CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD COLLEGE

Admin

- In lab today: check-ins about **Lab 4**
- **Midterm 1** handed out on Thursday (due the following Thursday – take in a 3 hour block)
- **Thursday**: review session in class
- **RIGHT NOW**: make sure you have
  - Midterm 1 Study Guide
  - Feedback form
  - Handout 9

Data Science and the Liberal Arts
A View from Down Under

Daniel Russo-Batterham
Senior Research Data Specialist
University of Melbourne

You are here!

Data Science talk recommendation!

Thursday, October 5, 2023
4:15pm in Lutnick 232

Sponsored by the Music Department, TriCo Data Science, and HAVERFORD LIBRARIES

Lab 4

- Lab 4 due TONIGHT

Midterm 1 Notes

- Handed out in class this Thursday, due the following Thursday.
- Timed exam: **3 hour limit**. DO NOT open the exam until you are ready to take it for 3 hours!
- You may use a one page (front and back) “study sheet”, handwritten, created by you
- You may also use a regular calculator
- Outside of your “study sheet” and calculator, **no other notes or resources**
- **As per the Honor Code, all work must be your own**

Outline for October 2

- Go over Lab 2
- Intro to probability
- Intro to Bayesian models
- Intro to algorithmic bias

Outline for October 2

- Go over Lab 2
- Intro to probability
- Intro to Bayesian models
- Intro to algorithmic bias

Lab 2: not posted online

Outline for October 2

- Go over Lab 2
- Intro to probability
- Intro to Bayesian models
- Intro to algorithmic bias

Intro to Probability

- The **probability** of an **event** e has a number of epistemological interpretations
- Assuming we have **data**, we can count the number of times e occurs in the dataset to estimate the probability of e, $P(e)$.

$$P(e) = \frac{\text{count}(e)}{\text{count}(\text{all events})}.$$

- If we put all events in a bag, shake it up, and choose one at random (called **sampling**), how likely are we to get e?

Intro to Probability

- Suppose we flip a fair coin
- What is the probability of heads, $P(e = H)$?

Intro to Probability

- Suppose we have a fair 6-sided die.

$$\frac{count(s)}{count(1) + count(2) + count(3) + \cdots + count(6)} = \frac{1}{1+1+1+1+1+1} = \frac{1}{6}$$

Intro to Probability

- What about a die with on ly three numbers $\{1, 2, 3\}$, each of which appears twice?
- What's the probability of getting "1"?

Intro to Probability

- The set of all probabilities for an event e is called a **probability distribution**
- Each die roll is an independent event (Bernoulli trial).

Materials by Alvin Grissom II

Intro to Probability

- Which is greater, $P(HHHHH)$ or $P(HHTHH)$?

Materials by Alvin Grissom II

Intro to Probability

Probability Axioms

1. Probabilities of events must be no less than 0. $P(e) \geq 0$ for all e.
2. The sum of all probabilities in a distribution must sum to 1. That is, $P(e_1) + P(e_2) + \ldots + P(e_n) = 1$. Or, more succinctly,

$$\sum_{e \in E} P(e) = 1.$$

Intro to Probability

Joint Probability

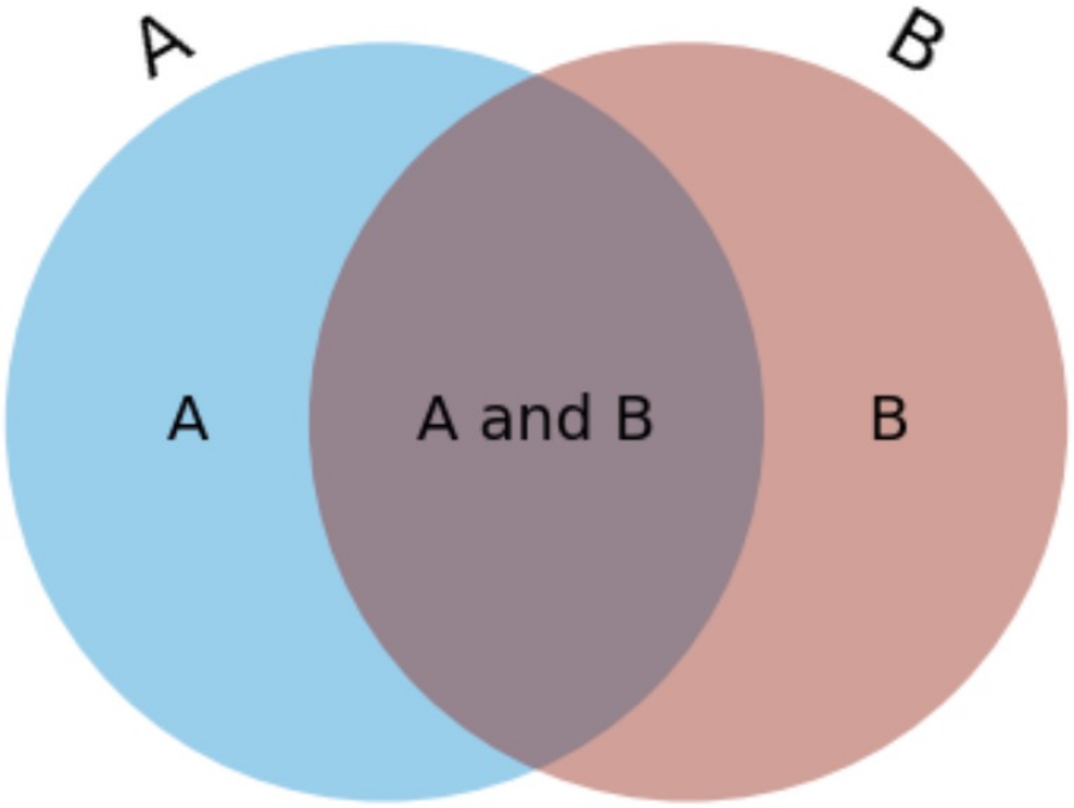
The probability that two independent events e_1 and e_2 *both* occur is given by their product.

$$P(e_1 \wedge e_2) = P(e_1 \cap e_2) = P(e_1)P(e_2) \text{ when } e_1 \cap e_2 = \emptyset$$

- Intuitively, think of every probability as a *scaling factor*.
- You can think of a probability as the fraction of the probability space occupied by an event e_1.
  - $P(e_1 \wedge e_2)$ is the fraction of of e_1's probability space wherein e_2 also occurs.
  - So, if $P(e_1) = \frac{1}{2}$ and $P(e_2) = \frac{1}{3}$, then $P(e_2, e_2)$ is a third of a half of the probability space or $\frac{1}{3} \times \frac{1}{2}$.

Intro to Probability

Joint Probability



Materials by Alvin Grissom II

Intro to Probability

Conditional Probability

- A **conditional probability** is the probability that one event occurs given that we take another for granted.
- The probability of e_2 given e_1 is $P(e_2 \mid e_1)$.
- This is the probability that e_2 will occur given that we take for granted that e_1 occurs.

Intro to Probability

Marginal Probability Distributions

Given a discrete joint probability distribution function $P(X, Y)$, how would we find $P(X)$?

- "Marginalize out" the Y (sum over all all $y \in Y$).
- Fix the X.
- Discrete Case: $p(x) = \sum_{y \in Y} P(x, y)$
- Continuous Case: $p(x) = \int p(x, y) dy$

example

R = rain

U = umbrella

If $P(R) = 20\%$ and

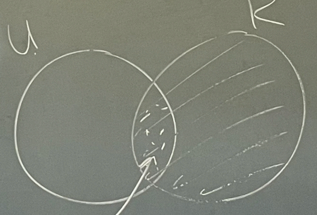
$P(R, U) = 15\%$, and (joint prob)

what is $P(U|R)$? given

(conditional probability)

Bayes Rule

$$\overbrace{P(U|R) = \frac{P(R, U)}{P(R)}}^{\text{Bayes Rule}} = \cancel{\frac{P(R)P(U)}{P(R)}}$$



$$= \frac{0.15}{0.2} = \boxed{0.75}$$

▨ = 0.2

⬚ = 0.15

P(R, U)

~~) P(u)~~

~~P(R)~~

Bayes Rule

$$P(A,B) = P(A)\,P(B \mid A)$$

$$P(A,B) = P(B)\,P(A \mid B)$$

$$P(A,B) =$$

$$\sum_{a \in \mathrm{vals}(A)} P(a) = 1 \quad , \quad \underbrace{P(R)}_{0.2} + \overbrace{\underbrace{P(\bar{R})}_{0.8}}^{\text{prob of } \underline{\text{not}}\ R} = 1$$

Independence

$$\underbrace{P(A,B) = P(A)\,P(B)}_{\text{not true in general!!}}$$

Conditional Independence

$$P(A \mid B, C) = P(A \mid C)$$

A: thunder, B: rain, C: lightning

"Thunder is independent of rain give lightning."

$$\cancel{P(A \mid B) = P(A)} \quad \text{Standard independence}$$

$P(B \mid A)$

Marginalizing

$$P(A) = \sum_{b \in \mathrm{val}(B)} P(A, B=b)$$

$$P(U) = P(R, U) + P(\bar{R}, U)$$

Example

$$\underbrace{P(\text{spam} \mid \underbrace{\text{words}}_{\text{email}})}_{\text{posterior}} = \frac{p(\text{spam}, \text{words})}{p(\text{words})}$$

very difficult!

"data"

$$\frac{p(\text{spam}, \text{words})}{p(\text{spam}, \text{words}) + p(\overline{\text{spam}}, \text{words})}$$

$$= \boxed{\frac{p(\text{spam})\, p(\text{words} \mid \text{spam})}{p(\text{spam})\, p(\text{words} \mid \text{spam}) + p(\overline{\text{spam}})\, p(\text{words} \mid \overline{\text{spam}})}}$$

prior

evidence

likelihood (generative)

X

Handout 9

Handout 9

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

①

$$\overbrace{P(D|pos)}^{\text{posterior}} = \frac{\overbrace{P(D)}^{\text{prior}}\,\overbrace{P(pos|D)}^{\text{likelihood}}}{\underbrace{P(pos)}_{\text{evidence}}}$$

$$= \frac{P(D)P(pos|D)}{P(D)P(pos|D) + \underbrace{P(H)}\,P(pos|H)}$$

true \ pred	neg −	pos +
H −	9/10	1/10
D +	1/10	9/10

$$\frac{\frac{1}{\cancel{100}}\cdot\frac{9}{\cancel{10}}}{\frac{1}{\cancel{100}}\cdot\frac{9}{\cancel{10}} + \frac{99}{\cancel{100}}\cdot\frac{1}{\cancel{10}}} = \frac{9}{9+99} = \frac{9}{108} = \frac{1}{12}$$

$$\approx \boxed{0.0825} \qquad \boxed{8\%}$$

Outline for October 2

- Go over Lab 2
- Intro to probability
- Intro to Bayesian models
- Intro to algorithmic bias

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$\boxed{p(y=k|\boldsymbol{x}) = \frac{p(y=k)p(\boldsymbol{x}|y=k)}{p(\boldsymbol{x})}}$$

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Evidence**: this is the data (features) we actually observe, which we think will help us predict the outcome we're interested in

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Prior**: without seeing any evidence (data), what is our prior believe about each outcome (intuition: what is the outcome in the population as a whole?)

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y=k|\boldsymbol{x}) = \frac{p(y=k)p(\boldsymbol{x}|y=k)}{p(\boldsymbol{x})}$$

- **Posterior**: this is the quantity we are actually interested in. **Given** the evidence, what is the probability of the outcome?

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Likelihood**: given an outcome, what is the probability of observing this set of features?

Examples

- Computing the probability an email message is **spam**, given the **words** of the email

- Another example: what is the probability of **Trisomy 21** (Down Syndrome), given the amount of sequencing of each chromosome?

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

Goal:

$$\mathbb{P}(T_{21}|\vec{q}\,) = \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,)}$$

$$= \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q}\,|T_{21}^C) \cdot \mathbb{P}(T_{21}^C)}$$

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

Goal:

Prior probability of T_{21}

$$\mathbb{P}(T_{21}|\vec{q}\,) = \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,)}$$

$$= \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q}\,|T_{21}^C) \cdot \mathbb{P}(T_{21}^C)}$$

Prior:

$P(T_{21})$

Maternal Age	Trisomy 21	All Trisomies
20	1 in 1,667	1 in 526
21	1 in 1,429	1 in 526
22	1 in 1,429	1 in 500
23	1 in 1,429	1 in 500
24	1 in 1,250	1 in 476
25	1 in 1,250	1 in 476
26	1 in 1,176	1 in 476
27	1 in 1,111	1 in 455
28	1 in 1,053	1 in 435
29	1 in 1,000	1 in 417
30	1 in 952	1 in 384
31	1 in 909	1 in 384
32	1 in 769	1 in 323
33	1 in 625	1 in 286
34	1 in 500	1 in 238
35	1 in 385	1 in 192
36	1 in 294	1 in 156
37	1 in 227	1 in 127
38	1 in 175	1 in 102
39	1 in 137	1 in 83
40	1 in 106	1 in 66
41	1 in 82	1 in 53
42	1 in 64	1 in 42
43	1 in 50	1 in 33
44	1 in 38	1 in 26
45	1 in 30	1 in 21
46	1 in 23	1 in 16
47	1 in 18	1 in 13
48	1 in 14	1 in 10
49	1 in 11	1 in 8

Anonymous feedback forms

Outline for October 2

- Go over Lab 2
- Intro to probability
- Intro to Bayesian models
- Intro to algorithmic bias

Next time!