

PROBABILITY

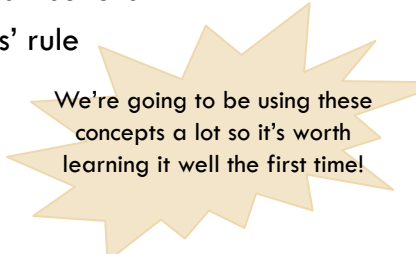
Progress Report

- We've finished Part I: Problem Solving!
- Part II: Reasoning with uncertainty
 - Probability
 - Bayesian networks
 - Reasoning over time (hidden Markov models)
 - Applications: tracking objects, speech recognition, medical diagnosis
- Part III: Learning

Today

- Reading
 - We're skipping to AIMA Chapter 13!

- Goals
 - Random variables
 - Joint, marginal, conditional distributions
 - Product rule, chain rule, Bayes' rule
 - Inference
 - Independence



We're going to be using these concepts a lot so it's worth learning it well the first time!

Handling Uncertainty

- The world is an **uncertain** place
 - Partially observable, non-deterministic
 - On the way to the bank, you get in a car crash!
 - Medical diagnosis
 - Driving to LAX (if you have to)
 - Sensors

- Probability theory gives us a language to reason about an uncertain world.

- Probability theory is beautiful!

Random variables

- A **random variable (rv)** is a variable (that captures some quantity of interest) whose value is random
 - X = the next word uttered by my professor (this is of great interest and importance)
 - Y = the number of people that enter this building on a given day
 - D = the time it will take to drive to LAX
 - W = today's weather
- Like variables in a CSP, random variables have **domains**
 - X in {the, a, of, is, in, if, when, up, on, ..., sky, shenanigans, ...}
 - Y in $[0, 1, 2, 3, 4, 5, 6, \dots, \infty)$
 - D in $[0, \infty)$
 - W in {sun, rain, cloudy, snowy}
- A **discrete rv** has a **countable** domain
- A **continuous rv** has an **uncountable** domain

Discrete Probability distribution

Each value (outcome) in the domain is associated with a real-valued number called a **probability** that reflects the chances of the random variable taking on that value

w	P(W = w)
sunny	0.6
rain	0.1
cloudy	0.29
snow	0.01

} probability distributions }

x	P(X = x)
the	.005
a	.002
of	.0001
...	...
shenanigans	10^{-9}

Constraints for a valid probability distribution:

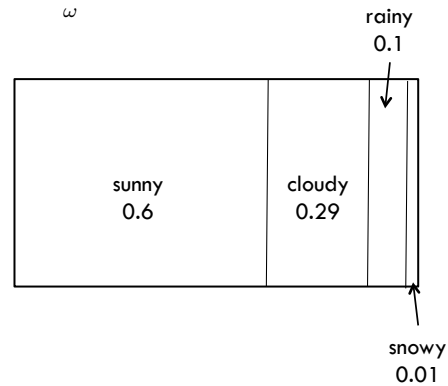
$$0 \leq p(\omega) \leq 1 \text{ such that } \sum_{\omega} p(\omega) = 1$$

Discrete Probability distribution

Constraints for a valid probability distribution:

$$0 \leq p(\omega) \leq 1 \text{ such that } \sum_{\omega} p(\omega) = 1$$

The total probability mass, which is 1, is divided among the possible outcomes



Joint probability distribution

- A **joint distribution** over a set of r.v.s $\{X_1, X_2, \dots, X_n\}$ assigns probabilities to each possible assignment:

$$p(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

$$p(x_1, x_2, \dots, x_n)$$

P(W = w, T = t)

w	t	P
sunny	hot	0.4
rain	hot	0.1
sunny	cold	0.2
rain	cold	0.3

Joint probability distribution

- A **joint distribution** over a set of r.v.s $\{X_1, X_2, \dots, X_n\}$ assigns probabilities to each possible assignment:

$$p(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

$$p(x_1, x_2, \dots, x_n)$$

P(W = w, T = t)

w	t	P
sunny	hot	0.4
rain	hot	0.1
sunny	cold	0.2
rain	cold	0.3

- Still subject to constraints:

$$0 \leq p(x_1, x_2, \dots, x_n) \leq 1 \quad \text{and} \quad \sum_{(x_1, \dots, x_n)} p(x_1, \dots, x_n) = 1$$

Joint probability distribution

- A **joint distribution** over a set of r.v.s $\{X_1, X_2, \dots, X_n\}$ assigns probabilities to each possible assignment:

$$p(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

$$p(x_1, x_2, \dots, x_n)$$

P(W = w, T = t)

w	t	P
sunny	hot	0.4
rain	hot	0.1
sunny	cold	0.2
rain	cold	0.3

- Still subject to constraints:

$$0 \leq p(x_1, x_2, \dots, x_n) \leq 1 \quad \text{and} \quad \sum_{(x_1, \dots, x_n)} p(x_1, \dots, x_n) = 1$$

If we have n variables with domain size d, what is the size of the probability distribution (the number of rows in the table)?

Events

- An **event** is a set E of outcomes
 - sunny AND hot = {(sunny, hot)}
 - sunny = {(sunny, hot), (sunny, cold)}
 - sunny OR hot = {(sunny, hot), (rainy, hot), (sunny, cold)}

Events

- An **event** is a set E of outcomes
 - sunny AND hot = {(sunny, hot)}
 - sunny = {(sunny, hot), (sunny, cold)}
 - sunny OR hot = {(sunny, hot), (rainy, hot), (sunny, cold)}
- The joint distribution can be used to calculate the probability of an event

$$p(E) = \sum_{(x_1, \dots, x_n) \in E} p(x_1, \dots, x_n)$$

w	t	P
sunny	hot	0.4
rain	hot	0.1
sunny	cold	0.2
rain	cold	0.3

The **probability of an event** is the sum of the probability of the outcomes in the set

Marginal Distributions

- Sometimes we have the joint distribution but we're only interested in the distribution of a subset of the variables
 - Called the **marginal distribution**
 - We "marginalize out" the other variables by summing over them
 - Corresponds to a sub-table created by summing over rows

w	t	P
sunny	hot	0.4
rain	hot	0.1
sunny	cold	0.2
rain	cold	0.3

$$p(X = x) = \sum_y P(X = x, Y = y)$$

→
Oftentimes, the events we're interested in are marginal distributions

w	P
sunny	
rain	

t	P
hot	
cold	

Marginal Distributions

- Sometimes we have the joint distribution but we're only interested in the distribution of a subset of the variables
 - Called the **marginal distribution**
 - We "marginalize out" the other variables by summing over them
 - Corresponds to a sub-table created by summing over rows

w	t	P
sunny	hot	0.4
rain	hot	0.1
sunny	cold	0.2
rain	cold	0.3

$$p(X = x) = \sum_y P(X = x, Y = y)$$

→
Oftentimes, the events we're interested in are marginal distributions

w	P
sunny	0.6
rain	0.4

t	P
hot	0.5
cold	0.5

Conditional (posterior) distribution

- Often, we observe some information (**evidence**) and we want to know the probability of an event conditioned on this evidence

$$p(W \mid T = \text{cold})$$



evidence

In all the worlds where $T = \text{cold}$, what is the probability that $W = \text{sunny}$?
That $W = \text{rainy}$?

- This is called the **conditional distribution** (or the **posterior distribution**), e.g. the distribution of W conditioned on the evidence $T = \text{cold}$

Conditional (posterior) distribution

- The conditional distribution is given by the equation

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

	hot	cold	
rainy	0.1	0.3	rainy
sunny	0.4		sunny

$p(W \mid T = \text{cold})?$

Whereas before the total probability mass was 1, the total probability mass is now $p(T = \text{cold})$. We compute everything in relation to this value.

Conditional (posterior) distribution

- The conditional distribution is given by the equation

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

	hot	cold	
rainy	0.1	0.3	rainy
sunny	0.4		

p(W | T=cold)?

Conditional (posterior) distribution

- The conditional distribution is given by the equation

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

	hot	cold	
rainy	0.1	0.3	rainy
sunny	0.4		

p(W | T=hot)?

Conditional (posterior) distribution

- The conditional distribution is given by the equation

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

	hot	cold	
rainy	0.1	0.3	rainy
sunny	0.4		

p(T | W=rainy)?

Conditional (posterior) distribution

- The conditional distribution is given by the equation

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

	hot	cold	
rainy	0.1	0.3	rainy
sunny	0.4		

p(T | W=sunny)?

Normalization constants

$$p(W = s|T = c) = \frac{p(W = s, T = c)}{p(T = c)} = \frac{0.2}{0.5} = 0.4$$

$$p(W = r|T = c) = \frac{p(W = r, T = c)}{p(T = c)} = \frac{0.3}{0.5} = 0.6$$

- Note that $p(T=c)$ is constant no matter the value of W
- We call $p(T=c)$ a **normalization constant** because:
 1. It is constant with respect to the distribution of interest $p(W|T=c)$
 2. It ensures that the distribution sums to 1 (i.e. it restores the distribution $p(W|T=c)$ back to the “normal” condition of summing to 1)

Normalization constants

- Even without the prob. of evidence, i.e. $p(T=c)$, we could calculate the conditional distribution using the joint distrib. alone
 - Step 1: Compute $Z = \text{sum of } p(W, T=c) \text{ for all values of } W$
 - Step 2: Divide each joint probability by Z
 - (All we've done is computed the prob. of evidence, i.e. $p(T=c)$, from the joint distribution by marginalizing over W)

$$p(X, Y) \propto p(X|Y)$$

$$\left\langle \frac{0.2}{0.2 + 0.3}, \frac{0.3}{0.2 + 0.3} \right\rangle = \langle 0.4, 0.6 \rangle$$

Normalization constant

- Normalize the following distribution:

$P(W = w, T = t)$

w	t	P
sunny	hot	20
rain	hot	5
sunny	cold	10
rain	cold	15

Step One:

$$Z = 20 + 5 + 10 + 15 \\ = 50$$

Step Two:

$$p(W, T) = \left\langle \frac{20}{50}, \frac{5}{50}, \frac{10}{50}, \frac{15}{50} \right\rangle = \langle .40, .10, .20, .30 \rangle$$

Summary of distributions so far

