CS 540 Introduction to Artificial Intelligence
Probability

University of Wisconsin-Madison

Spring 2026 Sections 1 & 2

Outline

- Probability
  - Basics: definitions and axioms
  - Random Variables (RVs) and 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

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

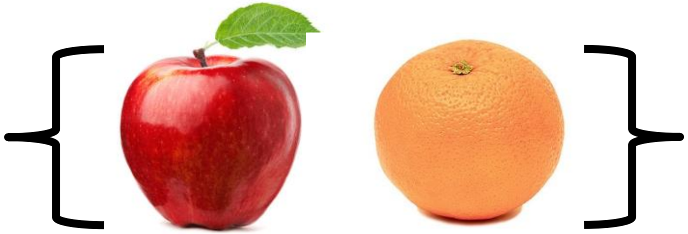
Basics: **Random Variables**

- Intuitively: a number X that's random
- Mathematically:

function that maps random outcomes to real values

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

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



Basics: **Random Variables**

$$\longrightarrow \quad X = 1 \quad \longrightarrow \quad P(X = 1)$$

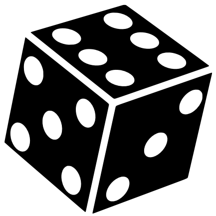
$$\longrightarrow \quad X = 2 \quad \longrightarrow \quad P(X = 2)$$

$$\longrightarrow \quad X = 3 \quad \longrightarrow \quad P(X = 3)$$

Basics: **CDF**

Cumulative Distribution Function (CDF)

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

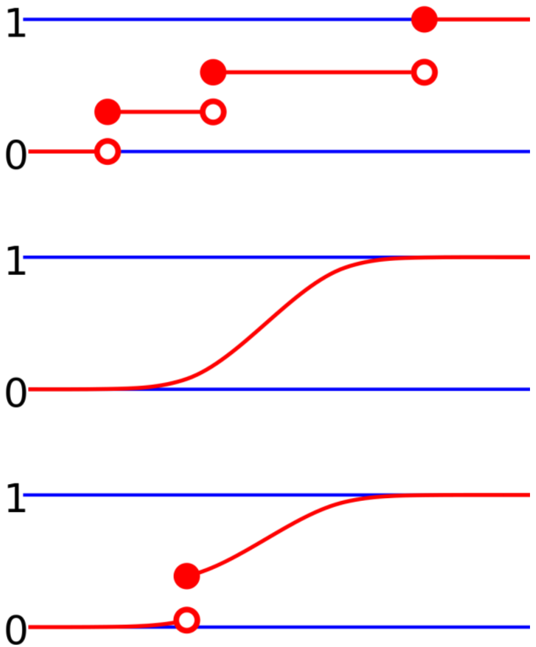


$$F_X(3) = 0.5$$
$$F_X(6) = 1$$

CDF for discrete probability distribution

CDF for continuous probability distribution

CDF for probability distribution with both discrete and continuous parts

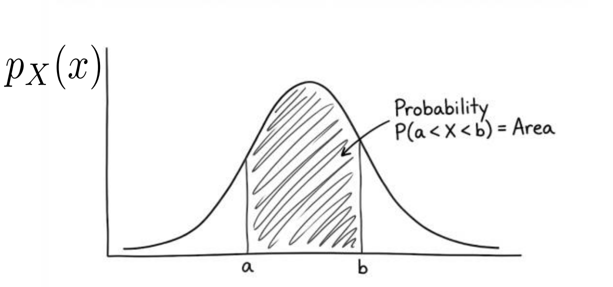


Wikipedia CDF

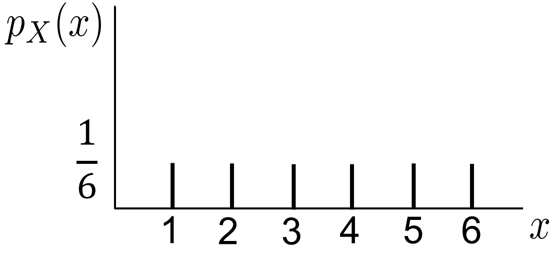
Basics: **PDF/PMF**

Probability density / mass function $p_X(x)$:

A mathematical function that tells you how likely different outcomes are.



example of a continuous
probability density function



example of a discrete
probability mass function

Basics: **Expectation**

Another advantage of RVs are ``summaries''

- Expectation:
  - The "average" $E[X] = \sum_a a \times P(X = a)$
- Example of a single toss of a fair coin:
  - **Success (Heads)** is assigned the value 1.
  - **Failure (Tails)** is assigned the value 0.

$$E(X) = \big(1 \times P(X = 1)\big) + \big(0 \times P(X = 0)\big)$$
$$E(X) = (1 \times 0.5) + (0 \times 0.5)$$
$$E(X) = 0.5$$

Basics: **Variance**

- Variance:
  - A measure of "spread"

$$Var[X] = E[(X - E[X])^2]$$

- Example of a single toss of a fair coin:

$$\text{Var}(X) = \big((1 - E[X])^2 \times \; P(X = 1)\big) + \big((0 - E[X])^2 \times \; P(X = 0)\big)$$

$$E(X) = (0.25 \times 0.5) + (0.25 \times 0.5)$$

$$E(X) = 0.25$$

Break & Quiz

Q 1.1: Consider a fair six-sided die where the probability of landing on any specific face (1, 2, 3, 4, 5, or 6) is exactly $P(x) = \frac{1}{6}$ Based on these probabilities, what is the expected value $E[X]$ for a single roll?:

A. 3.0
B. 3.5
C. 4.0
D. 21.0

Break & Quiz

Q 1.1: Consider a fair six-sided die where the probability of landing on any specific face (1, 2, 3, 4, 5, or 6) is exactly $P(x) = \frac{1}{6}$ Based on these probabilities, what is the expected value $E[X]$ for a single roll?:

A. 3.0
B. 3.5
C. 4.0
D. 21.0

$$E(X) = \left(1 \times \frac{1}{6}\right) + \left(2 \times \frac{1}{6}\right) + \left(3 \times \frac{1}{6}\right) + \left(4 \times \frac{1}{6}\right) + \left(5 \times \frac{1}{6}\right) + \left(6 \times \frac{1}{6}\right) = 3.5$$

Basics: **Joint Distributions**

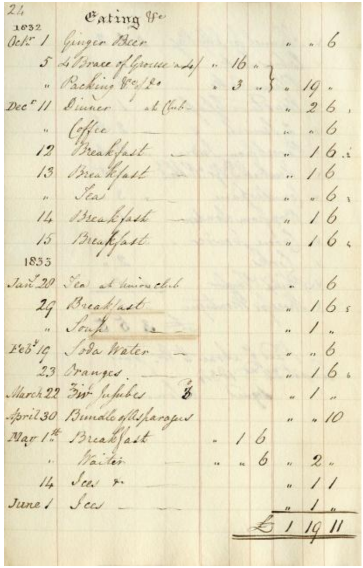
- Move from one variable to several
- Joint distribution: $P(X = a, Y = b)$
  - Why? Work with **multiple** types of uncertainty that correlate with each other

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.


24
Eating &c

1832			
Octr 1	Ginger Beer		" " 6
5	2 Brace of Grouse @ 4/	" 16 "	
"	Packing &c of do	" 3 "	" 19 "
Decr 11	Dinner at Club		" 2 6
"	Coffee		" " 6
12	Breakfast		" 1 6
13	Breakfast		" 1 6
"	Tea		" " 6
14	Breakfast		" 1 6
15	Breakfast		" 1 6
1833			
Jany 28	Tea at Union club		" 6
29	Breakfast		" 1 6
"	Soup		" 1 "
Feby 19	Soda Water		" " 6
23	Oranges		" 1 6
March 22	3iv Jujubes		" 1 "
April 30	Bundle of Asparagus		" " 10
May 1st	Breakfast	" 1 6	
"	Waiter	" " 6	" 2 "
14	Ices &c		" 1 1
June 1	Ices		" 1 "
			£ 1 19 11

Example: 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:

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

- Example: simultaneously toss a coin and roll a die
- 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

– It requires 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!

Chain Rule

Example drawing 3 Aces from a 52-card deck:

- Event A_1: The 1st card is an Ace.
- Event A_2: The 2nd card is an Ace.
- Event A_3: The 3rd card is an Ace.

$$P(A_1, A_2, A_3) = P(A_1)P(A_2 | A_1)P(A_3 | A_1, A_2)$$

Chain Rule

- Probability of the 1st Ace: $P(A_1) = \frac{4}{52}$
- Probability of the 2nd Ace : $P(A_2|A_1) = \frac{3}{51}$
- Probability of the 3rd Ace: $P(A_3|A_1, A_2) = \frac{2}{50}$
- Probability of drawing 3 Aces:

$$P(A_1, A_2, A_3) = P(A_1)P(A_2|\, A_1)P(A_3|A_1, A_2)$$

$$P(A_1, A_2, A_3) = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} = \frac{24}{132600} \approx 0.00018$$

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

Bayes' Rule

Theorem: For any events A and B we have

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

Proof: Apply the chain rule two different ways:

$$\left.\begin{aligned} P(A,B) &= P(\,A \mid B) \ \cdot P(B) \\ &= \ P(B \mid A) \cdot P(A) \end{aligned}\right\} \quad P(A|B) = \frac{P(\,B \mid A) \ \cdot P(A)}{P(B)}$$

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:

$$P(H|E) = \frac{P(E|H)\color{red}{P(H)}}{P(E)}$$ ⟵ **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:

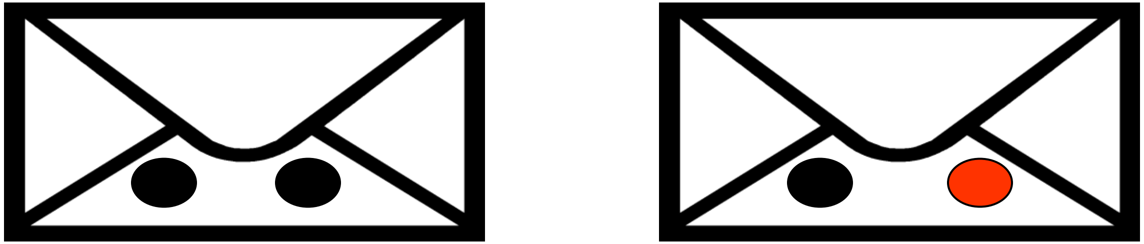
$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

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?**



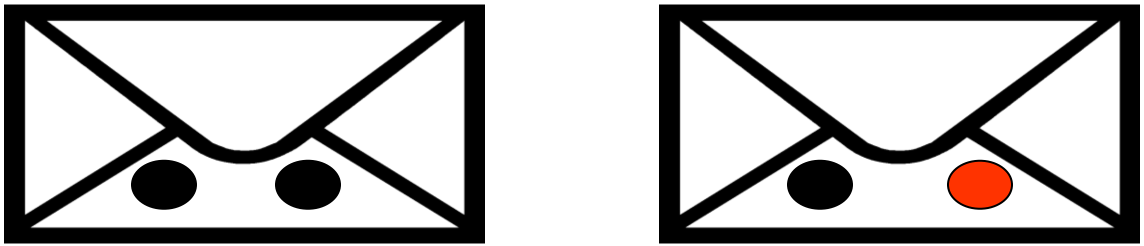
Two Envelopes Solution

- 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(DS,S)) = (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

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)