STA 35C: Statistical Data Science III

Lecture 2: Probability Review

Dogyoon Song

Spring 2025, UC Davis

Announcements

Office hours:

- Instructor: Wed 3:30–4:30 PM (MSB 4220)
- TA: Tue 12:30–2:30 PM (MSB 1143)

* Please also use Piazza for course questions

Important dates:

- Midterm 1: Fri, Apr 24 (in class)
- Midterm 2: Fri, May 15 (in class)
- Final: Fri, Jun 5, 1:00–3:00 PM

* *No make-up exams can be arranged other than SDC accommodations*

SDC accommodations: If you need accommodations, please submit requests through the Student Disability Center (SDC) as early as possible

* See the course webpage and syllabus for more details and additional information

Announcements: Homework 1

Homework 1 is now posted

- Due: Tue, April 7, 11:59 PM PT
    - Late submissions will not be accepted for any reason and will receive 0 points
- Submission instructions:
    - Upload a ***single PDF file*** to Canvas (*Assignments* → *Homework 1*)
    - Name the file using the prefix of your UC Davis email ID and homework number (e.g., `dgsong_hw1.pdf`)
    - Please make sure to include "STA 35C," your name, and the last four digits of your student ID on the front page
    - For coding problems, prepare your solutions in R Markdown; for non-coding problems, you may typeset your solutions in LaTeX (preferred), use a word processor, or handwrite and scan them; in all cases, make sure your work is legible and clearly organized

* See the syllabus for more details about homework policy and requirements

Agenda[1]

- Probability basics
  - Set theory
  - Probabilistic models
- Conditional probability
- Bayes' theorem

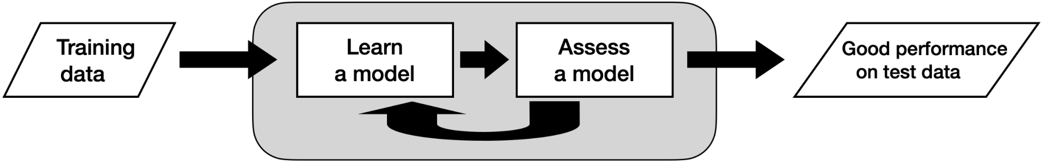
[1] Most of today's topics were covered in STA 35A; see Lectures 13–15

Motivation: Why probability?

Recall the goals of statistical learning:

- Predict Y from X (learn a function $f : \mathcal{X} \to \mathcal{Y}$)
- Identify patterns in data X

Standard workflow:



Key challenge:

- We aim for good predictions or insights on ***new, unseen* data**
- How can we reason about performance on future data, not just on the training data?

Solution: Probabilistic tools and viewpoints offer a formal way to handle uncertainty

Formalizing probability: Sample space and events

Ingredients to formalize probability:

- **Outcome:** a possible result of an experiment or trial
- **Sample space**: the set of all possible outcomes, often denoted by Ω
  - e.g., $\{H, T\}$, $\{1, 2, 3, 4, 5, 6\}$
- **Event:** a subset of Ω
  - e.g., for $\Omega = \{H, T\}$: $\emptyset$, $\{H\}$, $\{T\}$, $\{H, T\}$
  - e.g., for $\Omega = \{1, 2, 3, 4, 5, 6\}$: $\{6\}$, $\{1, 2\}$, $\{2, 4, 6\}$, ...
- **Probability[2]:** a map P that assigns each event A a number $P(A) \in [0, 1]$
  - We will shortly state the axioms that such assignments must satisfy

Probability theory makes extensive use of set notation and set operations!

[2] A formal, mathematically rigorous definition of probability measure is beyond the scope of STA 35C

Set theory: Notation and terminology

A **set** is a collection of distinct objects, called the **elements** of the set

- $x \in S$: x is an element of S
- $x \notin S$: x is not an element of S

The **empty set** is a set having no elements, denoted by $\emptyset$

Sets can be specified in various ways

- Roster notation (enumeration): $S = \{x_1, x_2, \ldots, x_n\}$, or $S = \{x_1, x_2, \ldots\}$

$$\text{e.g.,} \qquad \{H, T\}, \qquad \{1, 2, 3, 4, 5, 6\}, \qquad \{2, 4, 6, 8, \ldots\}$$

- Set-builder notation (logical formula): $S = \{x \mid x \text{ satisfies } Q\}$

$$\text{e.g.,} \qquad \{2k \mid k \text{ is a positive integer}\}, \qquad \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$$

Sets may be finite, countably infinite, or uncountable

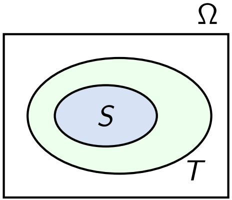
Set theory: Inclusion relations

S is a **subset** of T if every element of S is an element of T, i.e., $x \in S \implies x \in T$

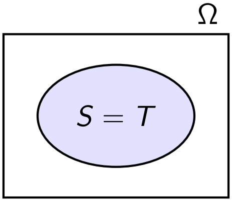
- We write $S \subseteq T$ (or $S \subset T$)
- Equivalently, T is a **superset** of S, denoted by $T \supseteq S$ (or $T \supset S$)

The two sets are **equal**, written $S = T$, if $S \subseteq T$ and $T \subseteq S$

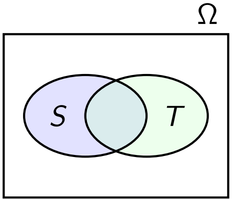
S is a **proper (strict)** subset of T, denoted by $S \subsetneq T$, if $S \subseteq T$ and $S \neq T$



$S \subsetneq T$



$S = T$



Neither $S \subseteq T$ nor $T \subseteq S$

Set theory: Set operations

Often, it is convenient to introduce the **universal set** Ω, containing all objects of interest

- $S^c = \{x \in \Omega \mid x \notin S\}$ is the **complement** of S with respect to Ω

Given two sets S and T,

- $S \cup T = \{x \mid x \in S \text{ or } x \in T\}$ is their **union**
- $S \cap T = \{x \mid x \in S \text{ and } x \in T\}$ is their **intersection**

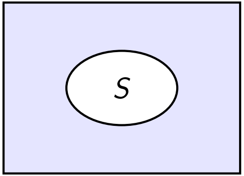
Two sets S and T are **disjoint** if $S \cap T = \emptyset$

- A collection of sets is **pairwise disjoint** if no two distinct sets have a common element

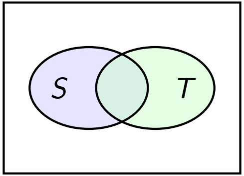
A collection of sets $\{S_1, S_2, \ldots\}$ is a **partition** of S if

- $\bigcup_n S_n = S$ and
- the sets $S_1, S_2, \ldots$ are pairwise disjoint

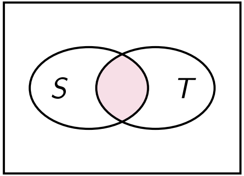
Illustration with Venn diagrams: Set operations



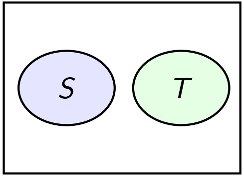
S^c



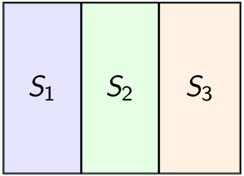
$S \cup T$



$S \cap T$



Disjoint



Partition of Ω

Set theory: The algebra of sets

Set operations have several properties, which are direct consequences of the definitions

- Commutativity:
$$S \cup T = T \cup S, \qquad S \cap T = T \cap S$$
- Associativity:
$$S \cup (T \cup U) = (S \cup T) \cup U, \qquad S \cap (T \cap U) = (S \cap T) \cap U$$
- Distributivity:
$$S \cap (T \cup U) = (S \cap T) \cup (S \cap U), \qquad S \cup (T \cap U) = (S \cup T) \cap (S \cup U)$$
- ... and more:
$$(S^c)^c = S, \qquad S \cap S^c = \emptyset, \qquad S \cup S^c = \Omega, \qquad S \cup \Omega = \Omega, \qquad S \cap \Omega = S$$

Elements of a probabilistic model

A **probabilistic model** is a mathematical description of an uncertain situation

- **Experiment**: an underlying process producing exactly one out of several possible outcomes

A probabilistic model consists of a sample space Ω and a probability law P

- **Sample space** Ω: the set of all possible outcomes
  - Event[3]: for this course, think of an event as any subset of Ω
- **Probability law** P: a map that assigns a number $P(A)$ encoding our knowledge or belief about the collective likelihood of the event A, satisfying *certain axioms*

Next we state the axioms that make P a valid probability law

[3] Strictly speaking, some sets have to be excluded, which involves measure theory. However, we can safely ignore pathological issues in this course.

Probability laws

Once the sample space Ω has been chosen, a probability law assigns to each event A a number $P(A)$, called the **probability** of A, subject to three axioms:

1. **Nonnegativity:** $P(A) \geq 0$ for every event A
2. **(Countable) Additivity:** if A and B are disjoint events, then

   $$P(A \cup B) = P(A) + P(B)$$

   More generally, if $A_1, A_2, \ldots$ are pairwise disjoint, then

   $$P\left(\bigcup_n A_n\right) = \sum_n P(A_n)$$

3. **Normalization:** $P(\Omega) = 1$

Some properties of probability laws

There are many natural properties of a probability law, derivable from the axioms

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$
- If $A \subseteq B$, then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leq P(A) + P(B)$ (union bound)
- $P(B \setminus A) = P(B) - P(A \cap B)$ $\quad B \setminus A := B \cap A^c$
- ...

It is often useful to visualize and verify these using a Venn diagram

A quick exercise: A die roll example

Setup:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $P(\{1\}) = P(\{2\}) = \cdots = P(\{6\}) = 1/6$
- $A = \{2, 3, 5\}$ (prime faces)
- $B = \{2, 4, 6\}$ (even faces)

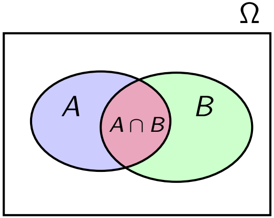
Questions:

- Draw a Venn diagram to visualize Ω, A and B
- Identify $A \cup B$, $A \cap B$ and $A \setminus B$ on the Venn diagram
- Compute $P(A \cup B)$, $P(A \cap B)$, and $P(A \setminus B)$

Conditional probability: Definition

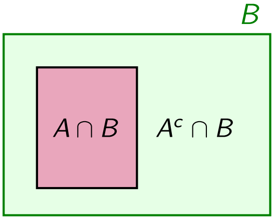
For an event[4] B with $P(B) > 0$, the **conditional probability** of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



Ordinary view inside Ω

$\Longrightarrow$ condition on B



Restrict attention to outcomes inside B

Example: Compare $P(A)$ vs $P(A \mid B)$ in the die roll example on the previous slide

- $P(A) = 3/6 = 1/2$, whereas $P(A \mid B) = 1/3$, since within $B = \{2, 4, 6\}$ only 2 is prime

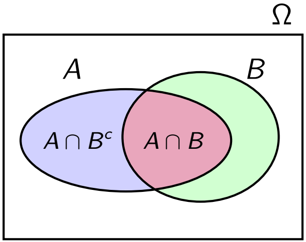
[4] In this course, we only condition on events with positive probability

Conditional probability: Key identities

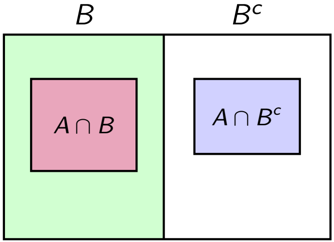
Two key identities for conditional probability:

- **Multiplication rule:** $P(A \cap B) = P(B)P(A \mid B)$
- **Law of total probability:** If $0 < P(B) < 1$, then

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) \\ &= P(A \mid B)P(B) + P(A \mid B^c)P(B^c) \end{aligned}$$



partition by B $\Longrightarrow$



A is split by the partition $\{B, B^c\}$

$A = (A \cap B) \cup (A \cap B^c)$ disjointly

Independence

Events A and B are **independent** if $P(A \cap B) = P(A)P(B)$

- Recall that $P(A \cap B) = P(B)P(A \mid B)$
- Thus, when $P(B) > 0$, independence implies $P(A \mid B) = P(A)$
  - That is, knowing that B occurred does not change the probability that A occurs
- Similarly, when $P(A) > 0$, independence implies $P(B \mid A) = P(B)$

Example: Flipping a coin and rolling a die

- Knowing the coin was heads does not help determine the outcome of a die roll

Counter-example: Seeing someone with an umbrella and the day being rainy are not independent

- If we see someone with an umbrella, it is more likely to be a rainy day

Bayes' theorem

Often, we know $P(B \mid A)$, but want $P(A \mid B)$

- A: a hypothesis/model/state of the world
- B: observed data or evidence
- Example: A=having cancer (or not), B=positive (negative) screening result

Assuming that we know

- *prior:* $P(A)$ and $P(A^c)$
- *likelihood:* $P(B \mid A)$ and $P(B \mid A^c)$

we update our belief about A after observing B

Bayes' theorem states that

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \quad \text{where} \quad P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$$

* *Posterior* = *Likelihood* × *Prior* / *Evidence*

Bayes' theorem: Examples

Example: Let A=cancer and B=positive screening result

- Suppose $P(A) = 0.01$
- $P(B \mid A) = 0.8$ (true positive)
- $P(B \mid A^c) = 0.1$ (false positive; positive screening though a person does not have cancer)

What is $P(A \mid B)$? How does observing B affect our "belief" on A?

$$P(A \mid B) = \frac{0.8(0.01)}{0.8(0.01) + 0.1(0.99)} \approx 0.075.$$

* A positive result raises the probability from 1% to about 7.5%, but it is still far from certain.

Food for thought: Let A=a model (or a set of models) and B=observed data

Example: A gambler is deciding whether a coin is fair ($\mathrm{Bernoulli}(1/2)$) or double-headed ($\mathrm{Bernoulli}(1)$); after tosses, Bayes' theorem updates the probability of each model

Wrap-up

- Events are sets, so set theory provides the basic language of probability
- A probabilistic model consists of a sample space Ω and a probability law P such that
  - $P(A) \geq 0$
  - $P\left(\bigcup_n A_n\right) = \sum_n P(A_n)$
  - $P(\Omega) = 1$
- Conditional probability $P(A \mid B)$ means we restrict attention to outcomes inside B
- Independence means conditioning on one event does not change the probability of the other
- The law of total probability decomposes $P(A)$, and Bayes' theorem reverses conditioning to update beliefs using data (observations)