

Introduction to Probability

Andrea Esuli



Probability

How likely is some event to happen?

- Toss a coin: probability of head?
- Toss a coin: probability of six consecutive tails?
- Lotto numbers: probability of the number 3
- Lotto numbers: probability of a sequence of five consecutive numbers?
- Weather: probability of sun?
- Weather: probability of rain in a sunny day?
- Weather: probability of rain in a cloudy day?
- Language: probability of a word to be a verb?
- Language: probability of "lemon" to be a verb?
- Language: probability of "lemon" to follow "the"?
- Language: probability of "the" to precede "lemon"?

Experiments and outcomes

An **experiment** is a repeatable process that produces an **outcome**.

The **sample space** is the set of all possible outcomes of a process Ω .

- Finite:
 - coin toss: {H,T},
 - dice: {1, 2, 3, 4, 5, 6},
 - lotto: sequences of five numbers from {1, 2, 3, 4, ..., 90}
- Infinite:
 - temperature: real number,
 - point on surface: x,y coordinates

Events

The **event space** is the set of all possible subsets of outcomes of the sample space, i.e., the **power set** $\mathcal{A}(\Omega)$ of outcomes.

Events are elements from the event space, i.e., a subset of outcomes, e.g.:

$$E_{\text{coin}} = \{\text{head}\}, E_{\text{dice}} = \{1,2,3\}, E_{\text{temp}} = 20, E_{\text{temp}} = 18 < T < 25, E_{xy} = \{x=1, y=2\}$$

- The event that include any possible outcome is the **sure** event, that is identified by whole set Ω .
- The event that does not include any outcome is the **impossible** event, i.e., an empty set \emptyset .

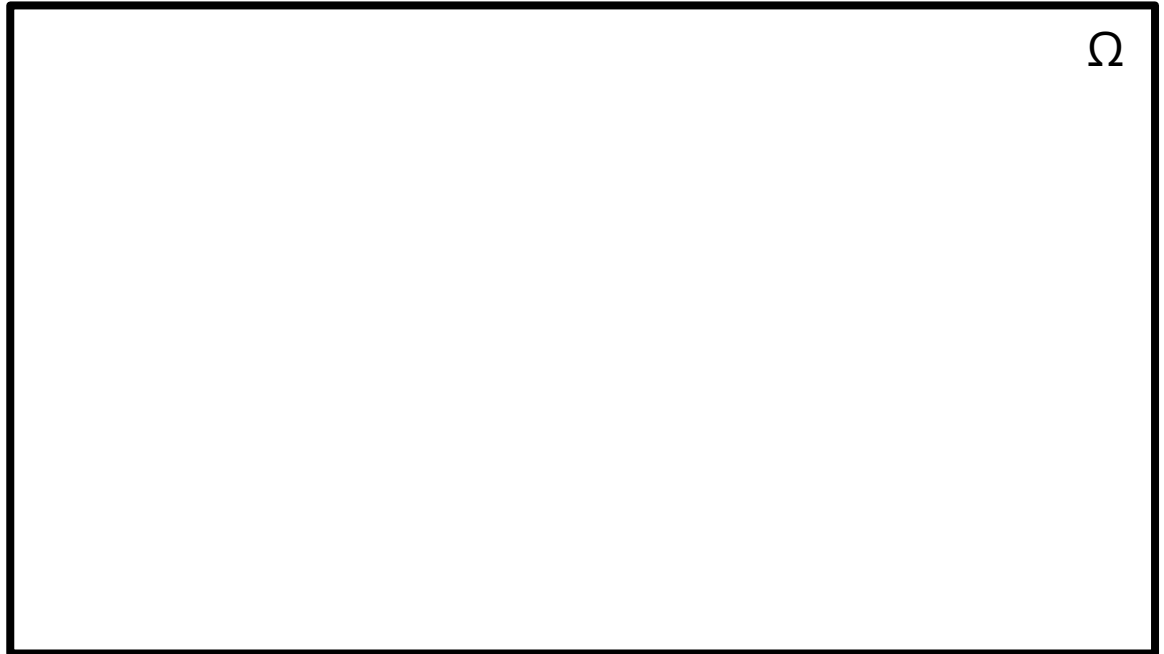
Events

- Set operations (union \cup , intersection \cap) among events define other events.
- Two events E and F are **mutually exclusive** if $E \cap F = \emptyset$
 - $\{1, 2, 6\} \cap \{3, 5\} = \emptyset$
- Two events E and F are **complementary** ($F = E^c = \neg E$) iff $E \cap F = \emptyset$ and $E \cup F = \Omega$
 - $\{1, 2, 6\} \cap \{3, 4, 5\} = \emptyset$ and $\{1, 2, 6\} \cup \{3, 4, 5\} = \Omega$

Venn diagrams

Visualization of events in the event space.

Ω is the event space.



Venn diagrams

Visualization of events in the event space.

A is an event, A^c is its complement



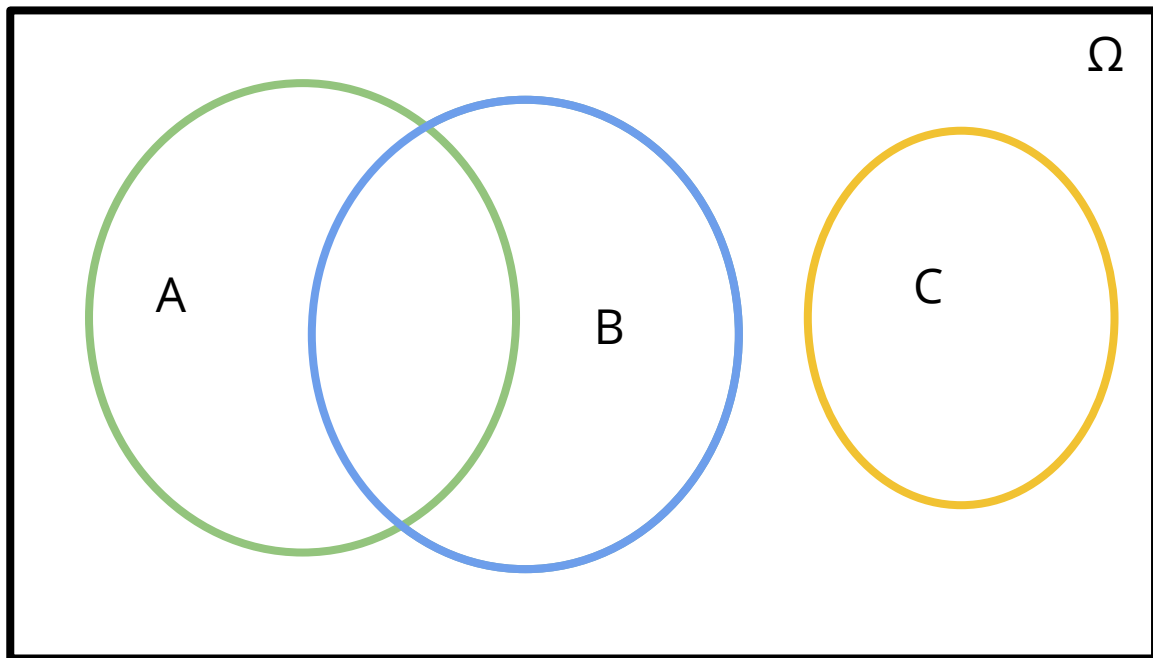
Venn diagrams

Visualization of events in the event space.

A and C are mutually exclusive.

B and C are mutually exclusive.

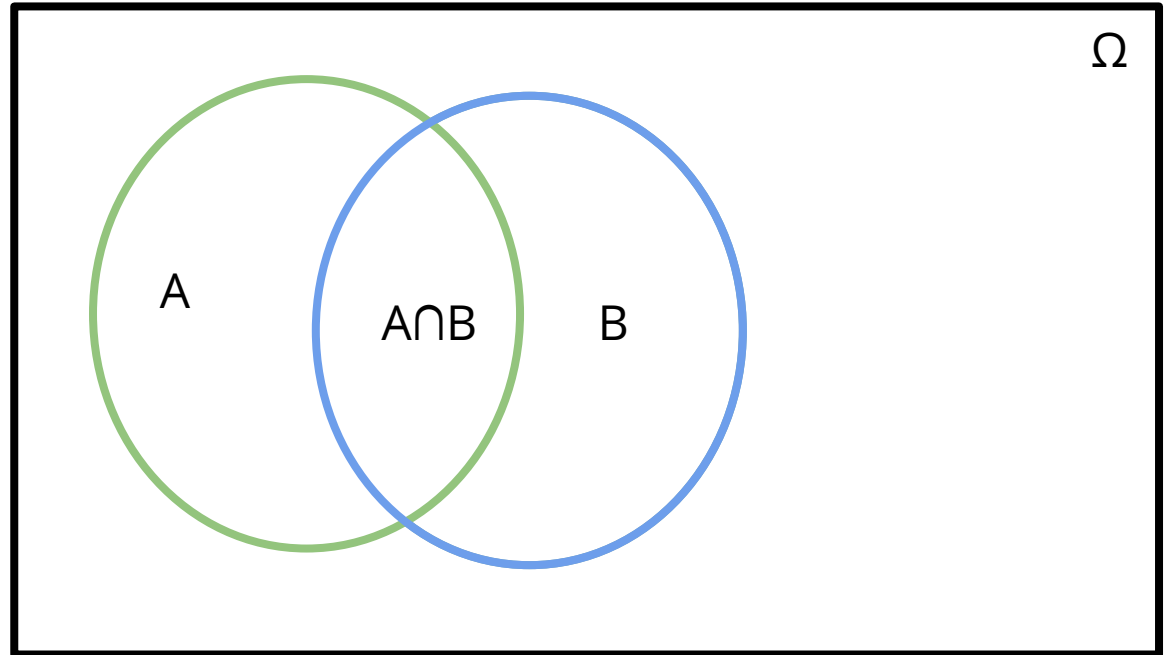
A and B are not mutually exclusive.



Venn diagrams

Visualization of events in the event space.

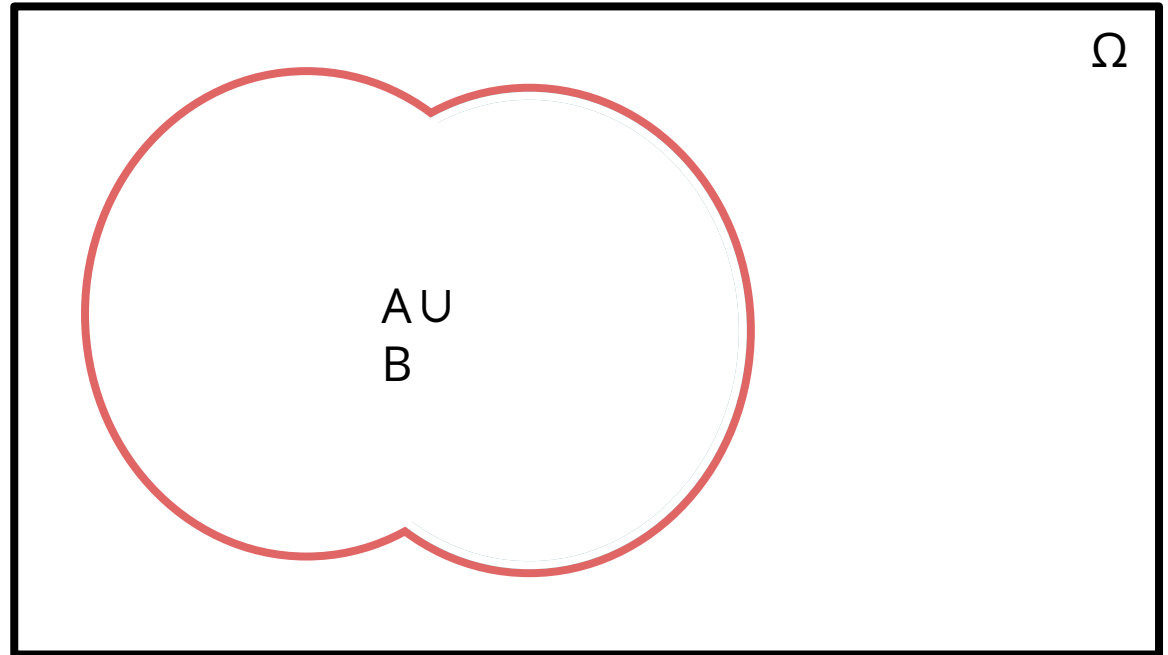
$A \cap B$ is the region where both A and B happen.



Venn diagrams

Visualization of events in the event space.

$A \cup B$ is the event where either A or B, or both, happen



Probability

A **probability law** P assigns to every event E a real number from 0 to 1.

- $0 \leq P(E) \leq 1$
- $P_{\text{coin}}(\{\text{tail}\})=0.5$
- $P_{\text{dice}}(\{1\}) = 1/6$
- 0 means that the event is *almost* impossible
- 1 means that the event is *almost* sure

Why almost? (hint: its related to infinite sample spaces)

A probability law must satisfy a set of axioms

Axioms of Probability

A **probability law** P assigns to every event E a number that models its likelihood to happen.

A probability law must satisfy three axioms:

- Non-negativity: $\forall E. P(E) \geq 0$
- Unitariness: $P(\Omega) = 1$
- σ -additivity:

For any set of mutually exclusive events $E_1, E_2 \dots E_n$ $P(\cup_i E_i) = \sum_i P(E_i)$

$\Omega, \mathcal{A}(\Omega)$ and P define the **probability space**.

Probability

- Monotonicity
 - $P(A) \leq P(B)$ for any $A \subseteq B$

Extreme cases:

- $P(E) = 0$ means that the event is *almost* impossible
- $P(E) = 1$ means that the event is *almost* sure

Why almost? (hint: its related to infinite sample spaces)

From counts to probability

When single outcomes of a process have **uniform probabilities**, i.e., they have all the same chance to happen, we can determine $P(E)$ of more complex event by **counting**.

$$P(E) = \text{number of positive outcomes} / \text{total number of outcomes}.$$

For many processes we can count the number of outcomes they can produce:

$$\#(\text{coin}) = 2 \quad \#(\text{dice}) = 6$$

If a process is repeated k times, each with n possible outcomes, the total number of outcomes is the n^k :

$$\#(\text{coin thrown } k \text{ time}) = 2^k \quad \#(\text{dice thrown } k \text{ time}) = 6^k$$

From counts to probability

More complex event spaces may require [combinatorial analysis](#).

Lotto is a sequence of five extractions **without replacement**:

$$\#(\text{cinquina at lotto}) = 90 \cdot 89 \cdot 88 \cdot 87 \cdot 86 / (5 \cdot 4 \cdot 3 \cdot 2 \cdot 1) = 43,949,268$$

Explanation:

- Numerator: the first number is extracted from 90 number, the second from 89, the third from 88...
- Denominator: order does not count. We can place the first number in five positions, the second in four, the third in three...

Side note: Italian lotto pays 6,000,000 times the bet (~7.3 times less the risk)

From counts to probability

To compute a probability we then have to count the number of positive events (or their complement):

$$\#(\text{two consecutive tosses with same outcome}) = \#(\{\text{HH,TT}\}) = 2$$

$$\#(\text{six consecutive tosses with same outcome}) = \#(\{\text{HHHHHH,TTTTTT}\}) = 2$$

$$\#(\text{two consecutive 6 with a dice}) = \#(\{66\}) = 1$$

$$\begin{aligned} \#(\text{two consecutive odd numbers with a dice}) &= \#(\{22,24,26,42,44,46,62,64,66\}) \\ &= 9 \end{aligned}$$

$$\#(\text{cinquina playing six numbers}) = 6$$

$$\#(\text{cinquina playing seven numbers}) = 6 \cdot 7 / 2 = 21 \text{ divided by two because the order does not count}$$

Uniform probabilities: from counts to probability

The probability is **the ratio** of **positive outcomes over the total number of outcomes**:

$$P(\text{two consecutive tosses with same outcome}) = 2/4 = 0.5$$

$$P(\text{six consecutive tosses with same outcome}) = 2/64 = 0.03125$$

$$P(\text{two consecutive 6 with a dice}) = 1/36 = 0.028$$

$$P(\text{two consecutive odd numbers with a dice}) = 9/36 = 0.25$$

$$P(\text{cinquina playing six numbers}) = 6/43,949,268 = 1.4 \cdot 10^{-7}$$

$$P(\text{cinquina playing seven numbers}) = 21/43,949,268 = 4.8 \cdot 10^{-7}$$

Properties of probabilities

- Probability from counts (for equiprobable outcomes):

$$P(E) = \#(E)/\#(\Omega)$$

- Union of independent events

$$P(A \text{ or } B) = P(A) + P(B)$$

- Intersection of independent events

$$P(A \text{ and } B) = P(A)*P(B)$$

- Probability of complement

$$P(A) = 1 - P(A^c) = 1 - P(\Omega \setminus A)$$

Non-uniform probabilities

Outcomes may be not equiprobable. In this case we cannot rely on counting, but we can exploit properties of probabilities to compute $P(E)$ for complex events.

A coin with $P(H) = 0.25$.

$$P(T) = 1 - P(H) = 0.75$$

$$P(\text{two heads or two tails}) = P(\text{two heads}) + P(\text{two tails}) = 0.25^2 + 0.75^2 = 0.625$$

$$P(\text{one head over two tosses}) = P(HH) + P(HT) + P(TH) = 0.25^2 + 0.75 * 0.25 * 2 = 0.4375$$

$$P(\text{one head over two tosses}) = 1 - P(\text{two tails}) = 1 - 0.75^2 = 0.4375$$

Probabilities and language

$P(E_1)$ = What's the probability of a word in a vocabulary to be a verb?

$P(E_2)$ = What's the probability to use a verb in a language?

$$P(E_1) =? P(E_2)$$

Probabilities and language

$P(E_1)$ = What's the probability of a word in a vocabulary to be a verb?

To determine E_1 we can take a **vocabulary** and count **how many of its words are verbs**.

$$P(E_1) = \#(\text{verbs in vocabulary}) / \#(\text{words in vocabulary})$$

Probabilities and language

$P(E_2)$ = What's the probability to use a verb in a language?

To determine E_2 we can take **a lot of text** and count **how many time a verb appears**.

$$P(E_2) = \#(\text{occurrences of verbs}) / \#(\text{occurrences of words})$$

Probabilities and language

$P(E_1)$ = What's the probability of a word in a vocabulary to be a verb?

$P(E_2)$ = What's the probability to use a verb in a language?

$$P(E_1) \neq P(E_2)$$

Language Models model **the use** of language from **large collections** of text.

Language modeling can be done in many ways:

What's the most probable word to appear after 'would'?

What's the probability of the letter 'h' to follow the letter 't'?

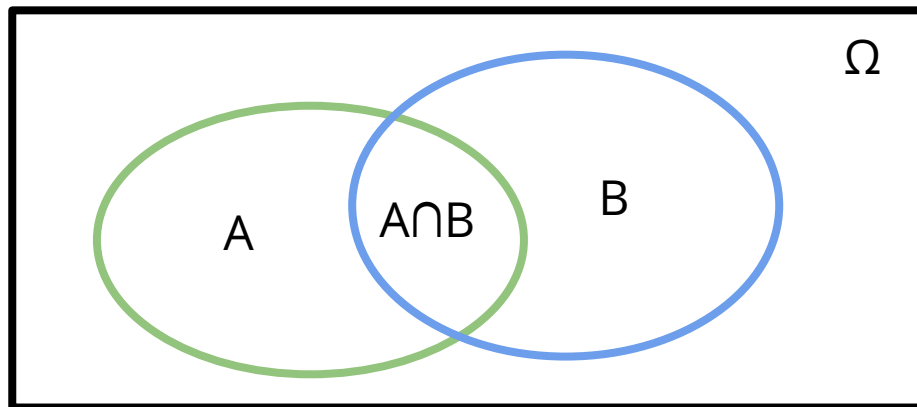
These are **conditional probabilities**.

Conditional probabilities

A **conditional probability** $P(A|B)$

determines the **probability** of the outcome to satisfy the **event A**

assuming that the outcome **satisfies** for sure the **event B**.



Conditional probabilities

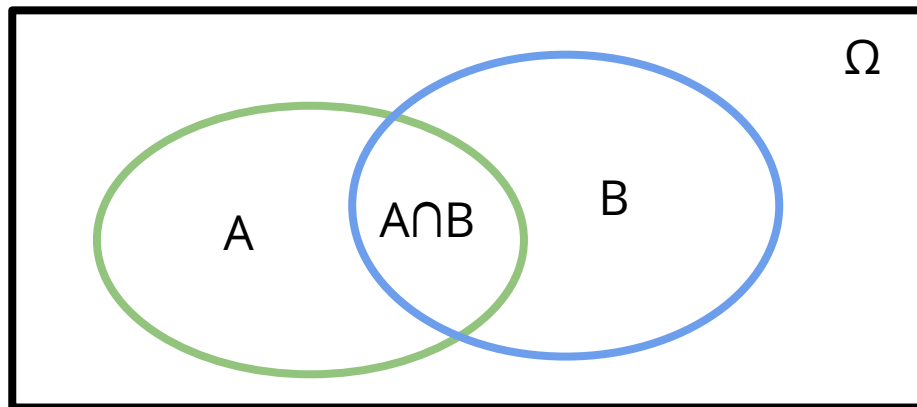
B is the set of outcome granted to happen.

Any outcome X outside B has $P(X) = 0$.

How much A does "cover" B?

The highest the coverage the highest the probability.

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



Conditional probabilities

$P('x') = ?$

$P(\text{red}) = ?$

$P('x' | \text{red}) = ?$

$P(\text{red} | 'x') = ?$

X	O	X	O	O	O	O	X	X
O	X	O	X	X	O	O	O	O
X	X	X	O	X	O	O	O	X
X	O	X	O	O	O	X	O	O

Conditional probabilities

$$P('x') = 15/36$$

$$P(\text{red}) = 17/36$$

$$P('x' | \text{red}) = 7/17$$

$$P(\text{red} | 'x') = 7/15$$

X	O	X	O	O	O	O	X	X
O	X	O	X	X	O	O	O	O
X	X	X	O	X	O	O	O	X
X	O	X	O	O	O	X	O	O

Conditional probabilities

Conditional probabilities allow to improve the prediction of the outcome of an experiment when partial information (evidence) is available.

Conditional probabilities are a key tool for **statistical inference**.

What's the probability of snow?

What's the probability of snow in Pisa?

What's the probability of snow in Pisa, when Temp = 30°C?

What's the most probable word to appear after 'would'?

What's the probability of the letter 'h' to follow the letter 't'?

What's the probability of a document containing words 'president', 'elections' and 'polls' to belong to topic 'cooking'?

From conditional probabilities...

$P(A,B)$ means the probability of events A **and** B to occur at the same time, i.e., $P(A,B) = P(A \cap B)$.

$$P(A | B) = P(A \cap B) / P(B) = P(A,B) / P(B)$$

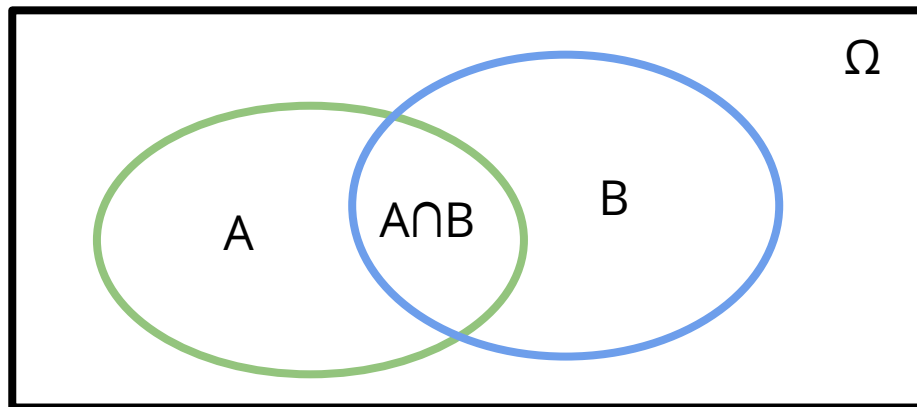
hence $P(A,B) = P(A | B)P(B)$

Similarly

$$P(B | A) = P(A \cap B) / P(A) = P(A,B) / P(A)$$

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

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



...to Bayes theorem

From $P(B | A)P(A) = P(A | B)P(B)$ we can derive:

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

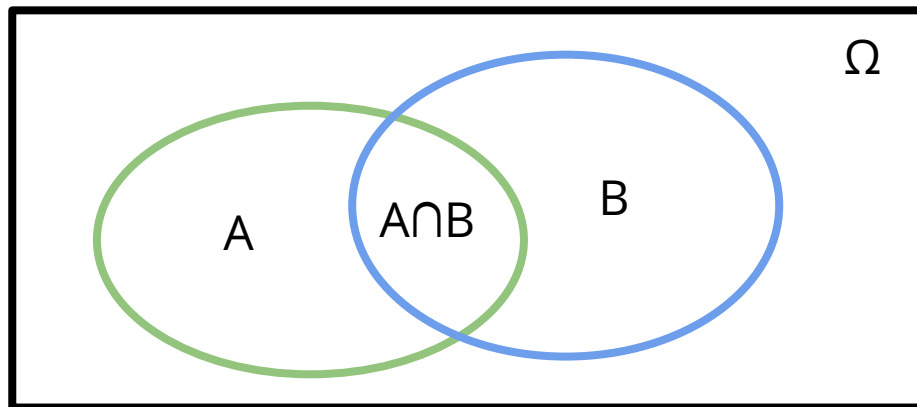
where

$P(B)$ = prior beliefs

$P(A | B)$ = likelihood

$P(A)$ = evidence

$P(B | A)$ = posterior beliefs



Bayesian classifier

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

posterior beliefs = prior beliefs * likelihood / evidence

Bayes theorem can be used to define a probabilistic classifier:

$$P(\text{class} | \text{document}) \propto P(\text{document} | \text{class}) P(\text{class})$$

The $P(\text{evidence})$ term can be discarded because it is constant when testing for different classes.

$P(\text{document} | \text{class})$ and $P(\text{class})$ can be **estimated** on a **training set** of **labeled documents**.

Bayesian classifier

We want to label document from a stream of news as either relevant for *cooking* or *politics*.

From a **training set** of **labeled news** from newspapers we can estimate:

$$P(\text{cooking}) = \#(\text{news about cooking})/\#(\text{all news}) = 0.01$$

$$P(\text{politics}) = \#(\text{news about politics})/\#(\text{all news}) = 0.12$$

These two probabilities the **prior** beliefs we have about the two labels.

If we have to label an unknown document, we have twelve times more chances of a correct classification if we label it with the label *politics*.

What if we are given the **evidence** that it contains the word '*zucchini*'?

Bayesian classifier

We can compute the **likelihood** of a document belonging to one of the label to contain the evidence, again using a **training set of labeled documents**.

$$P(\text{'zucchini'} \mid \text{cooking}) = \#(\text{cooking news with word 'zucchini'}) / \#(\text{cooking news}) = 0.05$$

$$P(\text{'zucchini'} \mid \text{politics}) = \#(\text{politics news with word 'zucchini'}) / \#(\text{politics news}) = 0.001$$

Multiplying priors with likelihoods we obtain the **posterior** beliefs, i.e., **the correction (update) of prior belief after observing the evidence**:

$$P(\text{cooking} \mid \text{'zucchini'}) \propto 0.05 * 0.01 = 0.0005$$

$$P(\text{politics} \mid \text{'zucchini'}) \propto 0.12 * 0.001 = 0.00012$$

For this document we have higher chances of a correct labeling for the label *cooking*, thus we assign this label.

Independent events

A and B are **independent events**

if and only if

the occurrence of one event **does not change** the probability of occurrence of the other event, i.e.:

$$P(A | B) = P(A) \text{ and } P(B | A) = P(B)$$

Which of these pairs of event are independent?

P_1 : (dice rolled even, dice rolled 1,2, or 3)

P_2 : (dice rolled even, dice rolled 1 or 2)

Independent events does not mean disjoint events.

$$P(A,B) = P(A)P(B)$$

Independent events

Assuming independence between events can simplify the modeling of probabilities of complex objects, e.g., text:

$$P(\text{text}) = P(\text{word}_1, \text{word}_2, \text{word}_3, \text{word}_4 \dots \text{word}_n) \stackrel{\text{independence}}{=} \prod_i P(\text{word}_i)$$

The [naïve bayesian classifier](#) uses the assumption of word independence to easily model language probabilities.