
warm	0.25
cold	0.75

Inference by Enumeration

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

S	T	R	P
summer	warm	sun	0.30
summer	warm	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	warm	sun	0.10
winter	warm	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: $H_1 \dots H_r$

  $X_1, X_2, \dots X_n$ *All variables*
- We want: $P(Y_1 \dots Y_m | e_1 \dots e_k)$
- The required summation of joint entries is done by summing out H:

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

- Then renormalizing

$$P(Y_1 \dots Y_m | e_1 \dots e_k) = \frac{P(Y_1 \dots Y_m, e_1 \dots e_k)}{P(e_1 \dots e_k)}$$

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

The Chain Rule I

- Sometimes joint P(X,Y) is easy to get
- Sometimes easier to get conditional P(X|Y)

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

- Example: P(Sun,Dry)?

$P(S)$

R	P
sun	0.8
rain	0.2

$P(D|S)$

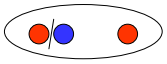
D	S	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$P(D,S)$

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

Lewis Carroll's Sack Problem

- Sack contains a red or blue ball, 50/50
- We add a red ball
- If we draw a red ball, what's the chance of drawing a second red ball?
- Variables:
  - F={r,b} is the original ball
  - D={r,b} is the ball we draw
- Query: P(F=r|D=r)



$P(F)$

F	P
r	0.5
b	0.5

$P(D|F)$

F	D	P
r	r	1.0
r	b	0.0
b	r	0.5
b	b	0.5

$P(F,D)$

F	D	P
r	r	
r	b	
b	r	
b	b	

Lewis Carroll's Sack Problem

- Now we have P(F,D)
- Want P(F|D=r)

F	D	P
r	r	0.5
r	b	0.0
b	r	0.25
b	b	0.25

Independence

- Two variables are *independent* if:

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

  - This says that their joint distribution *factors* into a product two simpler distributions
- Independence is a *modeling assumption*
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Sun, Dry, Toothache, Cavity}?
- How many parameters in the full joint model?
- How many parameters in the independent model?
- Independence is like something from CSPs: what?

Example: Independence

- N fair, independent coins:

$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

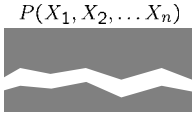
$\cdots$

$P(X_n)$

H	0.5
T	0.5

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

2^n



Example: Independence?

- Arbitrary joint distributions can be (poorly) modeled by independent factors

$P(T)$

T	P
warm	0.5
cold	0.5

$P(S)$

S	P
sun	0.6
rain	0.4

$P(T, S)$

T	S	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(T)P(S)$

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

Conditional Independence

- P(Toothache,Cavity,Catch) has 2^3 = 8 entries (7 independent entries)
- 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 haven't got 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 independence is very rare (two reasons: 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 factor any joint distribution as a product of incremental conditional distributions

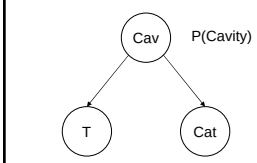
$$P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)\ldots$$

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

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

The Chain Rule III

- Write out full joint distribution using chain rule:
  - P(Toothache, Catch, Cavity)
    = P(Toothache | Catch, Cavity) P(Catch, Cavity)
    = P(Toothache | Catch, Cavity) P(Catch | Cavity) P(Cavity)
    = P(Toothache | Cavity) P(Catch | Cavity) P(Cavity)



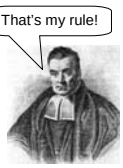
Graphical model notation:

- *Each variable is a node*
- *The parents of a node are the other variables which the decomposed joint conditions on*
- *MUCH more on this to come!*

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 invert a conditional distribution
  - Often the one conditional is tricky but the other simple
  - Foundation of many systems we'll see later (e.g. ASR, MT)
- In the running for most important AI equation!

More Bayes' Rule

- 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

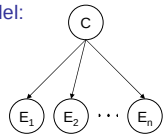
$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

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

Combining Evidence

P(Cavity| toothache, catch)
= α P(toothache, catch| Cavity) P(Cavity)
= α P(toothache | Cavity) P(catch | Cavity) P(Cavity)

- This is an example of a *naive Bayes* model:



$$P(\text{Cause}, \text{Effect}_1 \ldots \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i|\text{Cause})$$

- Total number of parameters is *linear* in n!
- We'll see much more of naïve Bayes next week

Expectations

- Real valued functions of random variables:

$$f : X \rightarrow R$$

- Expectation of a function a random variable

$$E_{P(X)}[f(X)] = \sum_x f(x)P(x)$$

- Example: Expected value of a fair die roll

X	P	f
1	1/6	1
2	1/6	2
3	1/6	3
4	1/6	4
5	1/6	5
6	1/6	6

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6}$$

$$= 3.5$$

Expectations

- Expected seconds wasted because of spam filter

Strict Filter

S	B	P	f
spam	block	0.45	0
spam	allow	0.10	10
ham	block	0.05	100
ham	allow	0.40	0

$$0 \times 0.45 + 10 \times 0.1 + 100 \times 0.05 + 0 \times 0.4 = 6$$

Lax Filter

S	B	P	f
spam	block	0.35	0
spam	allow	0.20	10
ham	block	0.02	100
ham	allow	0.43	0

$$0 \times 0.35 + 20 \times 0.1 + 100 \times 0.02 + 0 \times 0.43 = 4$$

- We'll use the expected cost of actions to drive classification, decision networks, and reinforcement learning…

Utilities

- Preview of utility theory (later)
- Utilities:
  - Function from events to real numbers (payoffs)
  - E.g. spam
  - E.g. airport

Estimation

- How to estimate the a distribution of a random variable X?
- *Maximum likelihood:*
  - Collect observations from the world
  - For each value x, look at the empirical rate of that value:

$$\hat{P}(x) = \frac{\text{count}(x)}{\text{total samples}}$$

  - This estimate is the one which maximizes the likelihood of the data
- *Elicitation:* ask a human!
  - Harder than it sounds
  - E.g. what's P(raining | cold)?
  - Usually need domain experts, and sophisticated ways of eliciting probabilities (e.g. betting games)

Estimation

- Problems with maximum likelihood estimates:
  - If I flip a coin once, and it's heads, what's the estimate for P(heads)?
  - What if I flip it 50 times with 27 heads?
  - What if I flip 10M times with 8M heads?
- Basic idea:
  - We have some prior expectation about parameters (here, the probability of heads)
  - Given little evidence, we should skew towards our prior
  - Given a lot of evidence, we should listen to the data
- How can we accomplish this? Stay tuned!