PROBABILITY

Progress Report

- We've finished Part I: Problem Solving!
- Part II: Reasoning with uncertainty
  - Probability
  - Bayesian networks
  - Reasoning over time (hidden Markov models)
- Part III: Machine Learning

Today

- Reading
  - AIMA Chapter 13.2-13.5
- 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,…, ∞)
  - D in [0, ∞)
  - 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
…	…
peregrinate	10^{-9}

Constraints for a valid probability distribution:

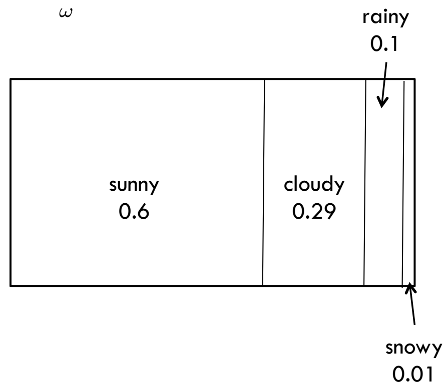
$$0 \le p(\omega) \le 1 \text{ such that } \sum_{\omega} p(\omega) = 1$$

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 $\{X_1, X_2, \ldots, X_n\}$ assigns probabilities to each possible assignment:

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

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

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:

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

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

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

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

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

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

P(W = w, T = t)

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

P(W = w)

w	P
sunny	
rain	

P(T = t)

t	P
hot	
cold	

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 | T = cold)

evidence

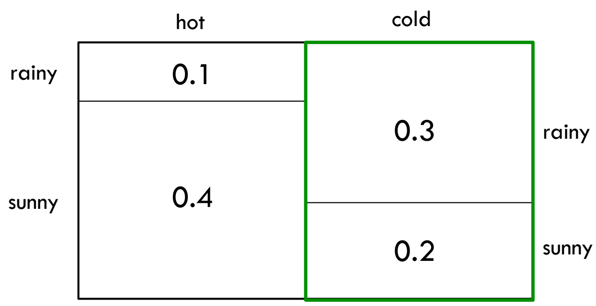
In all the worlds where T=cold, what is the probability that W = sunny? That W = rainy?

- This is called the conditional distribution, e.g. the distribution of W conditioned on the evidence T = cold

Conditional (posterior) distribution

- The conditional distribution is given by the equation

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



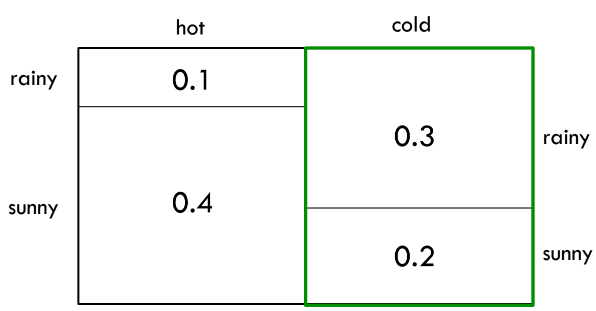
p(W| T=cold)?

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

Conditional (posterior) distribution

- The conditional distribution is given by the equation

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

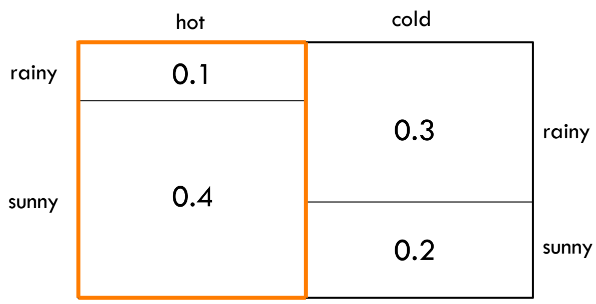


p(W| T=cold)?

Conditional (posterior) distribution

- The conditional distribution is given by the equation

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

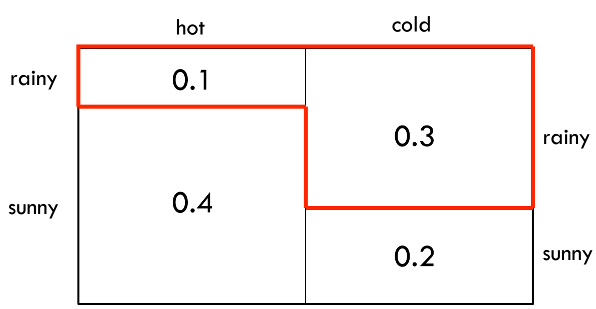


p(W| T=hot)?

Conditional (posterior) distribution

- The conditional distribution is given by the equation

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

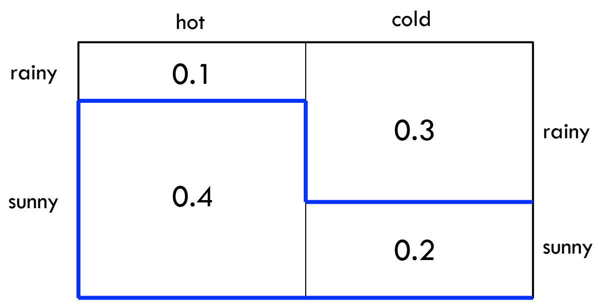


p(T| W=rainy)?

Conditional (posterior) distribution

- The conditional distribution is given by the equation

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



p(T| W=sunny)?

Conditional and Joint are just a constant apart!

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

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

- 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)

Conditional and Joint are just a constant apart!

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

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

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

"is proportional to"

Conditional and Joint are just a constant apart!

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

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

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

Normalization Trick

$$< \frac{0.2}{0.2+0.3}, \frac{0.3}{0.2+0.3} > = < 0.4, 0.6 >$$

Normalization Trick

- Step 1: Compute Z = sum of p(W, T=c) for all values of W
- Step 2: Divide each *joint probability* by Z
- (All we're doing is computing the prob. of evidence, i.e. p(T=c), from the joint distribution by marginalizing over W)

$$< \frac{0.2}{0.2+0.3}, \frac{0.3}{0.2+0.3} >=< 0.4, 0.6 >$$

Normalization Trick

- Normalize the following distributions:

P(W = w, T = t)

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

p(W| T=hot)?
p(W| T=cold)?
p(T| W=sunny)?
p(T| W=rainy)?

Summary of distributions so far