CS 540 Introduction to Artificial Intelligence
Probability

University of Wisconsin-Madison

Fall 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

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

$$P(\{1\}) = \frac{1}{6}, \qquad P(\{1,2,3\}) = \frac{3}{6} = \frac{1}{2}$$

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

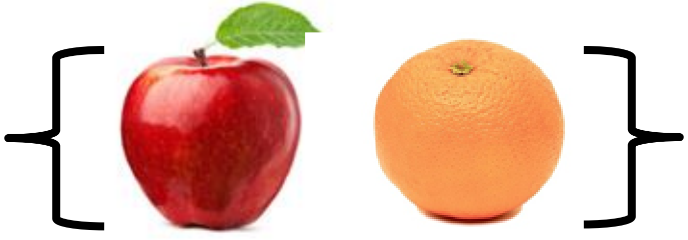
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Basics: **Random Variables**

- Intuitively: a number X that's random
- Mathematical definition: *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
  - Can define CDFs, expectations, variance, etc.



Basics: **Random Variables**

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

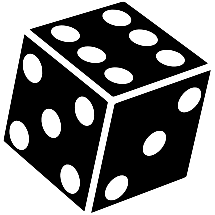
$X = 2 \longrightarrow P(X = 2)$

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

Basics: **CDF**

Cumulative Distribution Function (CDF)

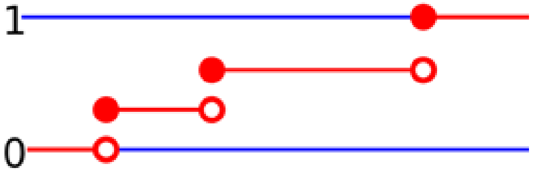
$$F_X(x) := P(X \leq x)$$



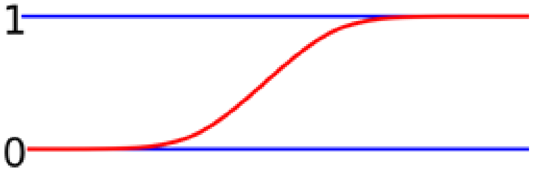
$$F_X(3) = 0.5$$

$$F_X(6) = 1$$

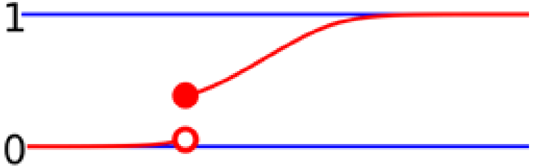
CDF for discrete probability distribution



CDF for continuous probability distribution



CDF for probability distribution with both discrete and continuous parts

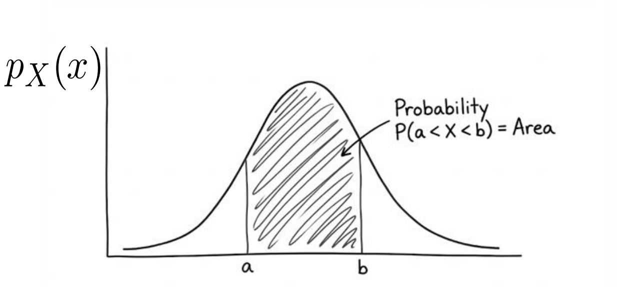


Wikipedia CDF

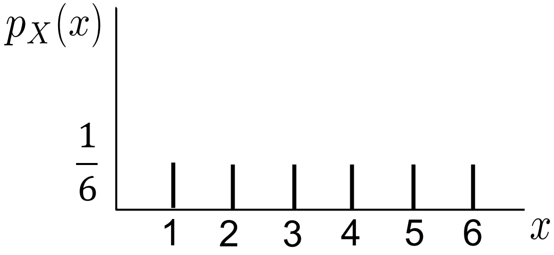
Basics: **PDF/PMF**

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

A function which 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 of a RV X : $E[X] = \sum_a a\, P(X = a)$
- Expectation of a function f of a RV X :

$$E[f(X)] = \sum_a f(a)\, P(X = a)$$

- Example of a single toss of a fair coin:
  - **Success (Heads)** is assigned the value 1.
  - **Failure (Tails)** is assigned the value 0.

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

Basics: **Variance**

- Variance:
  - A measure of "spread"

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

- Example of a single toss of a fair coin:

$$\text{Var}(X) = \big((1 - E[X])^2 \times\ P(X = 1)\big) + \big((0 - E[X])^2 \times\ P(X = 0)\big)$$
$$= (0.25 \times 0.5) + \ (0.25 \times 0.5) = 0.25.$$

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
    Eg: temperature and weather

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.

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

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

Independence

- Independence between RVs:

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

- Example: simultaneously toss a coin and roll a die. Then, $P(H, 6) = P(H)P(6)$
- Can express joint distribution as **product** of marginals

Probability Tables

- Write our distributions as tables

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

- How many entries?
  - If we have n variables with k values, we get k^n entries
  - **Big!**
  - Storing all terms in memory is near impossible.

- How many entries under independence?
  - Only need to record k entries per variable: nk

Conditional Probability

Reasoning under uncertainty **when we already know something**

(Y=b in the example below)

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

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

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

Conditional independence

Same as independence, but conditioned on something (an event, random variable etc)

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

Example: Ice-cream sales and drownings are correlated, but conditionally independent given the weather.

Chain Rule

- Apply conditional probability repeatedly,

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

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:

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

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

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:

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

Law of total probability

Let $B_1, B_2, \dots, B_n$ be a partition. That is, they are **mutually exclusive** (cannot happen at the same time) and **exhaustive** (cover all possibilities).

Then, $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$

Reasoning With Conditional Distributions

- Evaluating probabilities:
  - Wake up with a sore throat.
  - Do I have the flu?
- Logic approach: Sore throat $\Rightarrow$ Flu or Sore throat $\not\Rightarrow$ Flu
  - Too strong/restrictive to be practical
- **Inference**: compute probability given evidence $P(F|S)$

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:

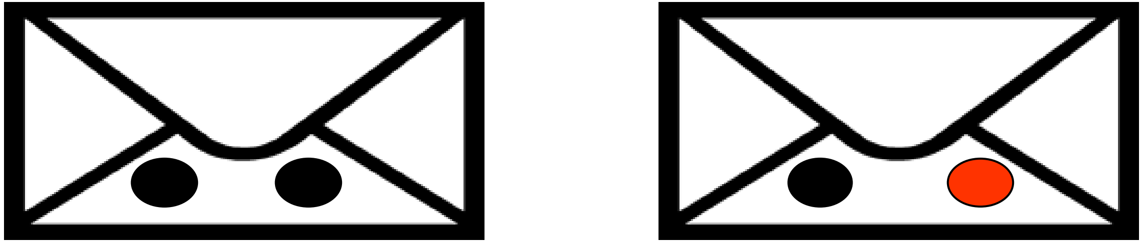
$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

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
  - If you choose the envelope with the red ball, you get \$100.
  - You get to choose an envelope, see one ball, and can then decide to switch if you wish you.
  - You see a black ball. **Do you switch?**



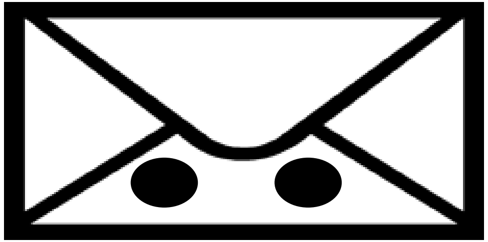
Two Envelopes Solution

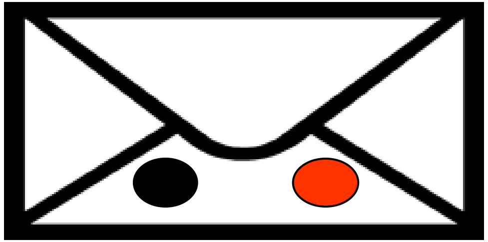
- Let's solve it. $$P(E_1|\text{Black ball}) = \frac{P(\text{Black ball}|E_1)P(E_1)}{P(\text{Black ball})}$$

- 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:

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

- **The naïve Bayes assumption** (a strong conditional independence assumption):

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

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")

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)