National Ph.D. Program in *Artificial Intelligence for Society*

Statistics for Machine Learning

Lesson 01 - Probabilities and independence

Andrea Pugnana, Salvatore Ruggieri

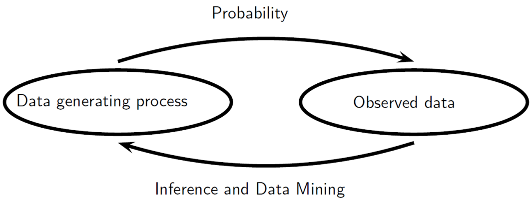
Department of Computer Science
University of Pisa, Italy
andrea.pugnana@di.unipi.it salvatore.ruggieri@unipi.it

Quick introduction

- Format: 20 hours
- Teachers:
  - Prof. Salvatore Ruggieri
  - Dott. Andrea Pugnana
- Theoretical introduction to probability/statistics
- For practical examples with R language see **Statistics for Data Science**

Why Statistics for Machine Learning?

We need grounded means for reasoning about data generated from real world with some degree of randomness.



What will you learn?

- Probability: properties of data generated by a known/assumed randomness model
- Statistics: properties of a randomness model that could have generated given data

Sample spaces and events

- An **experiment** is a measurement of a random process
- The **outcome** of an experiment takes values in some set Ω, called the **sample space**.

  Examples:
  - Tossing a coin: $\Omega = \{\text{H, T}\}$ *[Finite sample space]*
  - Month of birthdays $\Omega = \{\text{Jan}, \ldots, \text{Dec}\}$ *[Finite sample space]*
  - Population of a city $\Omega = \mathbb{N} = \{0, 1, 2, \ldots, \}$ *[Countably infinite sample space]*
  - Length of a street $\Omega = \mathbb{R}^+ = (0, \infty)$ *[Uncountably infinite sample space]*
  - Tossing a coin twice: $\Omega = \{\text{H, T}\} \times \{\text{H, T}\} = \{(\text{H, H}),(\text{H, T}),(\text{T, H}),(\text{T, T})\}$
  - Testing for Covid-19 (univariate): $\Omega = \{+, -\}$
  - Testing for Covid-19 (multivariate): $\Omega = \{\text{f, m}\} \times \mathbb{N} \times \{+, -\}$, e..g, $(\text{f, } 25, -) \in \Omega$
- An **event** is some subset of $A \subseteq \Omega$ of possible outcomes of an experiment.
  - $L = \{$ Jan, March, May, July, August, October, December $\}$ *a long month with 31 days*
- We say that an event A **occurs** if the outcome of the experiment belongs to the set A.
  - If the outcome is Jan then L occurs

Look at **seeing-theory.brown.edu**

Probability functions on finite sample space

A **probability function** is a mapping from events to **real numbers** that satisfies certain axioms. *Intuition: how likely is an event to occur.*

> DEFINITION. A *probability function* P on a finite sample space Ω assigns to each event A in Ω a number $\mathrm{P}(A)$ in [0,1] such that
> (i) $\mathrm{P}(\Omega) = 1$, and
> (ii) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$ if A and B are disjoint.
> The number $\mathrm{P}(A)$ is called the probability that A occurs.

- Fact: $P(\{a_1, \ldots, a_n\}) = P(\{a_1\}) + \ldots + P(\{a_n\})$ *[Generalized additivity]*
  - Assigning probability to a singleton is enough
- Examples:
  - $P(\{\mathrm{H}\}) = P(\{\mathrm{T}\}) = 1/2$
  - $P(\{\mathrm{Jan}\}) = 31/365$, $P(\{\mathrm{Feb}\}) = 28/365, \ldots P(\{\mathrm{Dec}\}) = 31/365$
  - $P(L) = 7/12$ or $31 \cdot 7/365$?
- $P(\{a\})$ often abbreviated as $P(a)$, e.g., $P(\mathrm{Jan})$ instead of $P(\{\mathrm{Jan}\})$

Properties of probability functions

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$ *[Impossible event]*
- $A \subseteq B \Rightarrow P(A) \leq P(B)$ *[Monotonicity]*
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ *[Inclusion-exclusion principle]*
- Example: $P(A \cup B) = P(A) + P(B \setminus A)$
- probability that at least one coin toss over two lands head?
  - Tossing a coin twice: $\Omega = \{\text{H, T}\} \times \{\text{H, T}\} = \{(\text{H, H}),(\text{H, T}),(\text{T, H}),(\text{T, T})\}$
  - $A = \{(H, H), (H, T)\}$ first coin is head
  - $B = \{(H, H), (T, H)\}$ second coin is head
  - Answer $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/2 - 1/4 = 3/4$

Defining probability functions

Assigning probability is **NOT** an easy task: a prob. function can be an approximation of reality

- **Frequentist** interpretation: probability measures a "*proportion of outcomes*".
  - A fair coin lands on heads 50% of times
  - $P(A) = {}^{|A|}/_{|\Omega|}$ *[Counting]*
  - $P(\{\text{ at least one H in two coin tosses}\}) = |\{(H,H),(H,T),(T,H)\}|/4 = {}^{3}/_{4}$
- **Bayesian** (or epistemological) interpretation: probability measures a "*degree of belief*".
  - (We believe that) Iliad and Odissey were composed by the same person at 90%

Probability functions on countably infinite sample space

DEFINITION. A *probability function* on an infinite (or finite) sample space Ω assigns to each event A in Ω a number $\mathrm{P}(A)$ in $[0,1]$ such that

(i) $\mathrm{P}(\Omega) = 1$, and

(ii) $\mathrm{P}(A_1 \cup A_2 \cup A_3 \cup \cdots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \mathrm{P}(A_3) + \cdots$ if $A_1, A_2, A_3, \ldots$ are disjoint events.

- (ii) is called countable additivity. It is equivalent to σ-additivity: for $A_1 \subseteq A_2 \subseteq \ldots$

$$P(\lim_{n\to\infty} A_i) = \lim_{n\to\infty} P(A_i)$$

- Example
  - Experiment: we toss a coin repeatedly until H turns up.
  - Outcome: the number of tosses needed.
  - $\Omega = \{1, 2, \ldots\} = \mathbb{N}^+$
  - Suppose: $P(H) = p$. Then: $P(n) = (1-p)^{n-1}p$
  - Is it a probability function? $P(\Omega) = \ldots$

Conditional probability

- Long months and months with 'r'
  - $L = \{$ Jan, Mar, May, July, Aug, Oct, Dec $\}$ *a long month with 31 days*
  - $R = \{$ Jan, Feb, Mar, Apr, Sep, Oct, Nov, Dec $\}$ *a month with 'r'*
  - $P(L) = 7/12 \quad P(R) = 8/12$
- Anna is born in a long month. What is the probability she is born in a month with 'r'?

$$P(R|L) = \frac{P(L \cap R)}{P(L)} = \frac{P(\{\text{Jan, Mar, Oct, Dec}\})}{P(L)} = \frac{4/12}{7/12} = \frac{4}{7}$$

- **Intuition:** probability of an event in the restricted sample space $\Omega \cap L$
  - *a-priori* probability $P(R) = 8/12$
  - *a-posteriori* probability $P(R|L) = 4/7 < 8/12$

Conditional probability

DEFINITION. The *conditional probability of A given C* is given by:

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

provided $P(C) > 0$.

Properties:

- $P(A|C) \neq P(C|A)$, in general
- $P(\Omega|C) = 1$
- if $A \cap B = \emptyset$ then $P(A \cup B|C) = P(A|C) + P(B|C)$

$P(\cdot|C)$ is a probability function

THE MULTIPLICATION RULE. For any events A and C:

$$P(A \cap C) = P(A \mid C) \cdot P(C) \,.$$

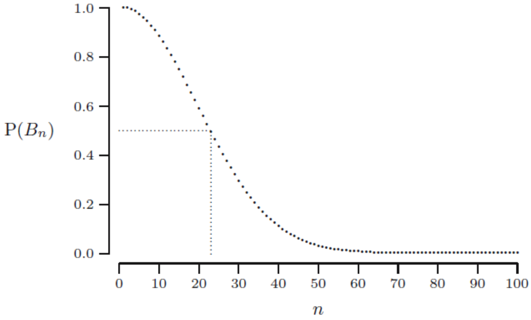
More generally, the **Chain Rule**:

$$P(A_1 \cap A_2 \cap A_3 \ldots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n| \cap_{i=1}^{n-1} A_i)$$

Example: no coincident birthdays

- $B_n = \{n \text{ different birthdays}\}$
- For $n = 1$, $P(B_1) = 1$
- For $n > 1$,

$$\begin{aligned} P(B_n) &= P(B_{n-1}) \cdot P(\{\text{the } n\text{-th person's birthday differs from the other } n-1\} | B_{n-1}) \\ &= P(B_{n-1}) \cdot (1 - \frac{n-1}{365}) = \ldots = \prod_{i=1}^{n-1} (1 - \frac{i}{365}) \end{aligned}$$



The law of total probability

THE LAW OF TOTAL PROBABILITY. Suppose $C_1, C_2, \ldots, C_m$ are disjoint events such that $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$. The probability of an arbitrary event A can be expressed as:

$$\mathrm{P}(A) = \mathrm{P}(A \mid C_1)\mathrm{P}(C_1) + \mathrm{P}(A \mid C_2)\mathrm{P}(C_2) + \cdots + \mathrm{P}(A \mid C_m)\mathrm{P}(C_m)\,.$$

- **Intuition:** case-based reasoning

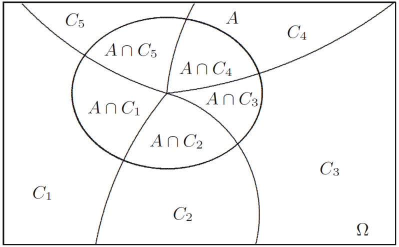


Fig. 3.2. The law of total probability (illustration for $m = 5$).

Example: case-based reasoning

Factory 1's light bulbs work for over 5000 hours in 99% of cases.
Factory 2's bulbs work for over 5000 hours in 95% of cases.
Factory 1 supplies 60% of the total bulbs on the market and Factory 2 supplies 40% of it.
Question: *What is the chance that a purchased bulb will work for longer than 5000 hours?*

- $A = \{\text{bulbs working for longer than 5000 hours}\}$
- $C_1 = \{\text{bulbs made by Factory 1}\}$, hence $C_2 = \{\text{bulbs made by Factory 2}\}$
- Since $\Omega = C_1 \cup C_2$ and $C_1 \cap C_2 = \emptyset$, by the multiplication rule:

$$P(A) = P(A|C_1) \cdot P(C_1) + P(A|C_2) \cdot P(C_2)$$

Answer: $P(A) = 0.99 \cdot 0.6 + 0.95 \cdot 0.4 = 0.974$

Independence of events

Intuition: whether one event provides any information about another.

Independence

An event A is independent of B, if $P(B) = 0$ or

$$P(A|B) = P(A)$$

- For $P(R|L) = 4/7 \neq 8/12 = PR(R)$ - knowing Anna was born in a long month change the probability she was born in a month with 'r' !
- Tossing 2 coins:
  - A_1 is "H on toss 1" and A_2 is "H on toss 2"
  - $P(A_1) = P(A_2) = 1/2$
  - $P(A_2|A_1) = P(A_2 \cap A_1)/P(A_1) = 1/4/1/2 = 1/2 = P(A_1)$
- Properties:
  - A independent of B iff $P(A \cap B) = P(A) \cdot P(B)$
  - A independent of B iff B independent of A *[Symmetry]*
  - A independent of B iff A^c independent of B

Physical independence and stochastic independence

Independence

An event A is independent of B, if $P(B) = 0$ or

$$P(A|B) = P(A)$$

- Physical independence implies stochastic independence
  - However, physical independence is quite a subtle matter (see the **butterfly effect**)
- But there are stochastic independent events that are physically dependent
  - Suppose a fair die is rolled twice.
  - $A =$ "a three is obtained on the second roll"
  - $B =$ "the sum of the two numbers obtained is less than or equal to 4"
  - **Exercise at home.** Prove that $P(A|B) = P(A)$

Conditional independence of events

Intuition: whether one event provides any information about another given a third event occurred. Technically, consider $P(\cdot|C)$ in independence.

Conditional independence

An event A is conditionally independent of B given C such that $P(C) > 0$, if $P(B|C) = 0$ or

$$P(A|B \cap C) = P(A|C)$$

- Properties:
  - A conditionally independent of B iff $P(A \cap B|C) = P(A|C) \cdot P(B|C)$
  - A conditionally independent of B iff B conditionally independent of A *[Symmetry]*
- **Exercise at home.** Prove or disprove:
  - If A is independent of B then A is conditionally independent of B given C

Independence of two or more events

INDEPENDENCE OF TWO OR MORE EVENTS. Events A_1, A_2, ..., A_m are called independent if

$$P(A_1 \cap A_2 \cap \cdots \cap A_m) = P(A_1)\,P(A_2)\cdots P(A_m)$$

and this statement *also* holds when any number of the events A_1, ..., A_m are replaced by their complements throughout the formula.

Alternative definition

Events $A_1, A_2, \ldots, A_m$ are called independent if for every $J \subseteq \{1, \ldots, m\}$:

$$P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i)$$

- **Exercise at home**: show the two definitions are equivalent

Independence of two or more events

Alternative definition

Events $A_1, A_2, \ldots, A_m$ are called independent if for every $J \subseteq \{1, \ldots, m\}$:

$$P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i)$$

- It is **stronger** than **pairwise independence**

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j) \text{ for } i \neq j \in \{1, \ldots, m\}$$

- Example: what is the probability of at least one head in the first 10 tosses of a coin?
  $A_i = \{\text{head in } i\text{-th toss}\}$

$$P(\bigcup_{i=1}^{10} A_i) = 1 - P(\bigcap_{i=1}^{10} A_i^c) = 1 - \prod_{i=1}^{10} P(A_i^c) = 1 - \prod_{i=1}^{10} (1 - P(A_i))$$

Bayes' Rule

BAYES' RULE. Suppose the events $C_1, C_2, \ldots, C_m$ are disjoint and $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$. The conditional probability of C_i, given an arbitrary event A, can be expressed as:

$$\mathrm{P}(C_i \mid A) = \frac{\mathrm{P}(A \mid C_i) \cdot \mathrm{P}(C_i)}{\mathrm{P}(A \mid C_1)\mathrm{P}(C_1) + \mathrm{P}(A \mid C_2)\mathrm{P}(C_2) + \cdots + \mathrm{P}(A \mid C_m)\mathrm{P}(C_m)}.$$

- It follows from $P(C_i|A) = \frac{P(A|C_i) \cdot P(C_i)}{P(A)}$ and the law of total probability
- Useful when:
  - $P(C_i|A)$ not easy to calculate
  - while $P(A|C_j)$ and $P(C_j)$ are known for $j = 1, \ldots, m$
  - E.g., in classification problems (see Bayesian classifiers from Data Mining)
- $P(C_i)$ is called the *prior* probability
- $P(C_i|A)$ is called the *posterior* probability (after seeing event A)

Testing for Covid-19

A new test for Covid-19 (or Mad-Cow desease, or drug use) has been developed.

- $\Omega = \{$ people aged 18 or higher $\}$
- $+ = \{$ people tested positive $\}$ $\quad - = \{$ people tested negative $\} = +^c$
- $C = \{$ people with Covid-19 $\}$ $\quad C^c = \{$ people without Covid-19 $\}$

In lab experiments, a sample of people with and without Covid-19 tested

- $P(+|C) = 0.99$ *[Sensitivity/Recall/True Positive Rate]*
- $P(-|C^c) = 0.99$ *[Specificity/True Negative Rate]*

What is the probability I really have Covid-19 given that I tested positive? *[Precision]*

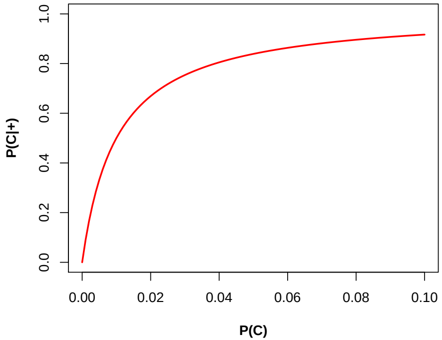
$$P(C|+) = \frac{P(C \cap +)}{P(+)} = \frac{P(+|C) \cdot P(C)}{P(+)} = \frac{P(+|C) \cdot P(C)}{P(+|C) \cdot P(C) + P(+|C^c) \cdot P(C^c)}$$

$$P(C|+) = \frac{0.99 \cdot P(C)}{0.99 \cdot P(C) + 0.01 \cdot (1 - P(C))}$$

$P(C)$ **is unknown!**

Testing for Covid-19

$P(C)$, the probability of having Covid-19, **is unknown**. Let's plot $P(C|+)$ over $P(C)$:



- For $P(C) = 0.02$, $P(C|+) = .67$
- For $P(C) = 0.06$, $P(C|+) = .86$
- For $P(C) = 0.10$, $P(C|+) = .92$

Optional references

Optional readings:

- [Sipka et al., 2022] survey methods for prior-shift adaptation (also when γ is unknown!).
- [Pozzolo et al., 2015] apply correction to the study of effectiveness of undersampling.

,

Tomáš Šipka, Milan Šulc, and Jiří Matas (2022)
The Hitchhiker's Guide to Prior-Shift Adaptation.
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 1516-1524.
https://arxiv.org/abs/2106.11695

Andrea Dal Pozzolo, Olivier Caelen, and Gianluca Bontempi (2015)
When is Undersampling Effective in Unbalanced Classification Tasks?
ECML/PKDD (1) 200–215.
Lecture Notes in Computer Science, volume 9284.
https://doi.org/10.1007/978-3-319-23528-8_13