CSCI 3022

intro to data science with probability & statistics

Sept 17, 2018

HW2 Posted! Due in 11 days.

1. Bayes' Rule
2. Random variables and probability distributions

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Public Service Announcements

- Don't miss 10 days in all your classes and shave 0.5-1.0 off your semester GPA!   (Or: don't put others at risk for the flu!)
- **Free Flu Shots for CU**: https://www.colorado.edu/healthcenter/flu !

- Register to vote, and vote—one of the most beautiful parts of Adulting.
- **Online Reg**: colorado.edu/registrar/students/registration/mycuinfo/register-vote
- Call your parents.

Previously, on CSCI 3022

Conditional probability: The probability that A occurs **given** that C has occurred is

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)}$$

Product rule: $P(A \cap C) = P(A \mid C)P(C)$

Independence: events A and B are independent if and only if:

1. $P(A \mid B) = P(A)$
2. $P(B \mid A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$

Law of total probability (LTP): If $C_1, C_2, \ldots, C_m$ are disjoint events such that $C_1 \cup C_2 \cup \ldots \cup C_m = \Omega$ then the probability of an arbitrary event A can be written as

$$P(A) = P(A \mid C_1)P(C_1) + P(A \mid C_2)P(C_2) + \ldots + P(A \mid C_m)P(C_m)$$

Bayes' Rule

- Recall the probability that two events intersect: $P(A \cap B) = P(A|B)P(B)$ ✓

$$\|$$

- But we can write it the other way, too: $P(A \cap B) = P(B|A)P(A)$ ✓

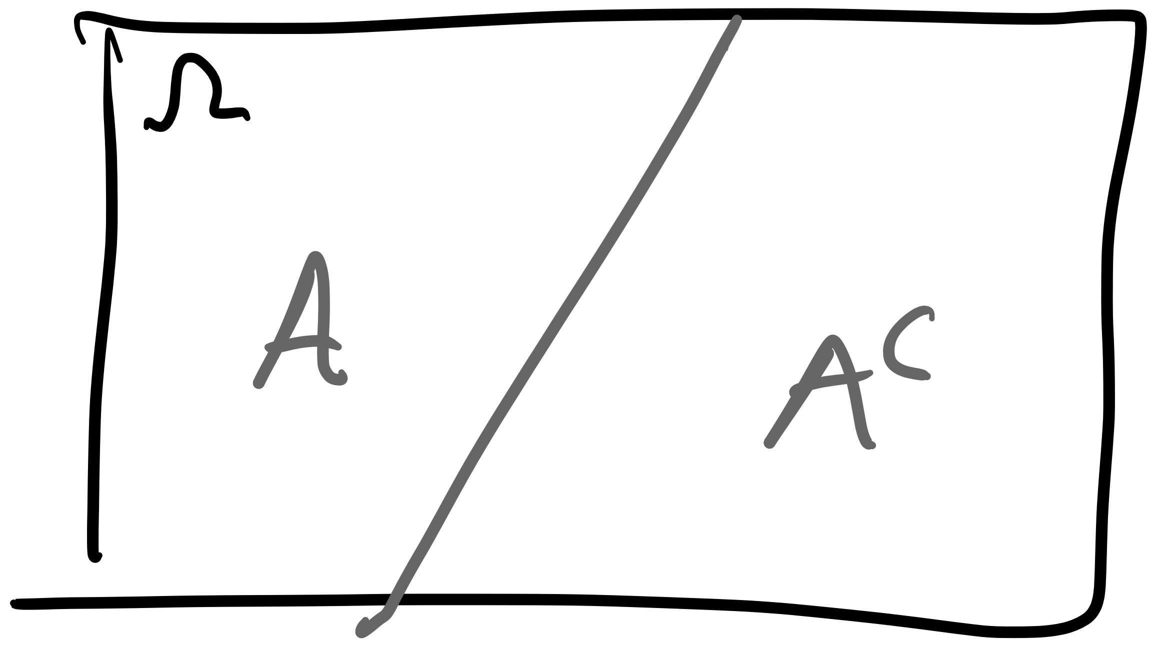
- And then… $P(A|B)P(B) = P(B|A)P(A)$ Assume $P(B) > 0$

- And this, is Bayes' Rule: $$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

- [Bonus: suppose A and B are independent…]

$P(A|B) = P(A)$ AND $P(B|A) = P(B)$ subs… $P(A) = \frac{\cancel{P(B)}P(A)}{\cancel{P(B)}}$

Bayes' Rule + Law of T. P.



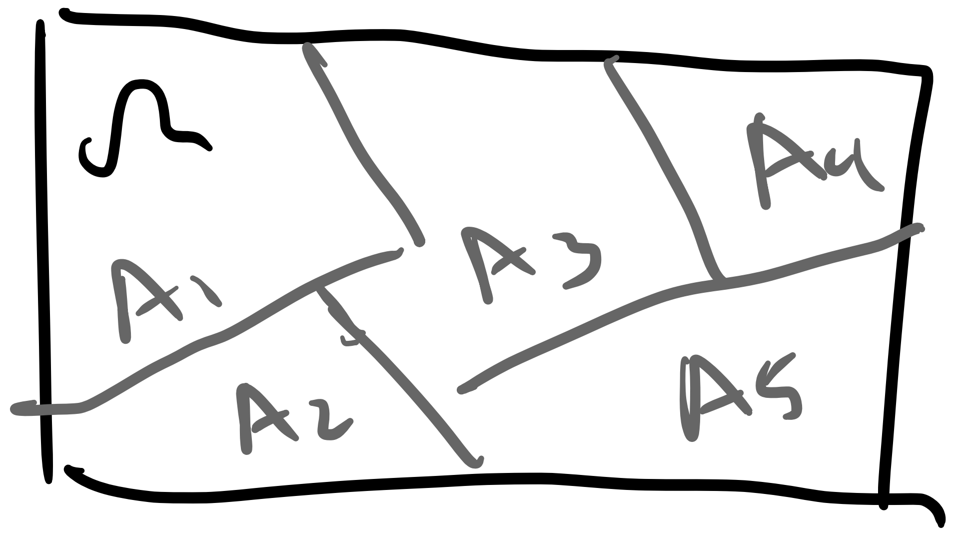
- Use Law of Total Probability to rewrite the denominator:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}} \overset{\text{LTP}}{=} \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

- Or, if B can be broken into K disjoint events:

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)} \overset{\text{LTP}}{=} \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_m)P(A_m)}$$

$$= \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}$$



Bayes' Rule + Law of T. P.

- Use Law of Total Probability to rewrite the denominator:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

- Or, if B can be broken into K disjoint events:

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots}$$

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_{k=1}^{K} P(B|A_k)P(A_k)}$$

Drugs, Prostates, TSA

- One of the most common applications of Bayes' rule is when we develop a test for something, but **the test is not always accurate**.
- Imagine that 1% of CS professors are using a drug *Bayes Salts*. There is a test that detects Bayes Salts on the breath of professors 98% of the time when a prof is using, and incorrectly calls BS only 1% of the time when a prof *is not* using.
- Suppose we test Professor Charles Xavier… he's **positive** for Bayes Salts! What is the probability that Professor Xavier is **actually** on Bayes Salts?

Drugs, Prostates, TSA

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- 1% of ~~math~~ CS professors *Bayes Salts* users. ⟵
- If on *Bayes Salts*, test says **+** 98% of the time
- If not on *Bayes Salts*, test says **+** only 1% of the time.
- What is probability that Prof. X is a user if he tests positive for *Bayes Salts*?

is user $= A$

not user $= A^c$

$$P(\text{user} \mid \text{test}\oplus) \overset{\text{Bayes}}{=} \frac{P(\text{test}\oplus \mid \text{user})P(\text{user})}{P(\text{test}\oplus)} \overset{\text{LTP}}{=} \frac{P(\text{test}\oplus \mid \text{user})P(\text{user})}{P(\text{test}\oplus \mid \text{user})P(\text{user}) + P(\text{test}\oplus \mid \text{not user})P(\text{not user})}$$

$$= \frac{0.98 \cdot \cancel{0.01}}{0.98 \cdot \cancel{0.01} + \cancel{0.01} \cdot 0.99} = \frac{0.98}{0.98 + 0.99} \approx \boxed{0.497}$$

Teach the controversy! 🫢

- Should we test men for prostate cancer?
- Bayes' Rule allows us to write down the probability that someone who tests positive for prostate cancer *actually has* prostate cancer.
- False positives may cause huge amounts of stress, heartache, and even unnecessary surgery!
- On the other hand, if you don't test for cancer, you may not discover it until it's too late.
- Things are slightly more complicated than this: age, PSA cutoffs, etc.

Flipping around a previous problem

- **Suppose I have two bags of marbles.** The first bag contains 6 white marbles and 4 black marbles. The second bag contains 3 white marbles and 7 black marbles. Now suppose I put the two bags in a box. If I close my eyes, grab a bag from the box, reach into the bag and pull out a white marble. What is the probability that I picked Bag 1?

$$P(\text{bag } 1 \mid \text{white marble}) = \frac{P(\text{white marble} \mid \text{bag } 1)\, P(\text{bag } 1)}{P(\text{white marble})}$$

Bayes!

$$= \frac{P(\text{white marble} \mid \text{bag } 1)\, P(\text{bag } 1)}{P(\text{white marble} \mid \text{bag } 1)\, P(\text{bag } 1) + P(\text{white marble} \mid \text{bag } 2)\, P(\text{bag } 2)}$$

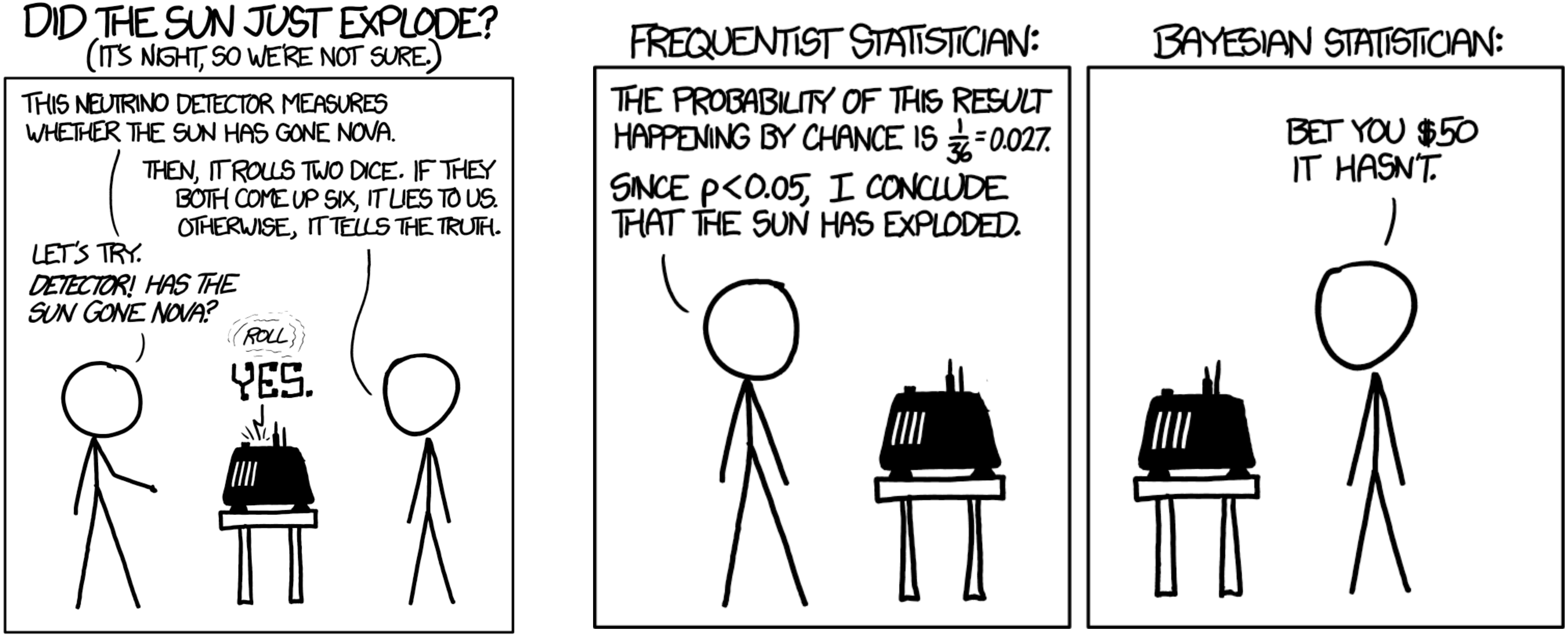
LtP

$$= \frac{\frac{6}{\cancel{10}} \cdot \cancel{\frac{1}{2}}}{\frac{6}{\cancel{10}} \cdot \cancel{\frac{1}{2}} + \frac{3}{\cancel{10}} \cdot \cancel{\frac{1}{2}}} = \frac{6}{6+3} = \frac{6}{9} = \boxed{\frac{2}{3}}$$

Bayes' Rule in the wild ☀️

xkcd
it is funny.

- Bayes' Rule is very helpful because it helps us incorporate our **prior knowledge** about probabilities into our conclusions.



Bayes' Rule in the wild

Likelihood

- Bayes' Rule is very helpful because it helps us incorporate our **prior knowledge** about probabilities into our conclusions.

Prior

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Posterior

evidence

- When we calculated the probability that Professor X was on Bayes Salts, which one of these terms was our prior knowledge of the background rate of Salts use?

1% background rate.

Bayes' Rule in machine learning

- Often, we have a **model with parameters** M and we have **data** D.

- Our goal is to learn the parameters M from the data. Yet we also have some beliefs about the parameters, and no particular beliefs about the data.

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

Flip coin, $\{9H, 1T\} = D$

M = bias of the coin

I tell you: "I'm pretty sure it's not biased" Prior: $p \sim 0.5$

Random variables

- Say I roll two dice.
- What's the most likely outcome? all same
- What's the most likely sum? 7

Random variables

- Say I roll two dice. What's the sample space? What are the tables of sums, differences, and maxima?

$P(\text{sum} = 7) = \frac{6}{36} = \frac{1}{6}$

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Sum

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

Random variables

- **The key**: the dice are random, so the sum is random!
- Let's *sidestep* the sample space entirely and just go straight to the thing we care about: the sum.
- We call the sum of the dice a *random variable*.

Discrete random variables

- **Definition**: a discrete random variable is a function that maps the elements of a sample space Ω to a finite number of values $a_1, a_2, \ldots, a_N$ or an infinite number of values $a_1, a_2, \ldots$

- Note: even if there are an infinite number, the values must be discrete.

- **Examples** of discrete random variables:
  - Sum of dice; difference of dice; maximum of dice.
  - Number of flips until we get a heads.

Probability mass functions

- **Definition**: a *probability mass function* is the map between the random variable’s values and the probabilities of those values.

$$f(a) = P(X = a)$$

PMF

- Called a “probability mass” function (PMF) because each of the random variable's values has some probability mass (or weight) attached to it.

- Since the PMF is a probability function, what is the sum of all the masses?

$P(\Omega) = 1$ (def'n of a prob function)

$a \in \Omega$

a \in \Omega

a is in the set Omega.

$$\sum_{a \in \Omega} f(a) = 1$$

Probability mass functions

- **Question**: what is the probability mass function for the number of coin flips until a biased coin comes up heads?

Want: function $f(n)$ that tells me the prob. that I wait n flips to get the first heads.

Cumulative distribution functions

- **Definition**: a *cumulative distribution function* (CDF) is a function whose value at point *a* is the cumulative sum of probability masses up until *a*.

$$F(a) = P(X \leq a)$$

- **Question**: what's the relationship between the PMF and the CDF?

Cumulative distribution functions

- **Example**: What is the probability that I roll two dice and they add up to at most 9?

Cumulative distribution functions

- **Example**: What is the probability that I roll two dice and they add up to between 3 and 6, inclusive?