Probabilities

Some material on these slides are derived from the slides created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley, available at http://ai.berkeley.edu.

Reviewing probabilities

- Probabilities are an important component of modern artificial intelligence.
- While this material you probably covered in other classes, it is useful for us to review it.
- Topics:
  - Random variables
  - Joint and marginal distributions
  - Conditional distributions
  - Product rule, chain rule, Bayes' rule
  - Independence

Uncertainty

- **Observed variables (evidence):** the agent makes observations about the world. These can be incomplete or noisy. They do not directly correspond to the state of the world!
    - Example: medical tests, lidar data
- **Unobserved variables:** the agent needs to reason about the state of the world to infer them
    - Example: medical diagnosis, location of the robot
- **(Probabilistic) model**: the agent knows something about how the known variables (observations) relate to the unknown variables
- **Probabilistic reasoning**: gives us a framework to manage beliefs and knowledge

Random variable

- An aspect of the world about which we might have uncertainty. Denoted with capital letters:
  - R = is it raining?
  - T = is it hot or cold?
  - S = stock market tomorrow
- They have **domains**, and can be **discrete** or **continuous**
  - $R \in \{true, false\}$
  - $T \in \{hot, cold\}$
  - $S \in [0, \infty)$

Probability distributions

- Associate a **probability** with each value in the domain
- A probability is a single number: $P(T = hot) = 0.4$
  - We can simplify it to $P(hot)$ if there is no chance of confusion
- A distribution is a table of probabilities of values

P(T)

T	P
hot	0.4
cold	0.6

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

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 (**outcome**)

$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$ or can be simplified to $P(x_1, x_2, \ldots x_n)$

- Must verify:

$$P(x_1, x_2, \ldots x_n) \geq 0$$
$$\sum P(x_1, x_2, \ldots x_n) = 1$$

Joint distribution example P(T,W)

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

Probabilistic model

- A **probabilistic model_** is a joint distribution over a set of random variables
    - A set of random variables with domains
    - Assignments are called **outcomes**
    - The joint distribution say which outcomes are more likely
    - **Normalized**: sum to 1.0
- Why is this a "model"?
    - It allows me to infer from the known variables the probabilities of the unknown (**"hidden"**) variables
- Ideally, only certain variables interact directly

Events

- An **event** is a set E of outcomes
- Typically, the events we care about are *partial assignments* eg. $P(T = hot)$
- If we have the joint distribution, we can calculate the probability of any event:
  $$P(E) = \sum_{(x_1, x_2, \dots x_n) \in E} P(x_1, x_2, \dots x_n)$$

Reasoning about events:

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

- What is the probability that it is hot AND sunny?
- Probability that it is hot?
- Probability that it is hot OR sunny?

Marginal distributions

- Marginal distributions are sub-tables that eliminate variables
- **Marginalization** (summing out) combine collapsed rows by adding

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

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

W	P
sun	0.6
rain	0.4

Conditional probabilities

- Definition:
  $P(a|b) = \frac{P(a,b)}{P(b)}$
- Example:
  $P(sun|cold) = P(sun, cold)/P(cold) = 0.2/0.5 = 0.4$

Conditional distributions

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Normalization trick

- How to calculate the conditional distribution table for a certain evidence
  - Start with the joint table
- Two steps:
  - **Select** the joint probabilities matching the evidence
  - **Normalize** the selection (make it sum to 1.0)
- Why does this work? Because the sum of selection is the $P(evidence)$

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

Example of the normalization trick

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

Normalization

- Procedure
  - Compute $Z =$ sum over all the entries
  - Divide every entry by Z (also called **normalization constant** or **partition function**)
- Premonition
  - Normalization is simple as a procedure...
  - ... but in more complex situations, it is a **huge problem**, because I need to sum over all possible alternatives...
  - eg. I want the probability of COVID infection
    - But now I have to sum the probabilities of every single other disease
    - Creative ways to avoid calculating the Z are a major research direction

Probabilistic inference

- In general: compute a certain probability we are interested in from other probabilities we know.
- Most of the time we are interested in a joint probability along the lines of $P(x|e_1, e_2, \ldots e_n)$
    - This is called the **belief** of the agent given the **evidence**.
    - Probabilities change with new evidence (can increase or decrease) $\rightarrow$ **updating beliefs**

Inference by Enumeration

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

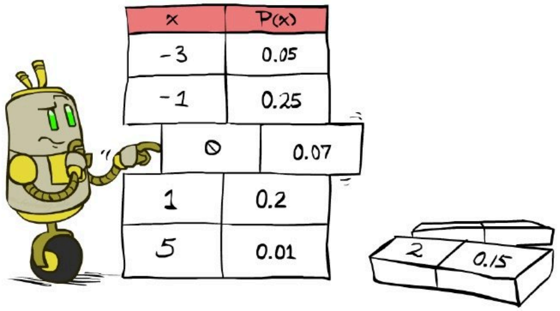
$$\left.\begin{array}{l} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{array}\right\} \begin{array}{l} X_1, X_2, \dots X_n \\ \textit{All variables} \end{array}$$

- We want:

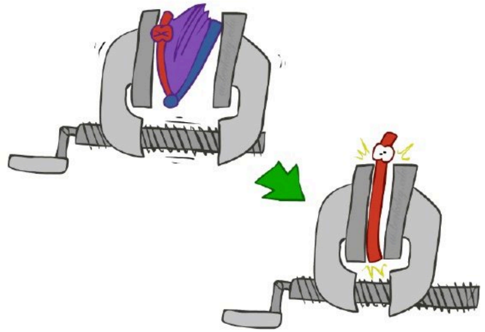
** Works fine with multiple query variables, too*

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

- Step 1: Select the entries consistent with the evidence



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



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

- Step 3: Normalize

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

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

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

Example

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

Example

$$P(W|winter)$$

$$P(W|winter, hot)$$

Inference by enumeration

- Did we just solve probabilistic reasoning once for all?
- Obvious problems:
  - Worst case time complexity $O(d^n)$
  - Space complexity $O(d^n)$ to store the joint distribution
  - ...and who will give us the joint table, anyhow?

The product rule

- $P(y)P(x|y) = P(x, y)$
- It is not really a "rule", just turning around the definition of the conditional

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

The chain rule

- Extending the product rule to multiple conditionals

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

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

Bayes' rule

- Let us start with the insight that the order of variables in the joint does not matter $P(x, y) = P(y, x)$
  - This is not the same for conditionals: $P(x|y) \neq P(y|x)$
- So, the product rule (or, for that matter the chain rule) can be written for any of these orders

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

- Dividing, we get Bayes' rule:

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

The importance of Bayes' rule

- Allows us to calculate $P(x|y)$ from $P(y|x)$
- There are many scenarios where one conditional is difficult, but the reverse is not
  - $P(covid|high\_fever)$ - how do we even start to think about this?
  - $P(high\_fever|covid)$ - easy peasy, about 0.9

Famous example of inference with Bayes' rule

- m: meningitis, s: stiff neck
- givens:

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

We want to calculate $P(+m|+s)$

Inference with Bayes' rule

Let us try to apply Bayes' rule

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)}$$

We don't have $P(+s)$! But we get it by marginalization

$$P(+s) = P(+s,+m) + P(+s,-m)$$

And get the joints from the conditionals using the product rule

$$P(+s,+m) = P(+s|+m)P(+m)$$
$$P(+s,+m) = P(+s|-m)P(-m)$$

Also, due to probabilities summing to 1.0

$$P(-m) = 1.0 - P(+m)$$

Inference with Bayes' rule

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

- It is about 0.00795, approximately 0.8%
- So, just because you have a stiff neck, the likelihood of meningitis is still low, because the overall occurrence of meningitis is low.
- This is a very big problem for many medical diagnosis systems, because a test for a symptom that appears for 99% for a certain disease might still not be a good predictor of the disease!