# Probability Theory and Bayes

Readings:

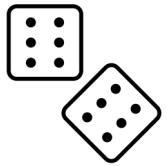
[KPM] 2.1 through example 2.2.3.1

Optional [JA6] 1.3 – 1.6

# Contents

- basic probability theory
- discrete and continuous outcomes
- union bound
- total probability
- Bayes' theorem
- conditional probabilities
- independence

# Probability Theory



- mathematical rules for dealing with uncertainty or randomness.

- basic elements of probability theory

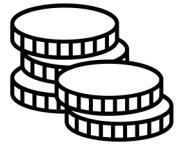
  1. sample space $\Omega$: the set of all possible outcomes of a random experiment
     - a sample space is *discrete* if $\Omega$ is *countable*
     - a sample space is *continuous* if $\Omega$ is *uncountably infinite*
  2. set of events $\mathcal{F}$: A collection of subsets of $\Omega$ denoted $A_1, A_2, \ldots$.
     - an event $A \subset \Omega$ occurs if the outcome $\omega \in A$
  3. probability measure $\mathbb{P}(\cdot)$: a function that maps events to $[0, 1]$ and satisfies **three axioms**:
     1. $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$
     2. $\mathbb{P}(\Omega) = 1$
     3. $A_i \bigcap A_j = \emptyset$ for $i \neq j$ then $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$

pairwise disjoint

- probability space consists of $(\Omega, \mathcal{F}, \mathbb{P})$

Mother nature chooses ω, but we often can't know ω. We can only know if $\omega \in A$ for $A \in \mathcal{F}$. And we know $\mathbb{P}(A)$.

# Discrete Example



- example
  - flipping a coin: $\Omega = \{H, T\}$

- $\mathcal{F}$ can be the set of all subsets of $\Omega$, i.e, the powerset of $\Omega$
  - flipping a coin: $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$

The probability $\mathbb{P}(A)$ is defined for all $A \in \mathcal{F}$. If $A \notin \mathcal{F}$, then $\mathbb{P}(A)$ may not be defined.
When the sample space $\Omega$ is finite or countably infinite, $\mathcal{F}$ is usually taken to be the power set.
In this case, $\{\omega\} \in \mathcal{F}$ for all $\omega \in \Omega$.

- requirements: $\mathbb{P}(\{\omega\}) \geq 0$ for all $\omega \in \Omega$, $\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1$
- flipping a coin:

$$\mathbb{P}(\{H\}) = p \qquad \mathbb{P}(\{T\}) = 1 - p$$

# More discrete examples

- rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- flipping a coin $n$ times: $\Omega = \{H, T\}^n$
- flipping a coin until the first head appears: $\Omega = \{H, TH, TTH, TTTH, \dots\}$
- asking someone to draw a '7', converting it a $28 \times 28$ black and white image: $\Omega = \{0, 1\}^{28 \times 28}$
- asking someone to say the word 'Alexa', and storing 1 second of 4 bit audio sampled at 10k samples per second: $\Omega = \{0, 1, \dots, 15\}^{10{,}000}$

# Continuous Example

- a sample space is continuous if $\Omega$ is uncountably infinite
- example
  - a random number between 0 and 1: $\Omega = (0, 1]$
  - `np.random.rand()` (finite precision, but useful to think of as continuous)
- can not longer define $\mathcal{F}$ as the powerset of $\Omega$
  - $\mathcal{F}$ must be a *sigma-algebra*:

    *i)* $\emptyset \in \mathcal{F}$

    *ii)* $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$

    *iii)* $A_1, A_2, ... \in \mathcal{F}$ implies that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$
  - *we'll except these complications and move on*
- when continuous, we can no longer just assign probabilities to each outcome in the sample space
  - for a *uniform* random number in $(0, 1]$:

$$\mathbb{P}((a, b)) = b - a \qquad \text{to all intervals}$$

# Consequences of the axioms of prob.

Prob. of a complement

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

monotonicity

- if $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

- union bound: a.k.a. countable sub-additivity
  - $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
  - $\mathbb{P}(\bigcup_{i=1}^{n} A_i) \leq \sum_i \mathbb{P}(A_i)$

- law of total probability:
  - if $A_1, A_2, \ldots A_n$ partition $\Omega$ then

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(A_i \cap B)$$

- $A_1, A_2, \ldots A_n$ partition $\Omega$ if $\bigcup_{i=1}^{n} A_i = \Omega$ and $A_i \cap A_j = \emptyset$ for $i \neq j$ , and $A_i \neq \emptyset$

# Conditional Probability and Bayes

- the conditional probability of an event $A$ given $B$ is defined as:

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

  - note $\mathbb{P}(B) \neq 0$ for this to make sense
  - $\mathbb{P}(A \cap B) = \mathbb{P}(A, B)$

- Bayes' theorem:

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

- *prior probability* of an event $A$ is just the unconditional probability $\mathbb{P}(A)$
- *posterior probability* of an event $A$ given $B$ is the conditional probability $\mathbb{P}(A|B)$

# Bayes and total probability

- $A_1, A_2, \ldots A_n$ partition $\Omega$
- know the prior $\mathbb{P}(A_i)$
- also know $\mathbb{P}(B|A_i)$
- how can we find the posterior $\mathbb{P}(A_i|B)$?

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\mathbb{P}(B)}$$

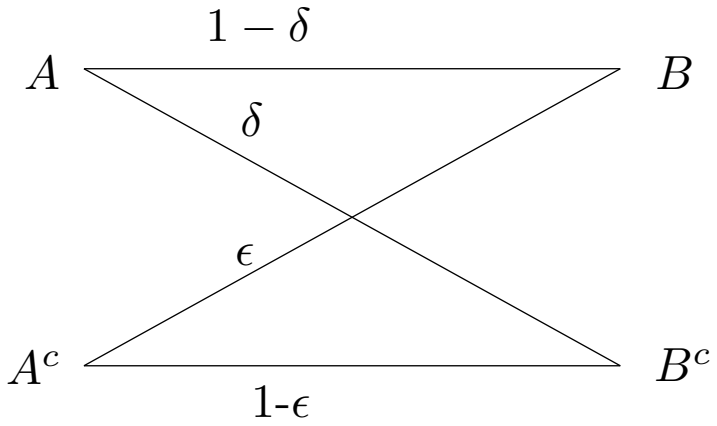
$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(A_i \cap B) \ = \sum_{i=1}^{n} \mathbb{P}(A_i)\mathbb{P}(B|A_i)$$

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(A_i)\mathbb{P}(B|A_i)}$$

# Example

event $A \Leftrightarrow$ 1s audio is 'Siri'

event $B \Leftrightarrow$ audio is classified as 'Siri'




- let $\mathbb{P}(A) = 1/100$
- $\mathbb{P}(B)$?

- $\mathbb{P}(A|B)$?

# Independence of events

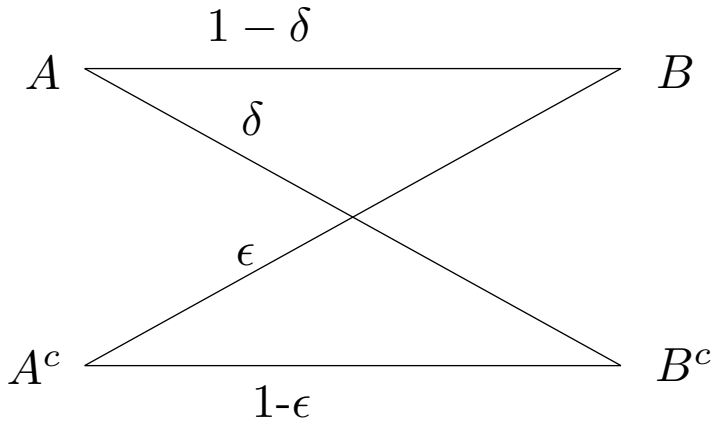
- two events are said to be *statistically independent* if and only if

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

- or equivalently

$$\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$$

- example: imagine the 'Siri' classifier is used twice on two independent 1 second recordings
  - define the events $E_1 = \{\text{first classification is in error}\}$
    $E_2 = \{\text{second classification is in error}\}$
  - $\mathbb{P}(E_1, E_2)$?



# Independence of events

- in general, events $A_1, A_2, \ldots A_n$ are *independent* if and only if

$$\mathbb{P}(A_{i_1}, A_{i_2}, \ldots, A_{i_k}) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}) \quad \text{for all subsets} \quad A_{i_1}, A_{i_2}, \ldots, A_{i_k}$$

- note: not enough to have $\mathbb{P}(A_1, A_2, \ldots A_n) = \prod_{i=1}^{n} \mathbb{P}(A_i)$

- example: roll two fair dice and define

$$A = \{\text{first is 1, 2 or 3}\}$$
$$B = \{\text{first is 2, 3 or 6}\}$$
$$C = \{\text{sum of outcomes is 9}\}$$

- $\mathbb{P}(A) = \frac{1}{2}$, $\mathbb{P}(B) = \frac{1}{2}$
- 36 fair outcomes, 4 outcomes add to 9
- $\mathbb{P}(C) = \frac{1}{9}$

- since $A \cap B \cap C = \{(3, 6)\}$

$$\mathbb{P}(A, B, C) = \frac{1}{36} \quad = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

- $\mathbb{P}(A, B) = \frac{2}{6} \neq \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B)$

A, B, C are not independent!