CS 540 Introduction to Artificial Intelligence
Probability

University of Wisconsin-Madison
Spring 2025

Announcements

- **HW 1**:
  - Writing assignment---nothing too stressful
- Piazza: https://piazza.com/wisc/spring2025/cd71

  Passcode: introtoai

- Class roadmap:

Probability	Mostly Foundations
Linear Algebra	Mostly Foundations
Statistics	Mostly Foundations
Logic	Mostly Foundations
NLP	

Probability: What is it good for?

- Language to express **uncertainty**

In AI/ML Context

- Quantify predictions

$[p(\text{lion}), p(\text{tiger})] = [0.98, 0.02]$

$[p(\text{lion}), p(\text{tiger})] = [0.01, 0.99]$

$[p(\text{lion}), p(\text{tiger})] = [0.43, 0.57]$

* If we know for sure the photo must contain either a lion or a tiger

Model Data Generation

- Model complex distributions

StyleGAN2 (Kerras et al ’20)

Win At Poker

- Wisconsin Ph.D. student Ye Yuan 5th in WSOP

Not unusual: probability began

as study of gambling techniques

Cardano

Liber de ludo aleae
Book on Games of Chance
1564!

pokernews.com

Outline

- Basics: definitions, axioms, RVs, joint distributions
- Independence, conditional probability, chain rule
- Bayes' Rule and Inference

Basics: Outcomes & Events

- **Outcomes:** possible results of an **experiment**

$$\Omega = \underbrace{\{1, 2, 3, 4, 5, 6\}}_{\text{outcomes}}$$

- **Events:** subsets of outcomes we're interested in

$$\underbrace{\emptyset, \{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \Omega}_{\text{events}}$$

- Always include $\emptyset, \Omega$

Basics: Probability Distribution

- We have outcomes and events.
- Assign **probabilities**: for each event $E, P(E) \in [0,1]$

Back to our example:

$$\underbrace{\emptyset, \{1\}, \{2\}, \ldots, \{1,2\}, \ldots, \Omega}_{\text{events}}$$

$$P(\{1,3,5\}) = 0.2, P(\{2,4,6\}) = 0.8$$

Basics: **Axioms**

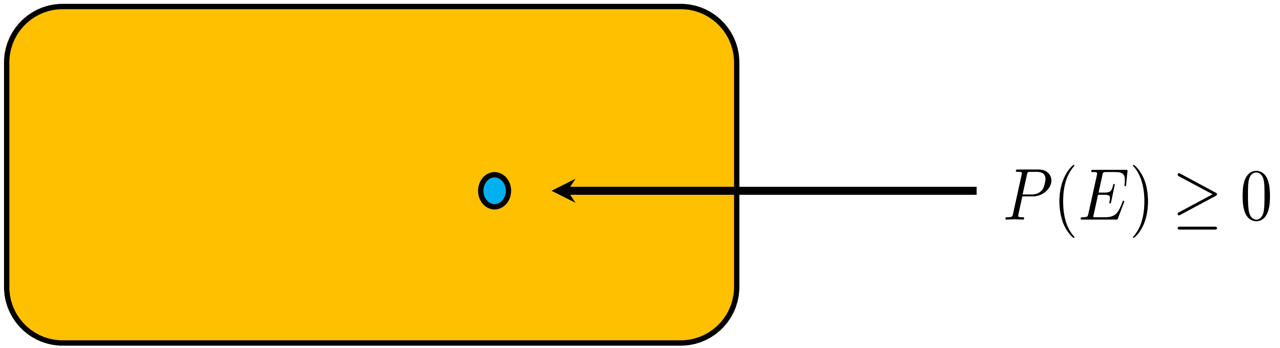
- Rules for probability:
  - For all events E, $P(E) \geq 0$
  - Always, $P(\emptyset) = 0, P(\Omega) = 1$
  - For disjoint events, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

- Easy to derive other laws. Ex: non-disjoint events

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

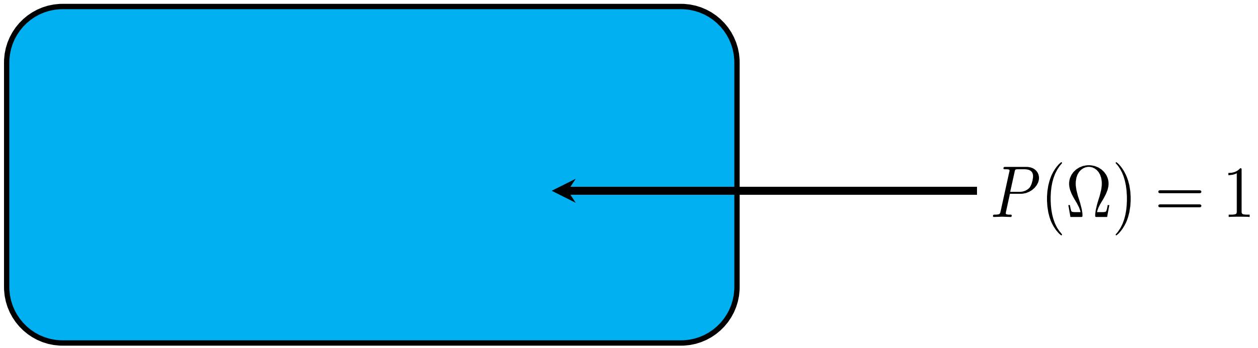
Visualizing the Axioms: I

- Axiom 1: for all events E , $P(E) \geq 0$



Visualizing the Axioms: **II**

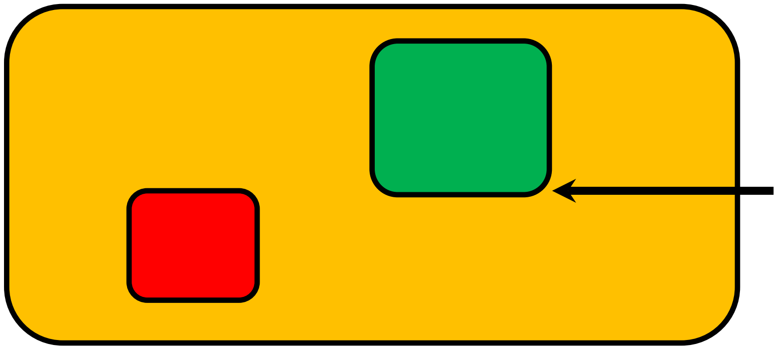
- Axiom 2: $P(\emptyset) = 0, P(\Omega) = 1$



Visualizing the Axioms: **III**

- Axiom 3: disjoint $$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

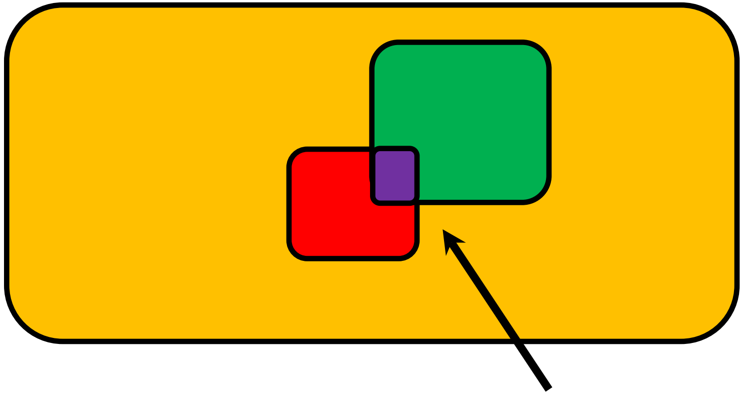
E_1 **or** E_2



$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Visualizing the Axioms

- Also, other laws:



E_1 **and** E_2

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Break & Quiz

- **Q 1.1**: We toss a biased coin. If P(heads) = 0.7, then P(tails) = ?
- A. 0.4
- B. 0.3
- C. 0.6
- D. 0.5

Break & Quiz

- **Q 1.1**: We toss a biased coin. If P(heads) = 0.7, then P(tails) = ?
- A. 0.4
- **B. 0.3**
- C. 0.6
- D. 0.5

Break & Quiz

- **Q 1.2**: There are exactly 3 candidates for a presidential election. We know X has a 30% chance of winning, B has a 35% chance. What's the probability that C wins?
- A. 0.35
- B. 0.23
- C. 0.333
- D. 0.8

Break & Quiz

- **Q 1.2**: There are exactly 3 candidates for a presidential election. We know X has a 30% chance of winning, B has a 35% chance. What's the probability that C wins?
- **A. 0.35**
- B. 0.23
- C. 0.333
- D. 0.8

Break & Quiz

- **Q 1.3**: What’s the probability of selecting a black card or a number 6 from a standard deck of 52 cards?
- A. 26/52
- B. 4/52
- C. 30/52
- D. 28/52

Break & Quiz

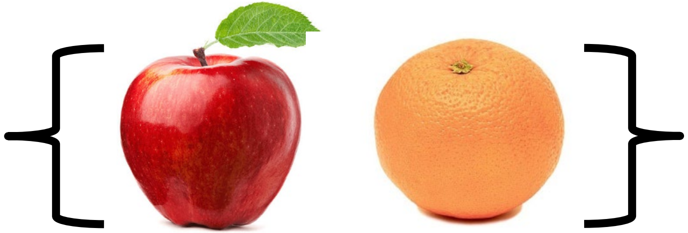
- **Q 1.3**: What's the probability of selecting a black card or a number 6 from a standard deck of 52 cards?
- A. 26/52
- B. 4/52
- C. 30/52
- **D. 28/52**

Basics: **Random Variables**

- Intuitively: a number X that's random
- Mathematically: map random outcomes to real values

$$X : \Omega \to \mathbb{R}$$

- Why?
  - Previously, everything is a set.
  - Real values are easier to work with



Basics: **CDF & PDF**

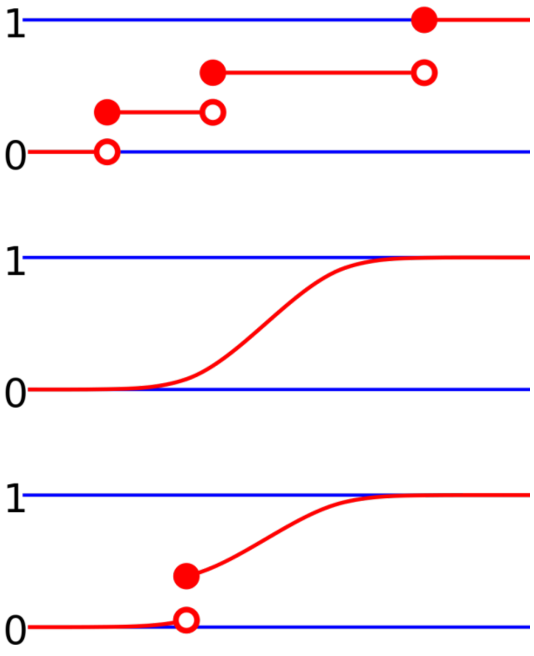
- Can still work with probabilities:

$$P(X = 3)$$

- Cumulative Distribution Func. (CDF)

$$F_X(x) := P(X \leq x)$$

- Density / mass function $p_X(x)$



Wikipedia CDF

Basics: **Expectation** & **Variance**

- Another advantage of RVs are ``summaries’’
- Expectation: $E[X] = \sum_a a \times P(x = a)$
  - The “average”
- Variance: $Var[X] = E[(X - E[X])^2]$
  - A measure of “spread”

Basics: **Joint Distributions**

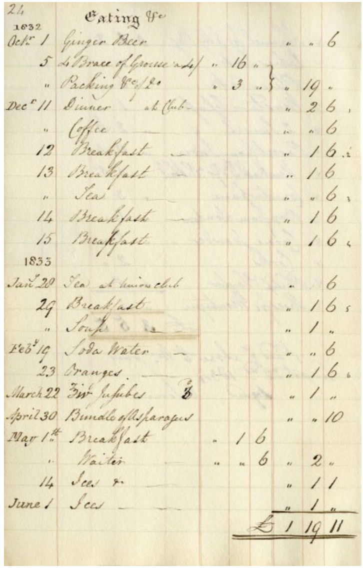
- Move from one variable to several
- Joint distribution: $P(X = a, Y = b)$
  - Why? Work with **multiple** types of uncertainty that correlate with each other

Basics: **Marginal** Probability

- Given a joint distribution $P(X = a, Y = b)$
  - Get the distribution in just one variable:

$$P(X = a) = \sum_b P(X = a, Y = b)$$

  - This is the “marginal” distribution.



Jerry's super blurry camera

- One pixel, 1-bit color sensor (green=trees, white=snow)
- Model T: comes with 1-bit temperature sensor (hot, cold)

Basics: **Marginal** Probability

$$P(X = a) = \sum_b P(X = a, Y = b)$$

	green	white
hot	150/365	45/365
cold	50/365	120/365

$$[P(\text{hot}), P(\text{cold})] = [\frac{195}{365}, \frac{170}{365}]$$

Probability Tables

- Write our distributions as tables

- # of entries? 4.
  - If we have n variables with k values, we get k^n entries
  - **Big!** For a 1080p screen, 12 bit color, size of table: $10^{7490589}$
  - No way of writing down all terms

Independence

- Independence between RVs:

$$P(X, Y) = P(X)P(Y)$$

- Why useful? Go from k^n entries in a table to $\sim kn$
- Expresses joint as **product** of marginals

- requires domain knowledge

Conditional Probability

- For when we know something (i.e. Y=b),

$$P(X = a|Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)}$$

	green	white
hot	150/365	45/365
cold	50/365	120/365

$$P(cold|white) = \frac{P(cold, white)}{P(white)} = \frac{120}{45 + 120} = 0.73$$

Conditional independence

Same as independence, but conditioned on something

- require domain knowledge

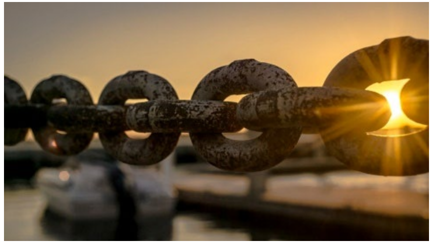
$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Chain Rule

- Apply repeatedly,

$$P(A_1, A_2, \ldots, A_n)$$
$$= P(A_1)P(A_2|A_1)P(A_3|A_2, A_1)\ldots P(A_n|A_{n-1}, \ldots, A_1)$$

- Note: still big!
  - If some **conditional independence**, can factor!
  - Leads to probabilistic graphical models



Break & Quiz

Q 2.1: Given joint distribution table:

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

What is the probability the temperature is hot given the weather is cloudy?

A. 40/365
B. 2/5
C. 3/5
D. 195/365

Break & Quiz

Q 2.1: Back to our joint distribution table:

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

What is the probability the temperature is hot given the weather is cloudy?

A. 40/365
B. 2/5
C. 3/5
D. 195/365

Break & Quiz

Q 2.2: Of a company's employees, 30% are women and 6% are married women. Suppose an employee is selected at random. If the employee selected is a woman, what is the probability that she is married?

A. 0.3
B. 0.06
C. 0.24
D. 0.2

Break & Quiz

Q 2.2: Of a company's employees, 30% are women and 6% are married women. Suppose an employee is selected at random. If the employee selected is a woman, what is the probability that she is married?

A. 0.3
B. 0.06
C. 0.24
D. 0.2

Reasoning With Conditional Distributions

- Evaluating probabilities:
  - Wake up with a sore throat.
  - Do I have the flu?
- Logic approach: $S \rightarrow F$
  - Too strong.
- **Inference**: compute probability given evidence $P(F|S)$
  - Can be much more complex!

Using **Bayes' Rule**

- Want: $P(F|S)$
- **Bayes' Rule:** $P(F|S) = \frac{P(F,S)}{P(S)} = \frac{P(S|F)P(F)}{P(S)}$
- Parts:
  - $P(S) = 0.1$ Sore throat rate
  - $P(F) = 0.01$ Flu rate
  - $P(S|F) = 0.9$ Sore throat rate among flu sufferers

So: $P(F|S) = 0.09$

Using Bayes' Rule

- Interpretation $P(F|S) = 0.09$
  - Much higher chance of flu than normal rate (0.01).
  - Very different from $P(S|F) = 0.9$
    - 90% of folks with flu have a sore throat
    - But, only 9% of folks with a sore throat have flu

- Idea: **update** probabilities from **evidence**

Bayesian **Inference**

- Fancy name for what we just did. Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- *H* is the hypothesis
- *E* is the evidence

Bayesian **Inference**

- Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \longleftarrow \textbf{Prior}$$

- Prior: estimate of the probability **without** evidence

Bayesian Inference

- Terminology:

Likelihood

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- Likelihood: probability of evidence **given a hypothesis**

Bayesian Inference

- Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

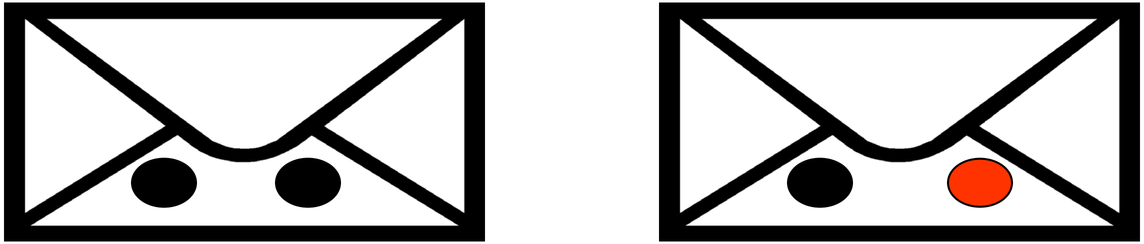
↑

Posterior

- Posterior: probability of hypothesis **given evidence**.

Two Envelopes Problem

- We have two envelopes:
  - E_1 has two black balls, E_2 has one black, one red
  - The **red** one is worth $100. Others, zero
  - Open an envelope, see one ball. Then, can switch (or not).
  - You see a black ball. **Switch?**



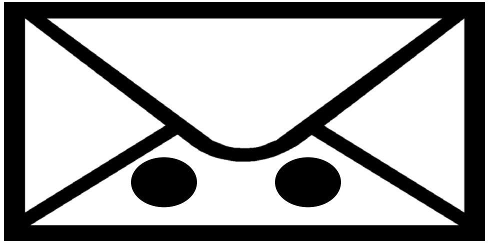
Two Envelopes Solution

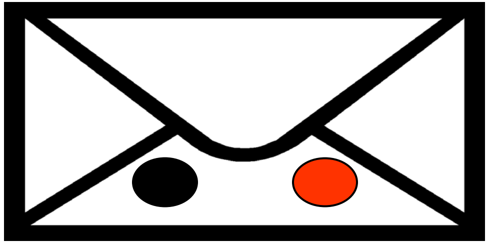
- Let's solve it. $$P(E_1|\text{Black ball}) = \frac{P(\text{Black ball}|E_1)P(E_1)}{P(\text{Black ball})}$$

- Now plug in: $$P(E_1|\text{Black ball}) = \frac{1 \times \frac{1}{2}}{P(\text{Black ball})}$$

$$P(E_2|\text{Black ball}) = \frac{\frac{1}{2} \times \frac{1}{2}}{P(\text{Black ball})}$$

So switch!





Naïve Bayes

- Conditional Probability & Bayes:

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1, \ldots, E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

- If we further make the **conditional independence assumption (a.k.a. Naïve Bayes)**

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1|H)P(E_2|H) \ \cdots \ P(E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

Naïve Bayes

- Expression

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1|H)P(E_2|H)\cdots, P(E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

- *H*: some class we'd like to infer from evidence
  - We know prior $P(H)$
  - Estimate $P(E_i|H)$ from data! ("training")
  - Very similar to envelopes problem.

Break & Quiz

Q 3.1: 50% of emails are spam. Software has been applied to filter spam. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a nonspam email?

A. 5/104
B. 95/100
C. 1/100
D. 1/2

Break & Quiz

Q 3.1: 50% of emails are spam. Software has been applied to filter spam. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a nonspam email?

A. 5/104
B. 95/100
C. 1/100
D. 1/2

S : Spam
NS: Not Spam
DS: Detected as Spam

P(S) = 50 % spam email
P(NS) = 50% not spam email
P(DS|NS) = 5% false positive, detected as spam but not spam
P(DS|S) = 99% detected as spam and it is spam

Applying Bayes Rule
P(NS|DS) = (P(DS|NS)*P(NS)) / P(DS) = (P(DS|NS)*P(NS)) / (P(DS|NS)*P(NS) + P(DS|S)*P(S)) = 5/104

Break & Quiz

Q 3.2: A fair coin is tossed three times. Find the probability of getting 2 heads and a tail

A. 1/8
B. 2/8
C. 3/8
D. 5/8

Break & Quiz

Q 3.2: A fair coin is tossed three times. Find the probability of getting 2 heads and a tail

A. 1/8
B. 2/8
C. 3/8
D. 5/8

S = { H H H, H H T, H T H, H T T, T H H, T H T, T T H, T T T}
P(2H, 1T)=(1/8) + (1/8) + (1/8) = 3/8

Readings

- Vast literature on intro probability and statistics.
- Local classes: **Math/Stat 431**

- **Suggested reading**:

  Probability and Statistics: The Science of Uncertainty,

  Michael J. Evans and Jeff S. Rosenthal

  http://www.utstat.toronto.edu/mikevans/jeffrosenthal/book.pdf

  (Chapters 1-3, excluding "advanced" sections)