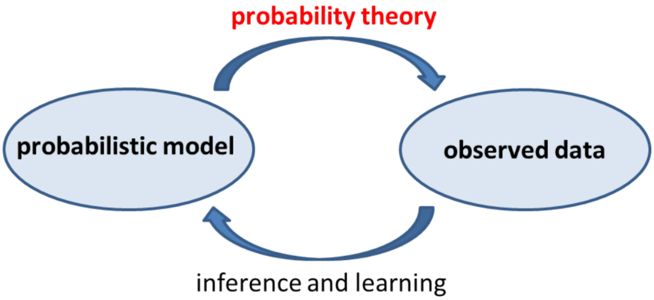
Probabilistic Machine Learning:
1. Probabilistic refresher

Tomasz Kajdanowicz, Piotr Bielak, Maciej Falkiewicz, Kacper Kania, Piotr Zieliński

Department of Computational Intelligence
Wroclaw University of Science and Technology

The presentation has been inspired and in some parts totally based on Prof. Mario A. T. Figueiredo presentation at LxMLS’2017, Instituto Superior Tecnico & Instituto de Telecomunicacoes, Lisboa, Portugal as well as working notes of Prof. Martin Ridout,University of Kent, UK.
Appropriate agreements to propagate their ideas have been acquired.

Probability theory

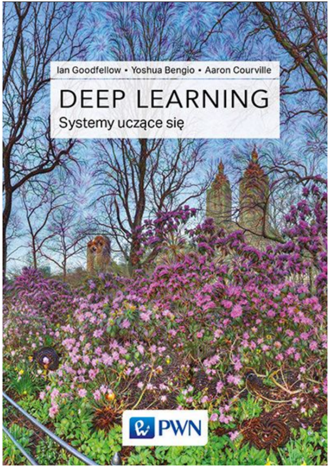


- has its origins in gambling
- great names: Fermat, Pascal, Bernoulli, Huygens, Laplace, Kolmogorov, Poisson, Cauchy, Boltzman, Bayes, Cardano, ...
- tool to handle uncertainty, information, knowledge, observations, ...
- ...thus also learning, decision making, inference, science, data science ...

Do we still need to know probability theory?

What book is this from?

Do we still need to know probability theory?



What is probability?

Probability theory is the branch of mathematics that treats those aspects of systems that have random or haphazard features.

Example: Clinical trial of a new drug

Old drug: 82% success rate.
New drug: 90% success rate claimed.
Action: Test the new drug on 100 patients, investigate the success rate.
Notes:

1. We will need to have a ‘reasonable’ number of patients. But how many must we use before we can be confident that we can rely on the results?
2. Different groups of patients will give different results, so there is need for some theory to describe and explain what happens.
3. How should one choose the patients who will receive the new drug?

Example: Opinion Polling

Political opinion polls are frequently used to assess the current political view of the electorate. Simple illustration: By-election, 2 candidates. [no ‘don’t knows’]

Action: Choose a sample of 100 individuals from the electorate, and ask them how they would vote ’if the by-election were tomorrow’.

Problems:

- possibility of unrepresentative (‘biased’) sample
- inherent error (uncertainty) of sampling

Notes:

1. We will need the concept of sampling at random . This in turn will require us to understand the concept of randomness.

Reminder: Probability theory is the branch of mathematics that treats those aspects of systems that have random or haphazard features.

Books

1. Schay, G. (2016). Introduction to Probability with Statistical Applications. In Introduction to Probability with Statistical Applications. https://doi.org/10.1007/978-3-319-30620-9
2. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. In Pattern Recognition (Vol. 4). https://doi.org/10.1117/1.2819119
3. Hastie, T., Tibshirani, R., Friedman, J. (2008). The Elements of Statistical Learning. In Springer series in Statistics. https://doi.org/10.1198/jasa.2004.s339
4. Barber, D. (2011). Bayesian Reasoning and Machine Learning. https://doi.org/10.1017/CBO9780511804779
5. Murphy, K. P. (2012). Machine Learning, A Probabilistic Perspective. In R&D Management (Vol. 16). https://doi.org/10.1111/j.1467-9310.1986.tb01158.x

Experiment and Event

In each previous examples **something which is done**
e.g. *drug given to patient, sample drawn from the electorate* ***has an uncertain result***
e.g. number cured, identity of sample members.

These are called

EXPERIMENT OR TRIAL** and **EVENT, OUTCOME or RESULT

Experiments

- *Experiments are considered to be repeatable, but at any repetition we do not know what the result will be.*
- *But we do know the set of all possible outcomes.*

The Sample Space

Definition

The Sample Space for an experiment is the set of all possible outcomes of that experiment.

The Sample Space: Expamples

1. Experiment: toss a coin 5 times and count the number of heads.
   Sample Space: 0, 1, 2, 3, 4, 5 .
2. Experiment: As above, but record the event in order.
   Sample Space: all 2^5 5-tuples of H and T, that is:
   HHHHH
   HHHHT
   HHHTH
   ...
   TTTTT

Remark: Both these sample spaces are finite. There are other possibilities.

The Sample Space: Expamples

4. Experiment: Toss a coin repeatedly, and record the number of *T*s before the first *H*, e.g.:
   TTH → 2
   TTTTTH → 5
   Sample space: 0, 1, 2, . . .
5. Record the duration (in seconds) of the next telephone call through the University switchboard.
   Sample Space: Set of all positive real numbers.

The Sample Spaces

Notes on Sample Spaces

- In example 4 the Sample Space is continuous. All the others are discrete.
- Usually an experiment involving measuring has a continuous Sample Space and one involving counting a discrete Sample Space.
- For most practical purposes we ignore the fact that strictly all Sample Spaces are discrete because of the finite precision with which numbers are recorded.
- The Sample Space of an experiment is not necessarily unique: there may well be more than one way to describe the outcomes. (c.f. earlier examples on coin tossing.)
- The elements of a Sample Space form a set of outcomes.

Standard notation: **Experiment** $\mathcal{E}$ and **Sample Space** $\mathcal{S}$.

Definition

An event is any subset of a sample space.

Examples of events

Sample Space	Event
number of heads	is even; exceeds 3; is 2 or 5
the 5-tuple	contains no pair of consecutive H's; is HT T T H; has 2nd element equal to T
length of phone call	between 100 and 200; <60; > 500

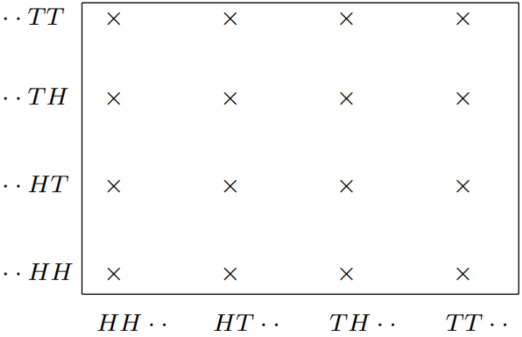
Venn diagrams

These diagrams are just ways of illustrating Sample Spaces, outcomes and events, and their relationships.

Notation

- The **Sample Space** $\mathcal{S}$ is represented by a rectangle.
- The **outcomes** in $\mathcal{S}$ are represented by points in the rectangle.
- **Events** are collections of outcomes, so are represented by collections of points, that is, by regions in the rectangle.
- Events are **subsets** of $\mathcal{S}$. They are identified by drawing a closed curve around the relevant points.

Example



$\mathcal{E}$: Toss 4 coins, record sequence.

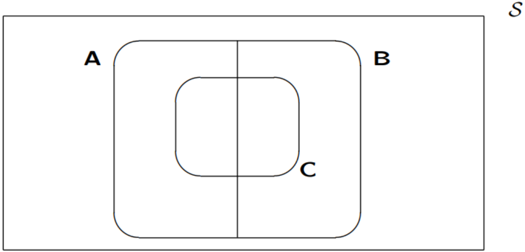
Events

- A: 3 or 4 H
- B: exactly 2 H
- C: 2nd and 3rd result H

Example cont.

The 4 coins example is a case, where:

- A and B do not overlap
- If C occurs then one of A and B must occur.



Events of special importance

1. $\mathcal{S}$ itself – the certain event.
2. 0 or $\emptyset$, a dummy event containing no outcomes, the impossible event.
3. x, where *x* is a single element of $\mathcal{S}$, i.e. an outcome.

Definition

An event comprising exactly one outcome is often termed a **simple event**; any other event is a **compound event**.

Reminder: Events can be viewed as sets - it's only the context and area of application which differ.

Relationships between events

Definition

Events $A_1, A_2, \ldots, A_n$ are **mutually exclusive** if no two can occur simultaneously, i.e. if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Definition

Events $A_1, A_2, \ldots, A_n$ are **mutually exhaustive** if at least one is certain to occur, i.e. if $A_1 \cup A_2 \cup, \ldots \cup A_n = \mathcal{S}$

Several important results are concerned with events that are mutually exclusive and exhaustive.

Probability

- To each event arising out of an experiment, a number (the probability of that event) is permanently assigned.
- For the various different events arising out of the same experiment, the probabilities need to be consistent.
- In principle, we can assign any self-consistent numbers to these events. But, for the probabilities to be useful, we insist that the 'more likely' the event is, the larger its probability should be.

Standard notation

The probability of an event A is denoted by $P(A)$ or $Pr(A)$.

Example

$\mathcal{E}$: Two dice are thrown
A: total score is 7
$Pr(A) = \frac{1}{6}$

Good practice

It is good practice to define and use clear notation, as in this example. For each event you consider, define notation such as $A, B, \ldots$ or $A_1, A_2, \ldots$ as appropriate. Then write the probabilities as $Pr(A), Pr(B), \ldots$ or $Pr(A_1), Pr(A_2), \ldots$

Confusion

It is possible to write statements such as $Pr(\text{total of 7 when two dice are thrown}) = 1/6$, but this confuses the definitions of the experiment and event.

Probability and relative frequency

- suppose that an experiment is performed n times
- suppose that event A occurs on n_A of these
- the **frequency** of A is n_A
- the **relative frequency** (r.f.) of A is $\frac{n_A}{n}$

The r.f. give us some idea of how likely A is to occur. But note that:

- the probability of A, $Pr(A)$, is a **fixed** quantity, but $\frac{n_A}{n}$ can and will change
- if another set of n trials were performed, $\frac{n_A}{n}$ would be very likely to be different
- if just one further trial is performed, the r.f. $\frac{n_A}{n}$ **will change** to $\frac{n_A+1}{n+1}$ or $\frac{n_A}{n+1}$

So the relative frequency $\frac{n_A}{n}$ cannot be used as the definition of $Pr(A)$.

So, what is probability?

Example

$\mathbb{P}$(randomly drawn card is $\heartsuit$) = $13/52$
$\mathbb{P}$(getting 1 in throwing a fair die) = $1/6$

- Classical definition: $\mathbb{P}(A) = \frac{N_A}{N}$
  ...with N mutually exclusive equally likely outcomes, N_A of which result in the occurrence of A.
- Frequentist definition: $\mathbb{P}(A) = \lim_{N\to\infty} \frac{N_A}{N}$
  ...relative frequency of occurrence of A in infinite number of trials
- Subjective probability:
  ...gives meaning to $\mathbb{P}$(“it will rain today”), or $\mathbb{P}$(“I’ll have passed the PUMa’s exam next winter”)

Kolmogorov's Axioms for Probability

- Probability is a function that maps events A into the interval $[0, 1]$.
  Kolmogorov's axioms (1933) for probability
  - For any A, $Pr(A) \geq 0$
  - $Pr(\mathcal{S}) = 1$
  - If $A_1, A_2, \cdots \subseteq \mathcal{S}$ are disjoint events, then $Pr(\bigcup_i A_i) = \sum Pr(A_i)$
- From these axioms, many results can be derived

Example

- $Pr(\emptyset) = 0$
- $C \subset D \implies Pr(C) \leq Pr(D)$
- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$
- $Pr(A \cup B) \leq Pr(A) + Pr(B)$

Numerical assessment of probabilities

There are three main ways of doing this:

- symmetry
- limiting relative frequency
- subjective (degrees of belief)

Symmetry

When to apply?

This can be applied only when all the outcomes of the experiment are known to be equally likely. The experiment is then symmetrical with respect to its outcomes.

Calculation:

- If there are n outcomes and if all are equally likely, each has probability $\frac{1}{n}$
- If an event A comprises a outcomes then $Pr(A) = \frac{a}{n}$

$$Pr(event) = \frac{\text{No. of outcomes favourable to event}}{\text{Total no. of outcomes}}$$

Limiting realative frequency

We have seen that relative frequencies cannot be used directly to define probabilities. We need to define $Pr(A)$, where A is some event based on the experiment $\mathcal{E}$.

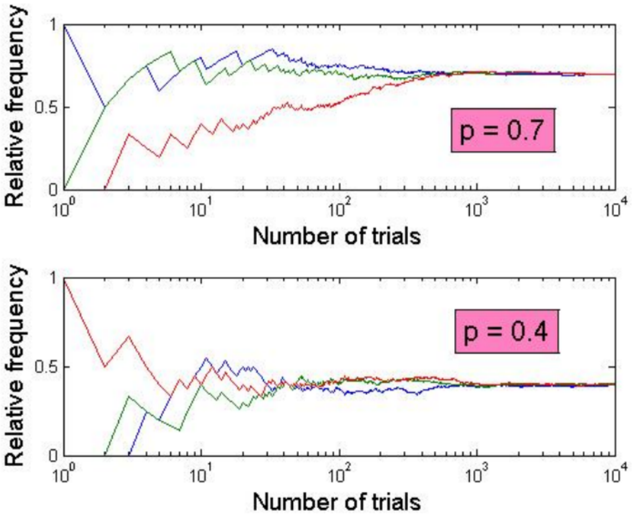
Relative frequency

Suppose that the experiment $\mathcal{E}$ is repeated over and over again. Let n_A denote the number of occurrences of A in the first n repetitions of the experiment.
The relative frequency of A, S_n, is given by

$$S_n = \frac{n_A}{n}$$

and we can plot a graph of S_n against n.

Limiting realative frequency



Limiting realative frequency

- graphs of this sort ‘settle down’ to some constant value
- The intuitive idea is that as n gets large S_n gets close to some constant, which we then define as the probability $Pr(A)$. (The form of convergence is not straightforward).

This definition is known as the **Limiting Relative Frequency** definition of probability.

Important

For any finite n, S_n is not the probability of A. It may be thought of as an estimate of $Pr(A)$.

Subjective probability

There are drawbacks with earlier definitions:

- **Symmetry:** restricted to experiments where we know the outcomes have equal probability.
- **Lim Rel Freq:** restricted to experiments which can be repeated over and over again.

If we define probability in some other way, we can extend the concept to experiments which do not have these properties.

Subjective probability

Subjective probability does this. It represents a person's degree of belief in some event.

Examples

- Trump will win the next general election
- the tram no 10 will go off the rails today

Summary so far

1. The subject of probability relates to experiments whose result is uncertain but must be one of a set of outcomes, the sample space.
2. Event: events are sets and can be manipulated ($\cup$, $\cap$, etc.)
3. To each event A is attached a number, its probability, denoted by $P(E)$ or $Pr(E)$.
4. Probabilities satisfy certain conditions – the axioms – on which the development of probability theory is based.
5. Several interpretations of probability obey the system of axioms.

Sample space with no structure

In many practical examples, the sample space has some structure: there are relationships between the outcomes. However, probability theory is developed without making such assumptions.

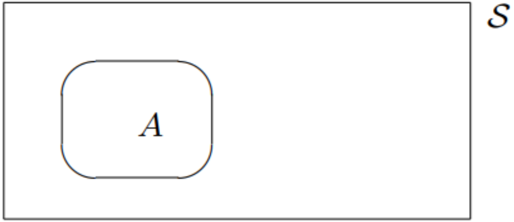
Deductions from Axioms

REMINDER

- A1. For any A, $Pr(A) \geq 0$
- A2. $Pr(\mathcal{S}) = 1$
- A3. If $A_1, A_2, \cdots \subseteq \mathcal{S}$ are disjoint events, then $Pr(\bigcup_i A_i) = \sum Pr(A_i)$

Theorem

$$Pr(\overline{A}) = 1 - Pr(A)$$



Deductions from Axioms

Proof.

Since $\overline{A}$ is the complement of A, $A \cup \overline{A} = \mathcal{S}$ and $A \cap \overline{A} = \emptyset$
Hence, using A3

$$Pr(A \cup \overline{A}) = Pr(A) + Pr(\overline{A})$$

i.e.

$$Pr(\overline{A}) = Pr(A \cup \overline{A}) - Pr(A)$$

but $A \cup \overline{A} = \mathcal{S}$ and by A2 $Pr(\mathcal{S}) = 1$:

$$Pr(\overline{A}) = 1 - Pr(A)$$

□

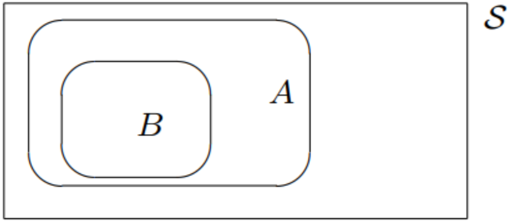
Homework 1:

- Proof that $Pr(\overline{\mathcal{S}}) = 0$
- Deduce if $Pr(\emptyset) = 0$

Deductions from Axioms

Theorem

If $A \supset B$ then $Pr(A) \geq Pr(B)$



Deductions from Axioms

Proof.

If $A \supset B$
then $A = B \cup (A \cap \overline{B})$ and $B \cap (A \cap \overline{B}) = \emptyset$
Hence, by A3, $Pr(A) = Pr(B) + Pr(A \cap \overline{B})$.
By A1, $Pr(A \cap \overline{B} \geq 0)$, so

$$Pr(A) \geq Pr(B)$$

$\square$

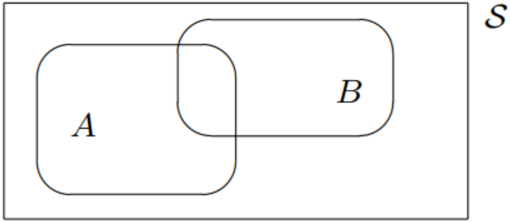
Note: Intuitively, this result is also obviously true.

Deductions from Axioms

Theorem

For any two events A and B

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$



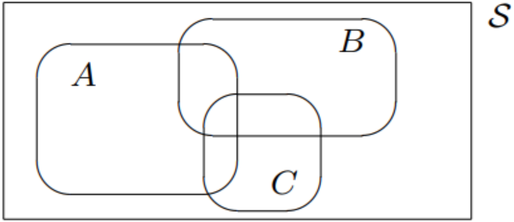
Deductions from Axioms

Proof.

$A \cup B = A \cup (\overline{A} \cap B)$
and
$B = (A \cap B) \cup (\overline{A} \cap B)$
RHS contains mutually exclusive events.
Use A3 and substitute for $Pr(\overline{A} \cap B)$ $\square$

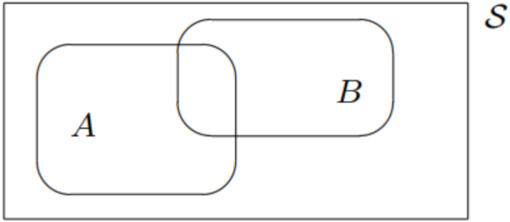
Homework 2

- Provide extension of the previous to 3 events: $Pr(A \cup B \cup C)$



Conditional Probability

This topic relates to two (or more) events associated with the same experiment.



Two events A and B divide S into 4 regions.
We now consider the form of relationship between these events.

Conditional Probability

Example

$\mathcal{E}$ – two cards are taken in sequence, without replacement, from a pack, at random.
We consider two events:

- A: the first card is an ace, $Pr(A) = 4/52$
- B: the second card is an ace, $Pr(B) = 4/52$.

Suppose now that the first card is examined and seen to be an ace. What is $Pr(B)$?
The answer **is not** $\frac{3}{51}$

Reminder: To each event arising out of an experiment, a number (the probability of that event) is permanently assigned.So what does the number $\frac{3}{51}$ represent?

The ratio $\frac{3}{51}$ does not arise from an experiment alone. It appears as a result of an experiment being performed and a particular condition being met.

- The **experiment** is that we choose two cards, at random, in sequence.
- The **condition** is that the first card chosen is an Ace.

We can say that the **conditional probability** of B given A is $\frac{3}{51}$.

Definition

If A and B are two events, then the **conditional probability** of B given A is defined as

$$\frac{Pr(A \cap B)}{Pr(A)}$$

for an event A such that $Pr(A) > 0$.

Notation

We write the conditional probability of B given A as $Pr(B|A)$. That is,

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

Example

Cards are selected at random without replacement from a pack. Define D as the event:
D = first ace appears at 2nd card.
Find Pr(D).

Solution

Define two events:

- A = first card chosen is not an ace
- B = second card chosen is an ace

Then $D \equiv A \cap B$.
Now $Pr(A) = \frac{48}{52}$, and $Pr(B|A) = \frac{4}{51}$.
Hence

$$Pr(D) = Pr(A \cap B) = Pr(A)Pr(B|A) = \frac{48}{52} \times \frac{4}{51} = \frac{16}{221} = 0.0724$$

Homework 3

- Provide extension of conditional probability to 3 events.

Independence

Definition

If for two events A and B

$$Pr(B|A) = Pr(B)$$

then we say that B **is independent of** A ($B \perp\!\!\!\perp A$)

Alternatively, we say that the events A and B are independent of each other.

Notes on independence

Independence is **reflexive**
If B is independent of A, then

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

Therefore

$$Pr(A \cap B) = Pr(A)Pr(B)$$

Dividing both sides by $Pr(B)$, we obtain:

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

So, if A is independent of B, then B **is independent of** A, and vice versa.

Notes on independence

Interpretation of independence
If A and B are not independent, then

$$Pr(B|A) \neq Pr(B)$$

Information that A has occurred changes our assessment of B.

But, if A and B are independent, knowledge about the occurrence of B does not affect our assessment of A.

Notes on independence

Theorem

If A and B are independent, then so are A and $\overline{B}$.

The Multiplication Law of probability

When events A and B are independent, then

$$Pr(A \cap B) = Pr(A)Pr(B)$$

In words: The multiplication law states that, if A and B are independent events, then their joint probability is the product of the individual probabilities.

Law of Total Probability

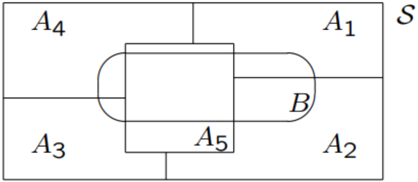
Consider a set $A_1, A_2, \ldots A_k$ of mutually exclusive and exhaustive events. Let B be some other event from the same experiment.

Law of Total Probability

$$Pr(B) = Pr(A_1)Pr(B|A_1) + Pr(A_2)Pr(B|A_2) + \ldots + Pr(A_k)Pr(B|A_k)$$
$$= \sum_{i=1}^{k} Pr(A_i)Pr(B|A_i)$$

Proof

PROOF: (illustrated for the case k = 5)



Each element of $\mathcal{S}$ is a member of one and only one of the A's.
We therefore obtain the result:

$$B \equiv (A_1 \cap B) \cup (A_2 \cap B) \cup \ldots \cup (A_k \cap B)$$

where the events on the RHS are mutually exclusive.
Hence $Pr(B) = \sum_{i=1}^{k} Pr(A_i \cap B)$.
But $Pr(A_i \cap B) = Pr(A_i)Pr(B|A_i)$
and so

$$Pr(B) = \sum_{i=1}^{k} Pr(A_i)Pr(B|A_i)$$

Homework 4

Three boxes contain certain items: box i contains n_i items, of which d_i are defective. In an experiment, one box is chosen at random. Then, one item is chosen at random from the chosen box. Find the probability that the chosen item is defective, when

$$n_1 = 50, n_2 = 100, n_3 = 100,$$

$$d_1 = 5, d_2 = 3, d_3 = 5$$

Bayes' theorem

Theorem

If $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive events, and if B is an event based on the same experiment, then

$$Pr(A_i|B) = \frac{Pr(A_i)Pr(B|A_i)}{\{\sum_{j=1}^{k} Pr(A_j)Pr(B|A_j)\}}$$

Proof.

By definition, $Pr(A_i \cap B) = Pr(A_i)Pr(B|A_i) = Pr(B)Pr(A_i|B)$. Equating the right-hand sides, we obtain:

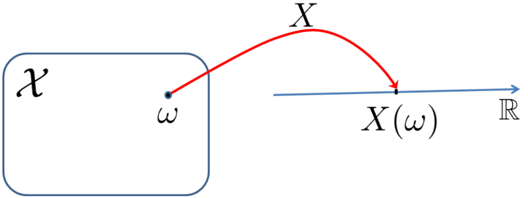
$$Pr(A_i|B) = \frac{Pr(A_i)Pr(B|Ai)}{Pr(B)}$$

$$= \frac{Pr(A_i)Pr(B|A_i)}{\{\sum_{j=1}^{k} Pr(A_j)Pr(B|A_j)\}}$$

using the law of total probability. □

Random Variables

- assume that $\mathcal{X}$ is a sample space
- A (real) random variable (RV) is a function: $X : \mathcal{X} \to \mathbb{R}$



  - Discrete RV: range of X is countable (e.g., $\mathbb{N}$ or $\{0, 1\}$)
  - Continuous RV: range of X is uncountable (e.g., $\mathbb{R}$ or $[0, 1]$)

Example

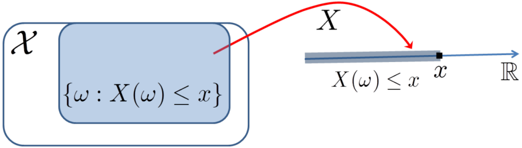
number of heads in tossing two coins, $\mathcal{X} = \{HH, HT, TH, TT\}$,
$X(HH) = 2$, $X(HT) = X(TH) = 1$, $X(TT) = 0$, range of $X = \{0, 1, 2\}$

Example

distance traveled by a tossed coin; range of $X = \mathbb{R}_+$

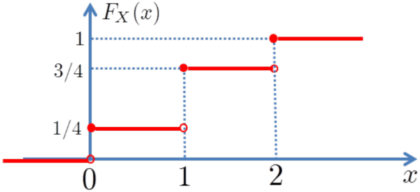
Random Variables: Distribution Function

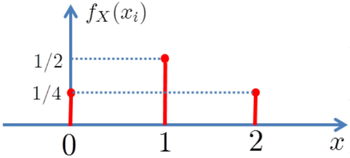
- **Distribution function**: $F_X(x) = Pr(\{\omega \in \mathcal{X} : X(\omega) \leq x\})$



Examples

number of heads in tossing 2 coins; range(X) = $\{0, 1, 2\}$





- **Probability mass function** (discrete RV): $f_X(x) = Pr(X = x)$, $F_X(x) = \sum_{x_i \leq x} f_X(x_i)$

Important Discrete Random Variables

- Uniform: $X \in \{x_1, \ldots, x_K\}$, pmf $f_X(x_i) = 1/K$

Examples

a fair roulette $X \in \{1, \ldots, 36\}$, with $f_X(x) = 1/36$
a fair die $X \in \{1, \ldots, 6\}$, with $f_X(x) = 1/6$

- Bernoulli RV: $X \in \{0, 1\}$, pmf $f_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$

  Compact form: $f_X(x) = p^x(1-p)^{1-x}$

Examples

an unfair coin (heads = 0, tails = 1), with $p \neq 1/2$.

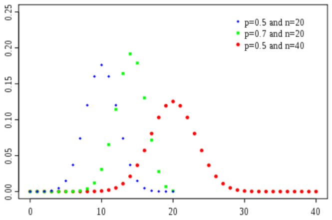
Important Discrete Random Variables

- Binomial RV: $X \in \{0, 1, \ldots, n\}$ (sum of n Bernoulli RVs)

  $$f_X(x) = \text{Binomial}(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

  Binomial coefficients ("n choose "x"):

  $$\binom{n}{x} = \frac{n!}{(n-x)!x!}$$



Example

number of heads in n coin tosses.

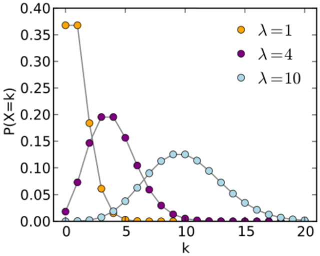
Other Important Discrete Random Variables

- Geometric(p): $X \in \mathbb{N}$, pmf $f_X(x) = p(1-p)^{x-1}$

Example

number of coin tosses until first heads

- Poisson(λ): $X \in \mathbb{N} \cup \{0\}$ pmf $f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$



Example

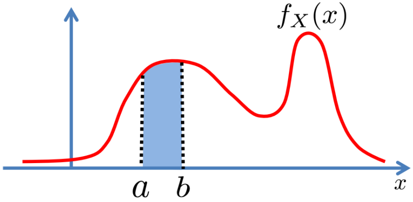
"...probability of the number of independent occurrences in a fixed (time/space) interval, if these occurrences have known average rate"

Continuous Random Variables

- Probability density function (pdf, continuous RV): $f_X(x)$

$$\int_{-\infty}^{\infty} f_X(x) = 1$$

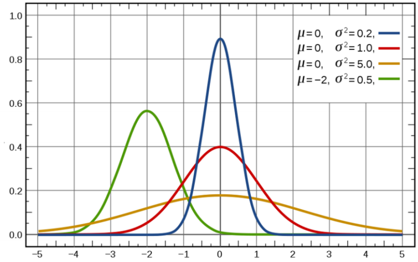
$$\mathbb{P}(X \in [a, b]) = \int_a^b f_X(x)dx$$

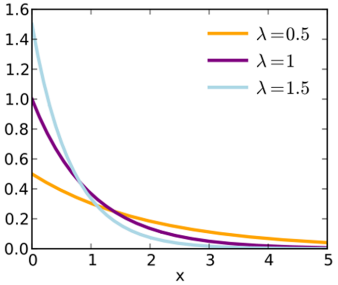


- Notice: $\mathbb{P}(X = c) = 0$

Important Continuous Random Variables

- Uniform: $f_X(x) = \text{Uniform}(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b] \end{cases}$
- Gaussian: $f_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$





- Exponential: $f_X(x) = \text{Exp}(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$