# CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD
COLLEGE

Materials by Sara Mathieson

# Admin

- Lab 4 due tonight

# Admin

- **Lab 3** grades & feedback will be posted on Wednesday
- **Midterm 1** will be handed out on Wednesday (due the following Wednesday – take in a 3 hour block)
- **Tuesday + Wednesday**: review sessions

# Midterm 1 Notes

- Handed out in class this Wednesday, due at the beginning of class the following Wednesday.
- Timed exam: **3 hour limit**. DO NOT open the exam until you are ready to take it for 3 hours!
- You may use one letter 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 today

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

# Outline for today

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

# Lab 2: not posted online

# Outline for today

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

# 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 $e$?

Materials by Alvin Grissom II

# Intro to Probability

- Suppose we have a fair 6-sided die.
- What's the probability of getting "1"?

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

# 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

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

$$P(e=1) = \frac{count(1)}{count(1) + count(2) + count(3)} = \frac{2}{2+2+2} = \frac{1}{3}.$$

# Intro to Probability

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

# Intro to Probability

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

# Intro to Probability

- Which is greater, $P(HHHHH)$ or $P(HHTHH)$?
- Since the events are independent, they're equal

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

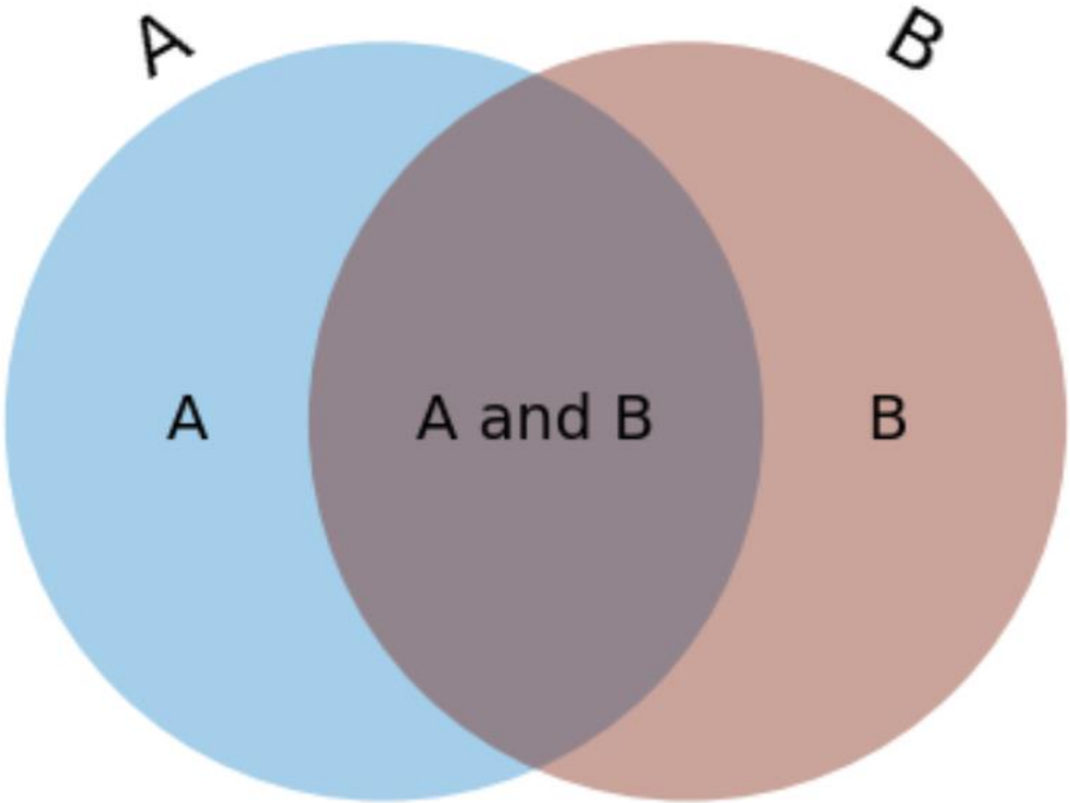
The probability that two independent events $e_1$ and $e_2$ *both* occur is given by their product.

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

- 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 $P(e_2, e_1)$ is a third of a half of the probability space or $\frac{1}{3} \times \frac{1}{2}$.

# Intro to Probability

## Joint Probability



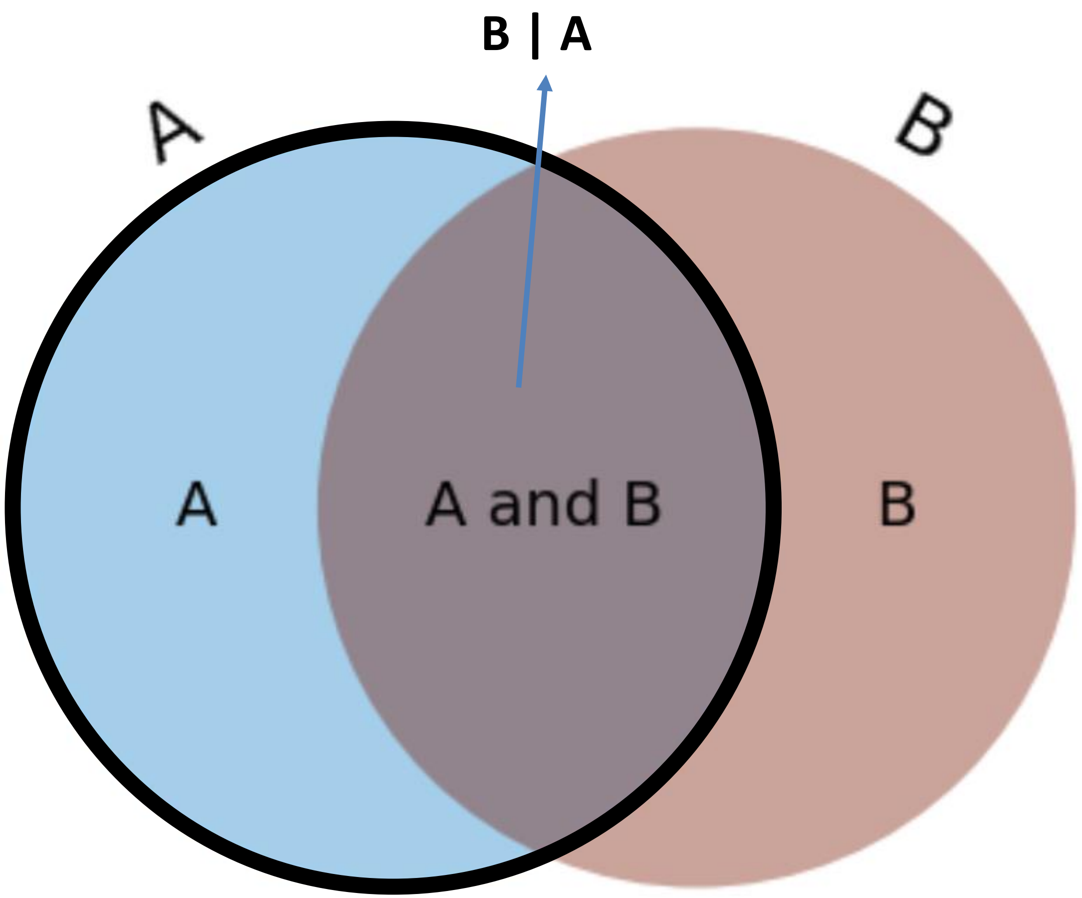

Materials by Alvin Grissom II

## 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

## Conditional Probability




Materials by Alvin Grissom II

## example

$R$ = rain

$U$ = umbrella

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

$P(R, U) = 15\%$, $\boxed{\text{and}}$ (joint prob)

what is $P(U|R)$?

given (conditional probability)

Bayes Rule

$$P(U|R) = \frac{P(R,U)}{P(R)} = \cancel{\frac{P(R)P(U)}{P(R)}}$$


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

$\square = 0.2$

$\square = 0.15$

$P(R, U)$.

$\Sigma$

$a \in$

Ind

# Bayes' Theorem

- $P(A,B) = P(A|B)P(B)$
- $P(A,B) = P(B|A)P(A)$

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

Independence: $P(A,B) = P(A)P(B)$ not true in general!!!

## Conditional Independence

$$P(\underset{\uparrow\atop\text{thunder}}{A} \mid \underset{\uparrow\atop\text{rain}}{B}, \underset{\uparrow\atop\text{lightning}}{C}) = P(A \mid C)$$

" Thunder is independent of rain
give lightning. "

$$\left.\cancel{P(A \mid B) = P(A)}\right\} \text{standard independence}$$

$P(B \mid A)$

## 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 $y \in Y$).
- Discrete Case: $p(x) = \sum_{y \in Y} P(x, y)$
- Continuous Case: $p(x) = \int p(x, y) dy$

## Marginalizing

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

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

## 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)

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

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

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

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

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

# Outline for today

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

Next time!