Machine Learning Applications: Introduction to Probability

Alessio Benavoli

CSIS
University of Limerick

Rules of probability

$p(x = \mathsf{x})$: the probability of variable x being in state x.

$$p(x = \mathsf{x}) = \begin{cases} 1 & \text{we are certain } x \text{ is in state x} \\ 0 & \text{we are certain } x \text{ is not in state x} \end{cases}$$

Values between 0 and 1 represent the degree of certainty of state occupancy.

---

domain (also called *possibility space*)

$\mathrm{dom}(x)$ denotes the states x can take. For example, $\mathrm{dom}(coin) = \{\mathsf{heads}, \mathsf{tails}\}$. When summing over a variable $\sum_x p(x)$, the interpretation is that all states of x are included, *i.e.* $\sum_x p(x) \equiv \sum_{\mathsf{s} \in \mathrm{dom}(x)} p(x = \mathsf{s})$.

---

distribution

Given a variable, x, its domain $\mathrm{dom}(x)$ and a full specification of the probability values for each of the variable states, $p(x)$, we have a **distribution** for x.

---

normalisation

The summation of the probability over all the states is 1:

$$\sum_{\mathsf{x} \in \mathrm{dom}(x)} p(x = \mathsf{x}) = 1$$

We will usually more conveniently write $\sum_x p(x) = 1$.

Example

Fair Dice

- x denotes the outcome of the dice
- $\mathrm{dom}(x) = \{1, 2, 3, 4, 5, 6\}$
- $p(x = 1) = p(x = 2) = \cdots = p(x = 6) = \frac{1}{6}$
- $\sum_x p(x) \equiv p(x = 1) + p(x = 2) + \cdots + p(x = 6) = 1$

Tricked Dice

- x denotes the outcome of the dice
- $\mathrm{dom}(x) = \{1, 2, 3, 4, 5, 6\}$
- $p(x = 1) = 0, p(x = 2) = \frac{1}{3}, p(x = 3) = 0, p(x = 4) = \frac{1}{3}, p(x = 5) = 0, p(x = 6) = \frac{1}{3}$
- $\sum_x p(x) \equiv p(x = 1) + p(x = 2) + \cdots + p(x = 6) = 1$

Score

- x denotes your final score for this module
- $\mathrm{dom}(x) = \{A_1, A_2, B_1, \ldots, F\}$
- $p(x = A_1) =? \; p(x = A_2) =? \cdots$
- $\sum_x p(x) = 1$

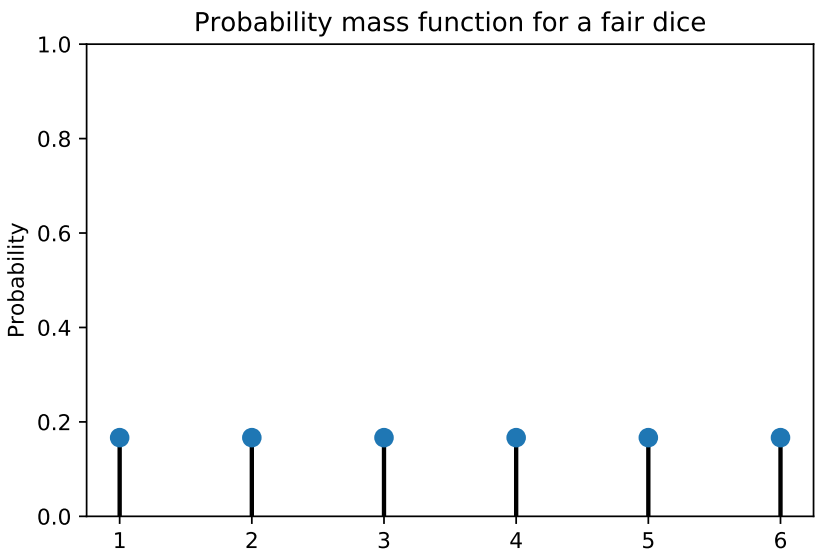
Probability is subjective (for instance, see the last example). However, no matter how you assign the numerical values $p(x = A_1), p(x = A_2)$ etc.., they must be nonnegative numbers and sum up to one!

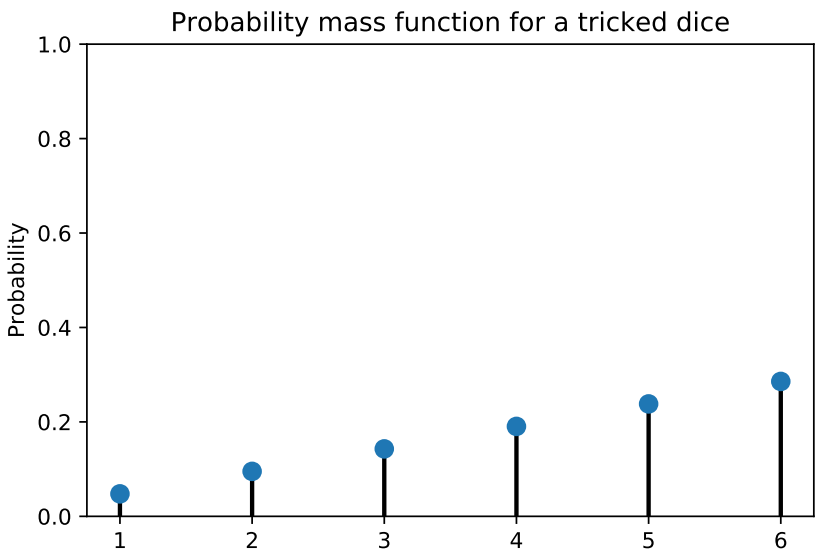
Probability mass function

In the (previous and) next slides, we will consider scenarios where the set of possible outcomes $\mathrm{dom}(x)$ is discrete, such as a coin toss or a roll of dice.
In this case, a (discrete) probability distribution can be fully specified by a discrete list of the probabilities of the outcomes, called *Probability Mass Function* (PMF).

A PMF is the list of all probabilities $p(x = \mathrm{x})$ for each element x of $\mathrm{dom}(x)$.

We can plot it, as shown in the example:





Probability mass function

From the PMF, we can compute the probability

$$p(x \in A)$$

for any $A \subseteq \mathrm{dom}(x)$.

Fair Dice

Probability of $A = \{2, 4, 6\}$ (the outcome is an even number) is

$$p(x \in A) = \sum_{\mathsf{x} \in A} p(x = \mathsf{x}) = p(x = 2) + p(x = 4) + p(x = 6) = \frac{1}{2}$$

Elementary Set theory

From the definition

$$p(x \in A) = \sum_{\mathsf{x} \in A} p(x = \mathsf{x})$$

and elementary set theory, we can derive that

$$p(x \in A \cup B) = p(x \in A) + p(x \in B) - p(x \in A \cap B)$$

for every $A, B \subseteq \mathrm{dom}(x)$, and

$$p(x \in A^c) = 1 - p(x \in A)$$

where A^c is the complementary of A, that is $A^c = \mathrm{dom}(x) \setminus A$. Here, $\setminus$ denotes the operation of set difference.

Boolean logic

Note that

$$x \in A, \quad x \in B, \quad x \in A^c$$

are events: they can be **true** or **false**.

Hence, we may want to compute

$$p(x \in A \text{ and } x \in B),\ p(x \in A \text{ or } x \in B),\ p(x \notin A)$$

Again from elementary Boolean logic and set theory, we can derive that

$$\begin{aligned} p(x \in A \text{ and } x \in B) &= p(x \in A \cap B) \\ p(x \in A \text{ or } x \in B) &= p(x \in A \cup B) \\ p(x \notin A) &= p(x \in A^c) \end{aligned}$$

Note that, $p(x \in A \text{ and } x \in B)$ is often denoted as $p(x \in A, x \in B)$.

Example

Fair Dice

$\text{dom}(x) = \{1, 2, 3, 4, 5, 6\}$, $A = \{1, 3, 5\}$ $B = \{1, 2\}$: Note that $B^c = \{3, 4, 5, 6\}$ and so

$$p(x \in B^c) = 1 - P(x \in B) = 1 - \frac{1}{3} = \frac{2}{3}$$

We can also compute

$$p(x \in A \text{ or } x \in B) = p(x \in A \cup B) = p(x \in \{1, 2, 3, 5\})$$

$$\begin{aligned} p(x \in A \cup B) &= p(x \in A) + p(x \in B) - p(x \in A \cap B) \\ &= p(x \in \{1, 3, 5\}) + p(x \in \{1, 2\}) - p(x \in \{1\}) \\ &= \frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3} \end{aligned}$$

$$p(x \in A \text{ and } x \in B) = p(x \in A \cap B) = p(x \in \{1\})$$

Conditional Probability and Bayes' Rule

The probability of event $x \in A$ conditioned on knowing event $x \in B$ (or more shortly, the probability of $x \in A$ given $x \in B$) is defined as

$$p(x \in A | x \in B) \equiv \frac{p(x \in A, x \in B)}{p(x \in B)} \quad \textit{this equation is also called Bayes' Rule}$$

Fair dice

$$p(x = 5 | x \textit{ is odd}) = \frac{p(x = 5, x \textit{ is odd})}{p(x \textit{ is odd})} = \frac{p(x = 5)}{p(x \textit{ is odd})} = \frac{1/6}{1/2} = 1/3$$

Throwing darts

$$p(\text{region } 5 | \text{not region } 20) = \frac{p(\text{region } 5, \text{not region } 20)}{p(\text{not region } 20)} = \frac{1/20}{19/20} = \frac{1}{19}$$

Interpretation

$p(x \in A | x \in B)$ should not be interpreted as 'Given the event $x \in B$ has occurred, $p(x \in A | x \in B)$ is the probability of the event $x \in A$ occurring'. The correct interpretation should be '$p(x \in A | x \in B)$' is the probability of x being in A under the **constraint** that x is in B.

The conditional probability is a probability

The conditional probability satisfies the rules of probability:

Definition

1. $p(\cdot|B) \in [0,1]$;
2. $p(x \in B|B) = 1$.
3. Given m events $A_1, \ldots, A_m$ in B such that $A_i \cap A_j = \emptyset$ for all $i \neq j = 1, \ldots, m$ it follows that:

$$p\left(x \in \bigcup_{i=1}^{m} A_i \middle| B\right) = \sum_{i=1}^{m} p\left(x \in A_i | B\right)$$

For instance, for a dice:

$$p(1 \cup 3|odd) = p(1|odd) + p(3|odd)$$

Similarly:

$$p(1 \cup 3 \cup 5|odd) = p(1|odd) + p(3|odd) + p(5|odd) = p(odd|odd) = 1$$

Example

Example

Consider a 52 cards deck. It consist of 4 suits: hearts, diamonds, spades and clubs. Each suit further contains 13 cards: 10 ace cards (A to 10) and 3 picture cards: Jack, Queen, and King.

Note that

$$p(hearts|queen) = \frac{1}{4}$$

while

$$p(queen|hearts) = \frac{1}{13}$$

In general, $p(A|B) \neq p(B|A)$. However, $p(A|B)$ and $p(B|A)$ seem to be related in some way. Bayes' rule tells us how.

Two or more variables

We will go back to Bayes' Rule later on.

Now we will consider the case we have two or more variables:

- two dices
- three coins
- scores and number of studying days.

Two dices

2 Fair Dices

- z denotes the outcomes of the two dices
- $\mathrm{dom}(z) = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\ldots,(6,4),(6,5),(6,6)\}$ (it has 36 elements).
- $p(z=(1,1)) = p(z=(1,2)) = \cdots = p(z=(6,6)) = \frac{1}{36}$
- $\sum_z p(z) = 1$

Note that, in this case, we cannot talk about "the outcome of the red dice" because for instance in $(1,2)$ we do not know if 1 is for red or blue.
To solve this problem, we can introduce two variables

- x denotes the outcome of the red dice
- $\mathrm{dom}(x) = \{1,2,3,4,5,6\}$
- y denotes the outcome of the blue dice
- $\mathrm{dom}(y) = \{1,2,3,4,5,6\}$

We can then define

- $z = (x,y)$ (all possible pairs, but now we know that (1,2) means 1 on red dice and 2 on blue dice)
- $\mathrm{dom}(z) = \mathrm{dom}(x) \times \mathrm{dom}(y)$, where $\times$ denotes the Cartesian product.

Marginal and Joint distribution

---

joint

Given two (or more) variables $p(x, y)$ is called the joint probability of x and y

---

marginalisation

Given a joint distr. $p(x, y)$ the marginal distr. of x is defined by

$$p(x = \mathsf{x}) = \sum_{\mathsf{y} \in \mathrm{dom}(y)} p(x = \mathsf{x}, y = \mathsf{y})$$

More generally,

$$p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \sum_{x_i} p(x_1, \ldots, x_n)$$

$$p(x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \sum_{x_1, x_i} p(x_1, \ldots, x_n)$$

$$p(x_1) = \sum_{x_2, \ldots, x_n} p(x_1, \ldots, x_n)$$

Example

Two fair dices:

$$p(x = 1, y = 6) \equiv p(x = 1 \text{ AND } y = 6)$$

They are equivalent ways to denote "probability that $x = 1$ and $y = 6$.

Given the PMF $p(x = i, y = j)$ for $i, j = 1, 2, 3, 4, 5, 6$, how can we compute $p(x = 1)$?

$p(x = 1)$ is called *marginal probability* and can be obtained as

$$\begin{aligned} p(x = 1) = \sum_y p(x = 1, y) \equiv\ & p(x = 1, y = 1) + p(x = 1, y = 2) \\ & + p(x = 1, y = 3) + p(x = 1, y = 4) \\ & + p(x = 1, y = 5) + p(x = 1, y = 6) \end{aligned}$$

Independence

Variables x and y are independent if knowing one event gives no extra information about the other event. Mathematically, this is expressed by

$$p(x = \mathsf{x}, y = \mathsf{y}) = p(x = \mathsf{x})p(y = \mathsf{y})$$

Independence of x and y is equivalent to

$$p(x = \mathsf{x}|y = \mathsf{y}) = p(x = \mathsf{x}) \Leftrightarrow p(y = \mathsf{y}|x = \mathsf{x}) = p(y = \mathsf{y})$$

for all $\mathsf{x} \in \mathrm{dom}(x)$, $\mathsf{y} \in \mathrm{dom}(y)$, then the variables x and y are said to be independent. We write then $x \perp\!\!\!\perp y$.

---

interpretation

Note that $x \perp\!\!\!\perp y$ doesn't mean that, given y, we have no information about x. It means the only information we have about x is contained in $p(x)$.

---

factorisation

If

$$p(x, y) = kf(x)g(y)$$

for some constant k, and positive functions $f(\cdot)$ and $g(\cdot)$ then x and y are independent.

Conditional Independence

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$$

denotes that the two sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are independent of each other given the state of the set of variables $\mathcal{Z}$. This means that

$$p(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})p(\mathcal{Y}|\mathcal{Z}) \text{ and } p(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})$$

for all states of $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. In case the conditioning set is empty we may also write $\mathcal{X} \perp\!\!\!\perp \mathcal{Y}$ for $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \emptyset$, in which case $\mathcal{X}$ is (unconditionally) independent of $\mathcal{Y}$.

Conditional independence does not imply marginal independence

$$p(x, y) = \sum_z \underbrace{p(x|z)p(y|z)}_{\text{cond. indep.}} p(z) \neq \underbrace{\sum_z p(x|z)p(z)}_{p(x)} \underbrace{\sum_z p(y|z)p(z)}_{p(y)}$$

Conditional dependence

If $\mathcal{X}$ and $\mathcal{Y}$ are not conditionally independent, they are conditionally dependent. This is written

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} | \mathcal{Z}$$

Conditional Probability and Bayes' Rule

We can generalise Bayes' Rule to two (or more variables).

Fair dice

Two dices: x denotes the outcome of the red dice and y the outcome of the blue dice.

$$p(x=5, y=3|x+y=8) = \frac{p(x=5, y=3, x+y=8)}{p(x+y=8)} = \frac{p(x=5, y=3)}{p(x+y=8)} = \frac{1/36}{5/36} = \frac{1}{5}$$

Remember the Interpretation

$p(A = \text{a}|B = \text{b})$ should not be interpreted as 'Given the event $B = \text{b}$ has occurred, $p(A = \text{a}|B = \text{b})$ is the probability of the event $A = \text{a}$ occurring'. The correct interpretation should be '$p(A = \text{a}|B = \text{b})$ is the probability of A being in state a under the constraint that B is in state b'.

Conditional Probability and Bayes' Rule

The probability of event x conditioned on knowing event y (or more shortly, the probability of x given y) is defined as

$$p(x|y) \equiv \frac{p(x,y)}{p(y)}$$

The probability of event y conditioned on knowing event x (or more shortly, the probability of y given x) is defined as

$$p(y|x) \equiv \frac{p(x,y)}{p(x)}$$

From the latter, we can derive that

$$p(x,y) = p(y|x)p(x)$$

and if we replace it in the first one, we have

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\sum_x p(x,y)} = \frac{p(y|x)p(x)}{\sum_x p(y|x)p(x)}$$

This tells us $p(x|y)$ and $p(y|x)$ are related.

Scientific Inference

Much of science deals with problems of the form : tell me something about the variable θ given that I have observed data $\mathcal{D}$ and have some knowledge of the underlying data generating mechanism. Our interest is then the quantity

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

This shows how from a forward or *generative model* $p(\mathcal{D}|\theta)$ of the dataset, and coupled with a *prior* belief $p(\theta)$ about which variable values are appropriate, we can infer the *posterior* distribution $p(\theta|\mathcal{D})$ of the variable in light of the observed data.

Generative models in science

This use of a generative model sits well with physical models of the world which typically postulate how to generate observed phenomena, assuming we know the model. For example, one might postulate how to generate a time-series of displacements for a swinging pendulum but with unknown mass, length and damping constant. Using this generative model, and given only the displacements, we could infer the unknown physical properties of the pendulum, such as its mass, length and friction damping constant.

Prior, Likelihood and Posterior

For data $\mathcal{D}$ and variable θ, Bayes' rule tells us how to update our prior beliefs about the variable θ in light of the data to a posterior belief:

$$\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{p(\mathcal{D}|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathcal{D})}_{\text{evidence}}}$$

The evidence is also called the marginal likelihood.

$$p(\mathcal{D}) = \sum_{\theta \in \mathrm{dom}(\theta)} p(\mathcal{D}|\theta)p(\theta)$$

The term likelihood is used for the probability that a model generates observed data.

Is the coin fair?

- Denote with C a certain coin.
- Denote its bias with θ (the probability of getting Heads).
- **Query:** We want to know if C is fair (equivalently if $\theta = 0.5$)?
- For the moment, we consider the case in which only three values of θ are possible:
  - $\theta = 0.25$, $\theta = 0.5$, $\theta = 0.75$
- We **assume** that $p(\theta = 0.25) = 0.25$, $p(\theta = 0.5) = 0.5$, $p(\theta = 0.75) = 0.25$ (*prior model*).
- We cannot observe θ directly but we can toss the coin how many times we want.

Is the coin fair?

- We have that
  - $P(H|\theta) = \theta$;
  - $P(T|\theta) = 1 - \theta$.

  this is the *likelihood* model.
- Suppose we have tossed the coin 12 times with the following results: 3 Heads and 9 Tails.
- **Assume** that the tosses are independent, that is the result of one toss does not dependent on the result of the previous toss.
  Example:

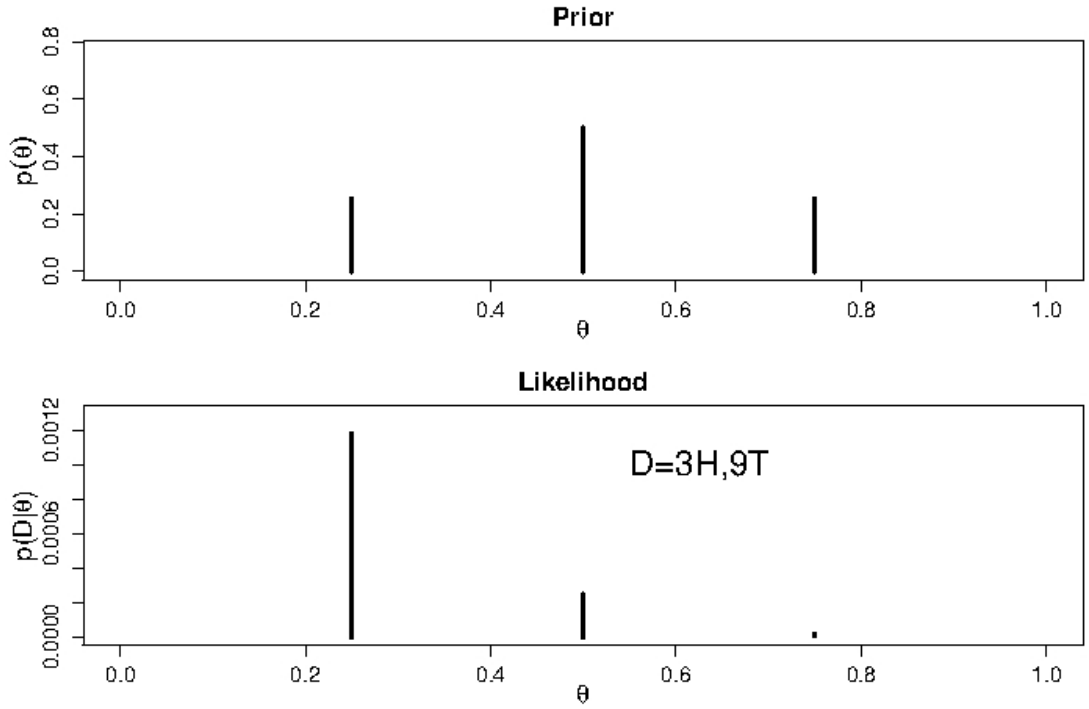
$$P(H, H|\theta) = \theta\theta, \quad P(H, T|\theta) = \theta(1-\theta), \quad P(H, H, T|\theta) = \theta^2(1-\theta)$$

- Then, the probability

$$P(3H, 9T|\theta) = \theta^3(1-\theta)^9$$

  We denote the observations (data) with D, in this case $D = \{3H, 9T\}$.

Is the coin fair?



For $\theta = 0.25$, we have that

$$P(3H, 9T|\theta = 0.25) = 0.25^3(0.75)^9 \approx 0.00117$$

These are the values in the second plot.

Bayes' rule

- We can apply Bayes' rule:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

- For instance, for $\theta = 0.25$.

$$P(\theta = 0.25|D) = \frac{P(D|\theta = 0.25)P(\theta = 0.25)}{P(D)} = \frac{(0.25)^3(0.75)^9 0.25}{P(D)} = \frac{0.0002933}{P(D)}$$

where we have seen that

$$\begin{aligned} P(D) &= P(D|\theta = 0.25)P(\theta = 0.25) + P(D|\theta = 0.5)P(\theta = 0.5) + P(D|\theta = 0.75)P(\theta = 0.75) \\ &= (0.25)^3(0.75)^9 0.25 + (0.5)^3(0.5)^9 0.5 + (0.75)^3(0.25)^9 0.25 = 0.0004556 \end{aligned}$$

- We can then get

$$P(\theta = 0.25|D) = \frac{0.0002933}{0.0004556} = 0.6438$$

Is the coin fair

- Similarly for the other values

$$P(\theta = 0.5|D) = \frac{P(D|\theta = 0.5)P(\theta = 0.5)}{P(D)} = \frac{(0.5)^3(0.5)^9 0.5}{P(D)} = \frac{0.0001954}{P(D)} = 0.2679$$

- Finally

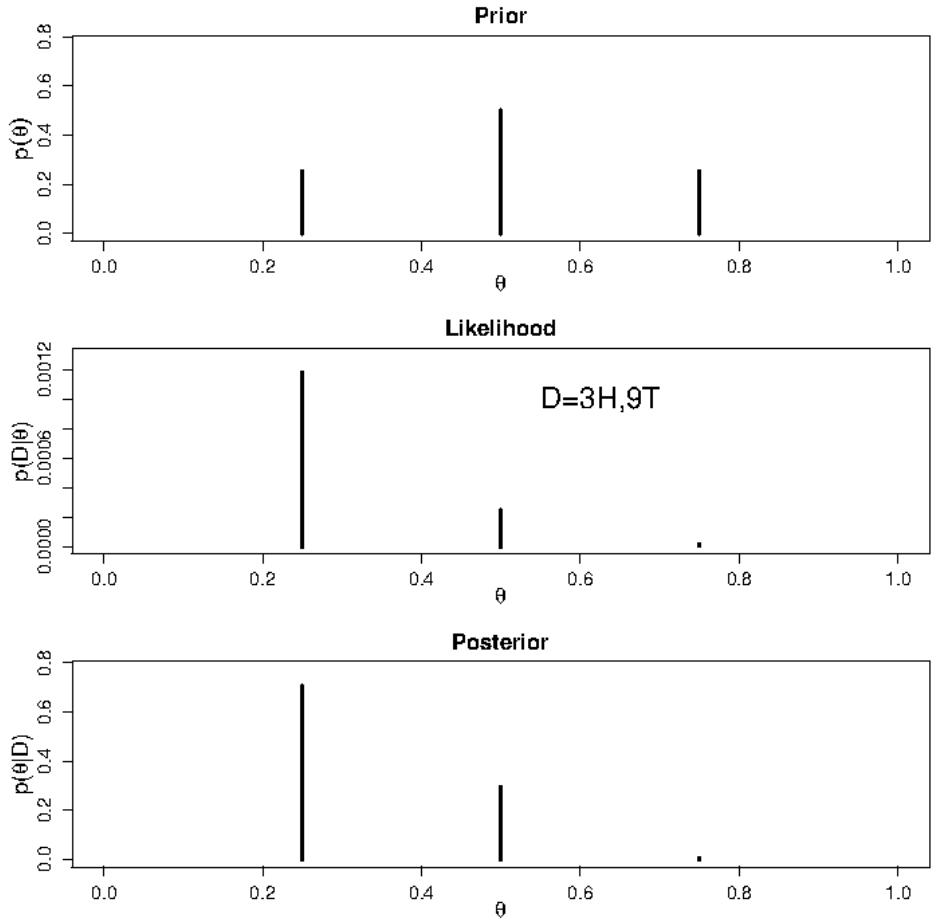
$$P(\theta = 0.75|D) = 1 - P(\theta = 0.5|D) - P(\theta = 0.25|D) = 0.0883$$

- Summing up

$$P(\theta = 0.25|D) = 0.6438, \quad P(\theta = 0.5|D) = 0.2679 \quad P(\theta = 0.75|D) = 0.0883$$

Therefore, given the data, we can say that the coin is biased towards Tail with higher probability.

Is the coin fair?



Expectation

Give a function $f(x)$ of the variable x, its expectation with respect the probability distribution of x is defined as:

$$E[f(x)] = \sum_{\mathsf{x} \in dom(x)} f(\mathsf{x}) p(x = \mathsf{x})$$

If

- $f(x) = x$ (identity function) then $E[x]$ is the mean of x;
- $f(x) = (x - E[x])^2$ then $E[(x - E[x])^2] = Var(x)$ is called the variance of x;
- $f(x) = (x == a)$ then $E[(x == a)]$ is the probability of $x = a$;
- $f(x) = (x >= a)$ then $E[(x >= a)]$ is the probability of $x \geq a$;

These definitions can easily be generalised to more than one variable.

Example Fair Dice

$$E[f(x)] = \sum_{\mathsf{x} \in dom(x)} f(\mathsf{x}) p(x = \mathsf{x})$$

$$E[x] = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$$

$$E[x^2] = 1\frac{1}{6} + 4\frac{1}{6} + 9\frac{1}{6} + 16\frac{1}{6} + 25\frac{1}{6} + 36\frac{1}{6} = 13.5$$

$$E[(x-3.5)^2] = (1-3.5)^2\frac{1}{6} + (2-3.5)^2\frac{1}{6} + (3-3.5)^2\frac{1}{6} + (4-3.5)^2\frac{1}{6} + (5-3.5)^2\frac{1}{6} + (6-3.5)^2\frac{1}{6} = \frac{35}{12}$$

$$E[x == 1] = 1\frac{1}{6} + 0\frac{1}{6} + 0\frac{1}{6} + 0\frac{1}{6} + 0\frac{1}{6} + 0\frac{1}{6} = \frac{1}{6}$$

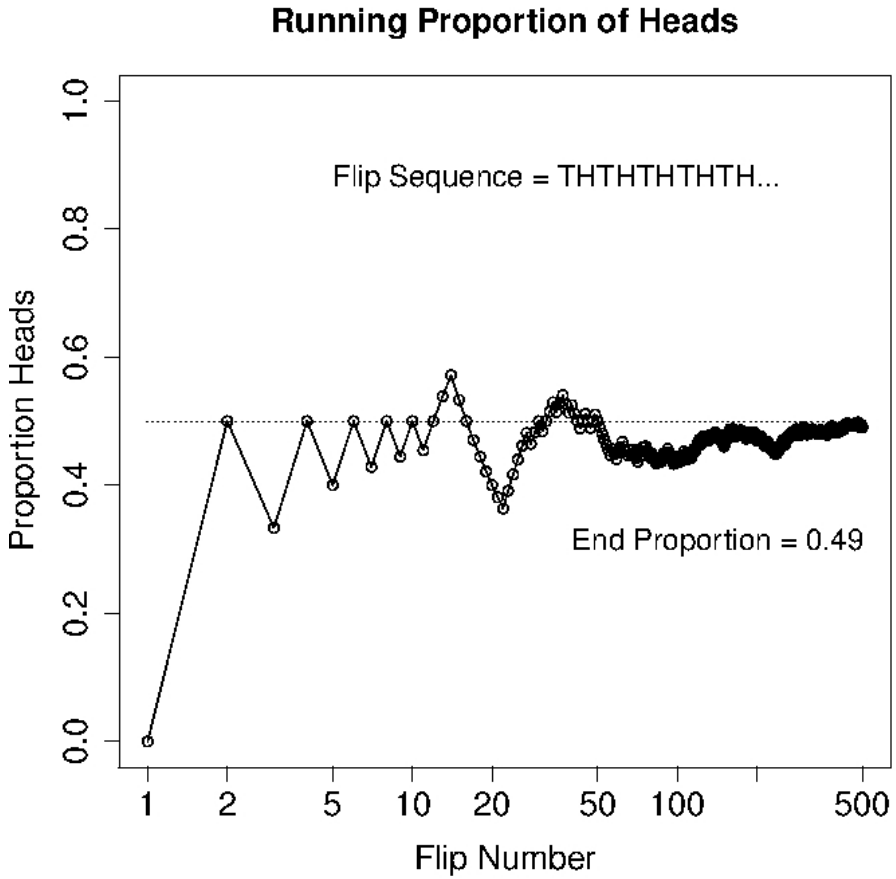
$$E[x >= 2] = 0\frac{1}{6} + 1\frac{1}{6} + 1\frac{1}{6} + 1\frac{1}{6} + 1\frac{1}{6} + 1\frac{1}{6} = \frac{5}{6}$$

What is the meaning of this number between 0 and 1 we call probability?

- There are two main interpretations of the meaning of probability:
  1. frequentist;
  2. subjective (or Bayesian).

- `https://plato.stanford.edu/entries/probability-interpret/`

Frequentist



$$p(x \in A) = \lim_{n \to \infty} \frac{n_A}{n}$$

where n_A is the number of times the event $x \in A$ happens and n is the total number of trials.

Subjective

Probability is a degree of belief!

It represents our uncertainty about some fact (variable) of the world (domain).

It allows us to express probability for events that are not repeatable:

- What is the probability of Hard-Brexit?
- What is the probability that you will score more than B3?

In spite of the fact that we are considering subjective beliefs, different agents can reach the same conclusion conditional on some evidence.

Frequentist vs. Subjective

Is it just a matter of interpretation?
Unfortunately No, statistics methods based on the frequentist interpretation are flawed, example:

- hypothesis testing based p-values;
- maximum likelihood estimation.

Does this affect ML?
Yes, standard ML approaches derive from the frequentist interpretation and, therefore, they inherit those flaws. Overfitting is an example of that and regularisation and dropout are patch solutions.

$$\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{\underbrace{p(\mathcal{D}|\theta)}_{\text{likelihood}}\underbrace{p(\theta)}_{\text{prior}}}^{}}{\underbrace{p(\mathcal{D})}_{\text{evidence}}}$$

Standard ML methodologies only consider $\arg\max_\theta p(\mathcal{D}|\theta)$, while instead we should consider $p(\theta|\mathcal{D})$.

Inspector Clouseau

Inspector Clouseau arrives at the scene of a crime. The Butler (B) and Maid (M) are his main suspects. The inspector has a prior belief of 0.6 that the Butler is the murderer, and a prior belief of 0.2 that the Maid is the murderer. These probabilities are independent in the sense that $p(B, M) = p(B)p(M)$. (It is possible that both the Butler and the Maid murdered the victim or neither). The inspector's *prior* criminal knowledge can be formulated mathematically as follows:

$$\mathrm{dom}(B) = \mathrm{dom}(M) = \{\text{murderer}, \text{not murderer}\}$$

$$\mathrm{dom}(K) = \{\text{knife used}, \text{knife not used}\}$$

$$p(B = \text{murderer}) = 0.6, \qquad p(M = \text{murderer}) = 0.2$$

$$\begin{aligned}
p(\text{knife used}|B = \text{not murderer}, \quad & M = \text{not murderer}) && = 0.3 \\
p(\text{knife used}|B = \text{not murderer}, \quad & M = \text{murderer}) && = 0.2 \\
p(\text{knife used}|B = \text{murderer}, \quad & M = \text{not murderer}) && = 0.6 \\
p(\text{knife used}|B = \text{murderer}, \quad & M = \text{murderer}) && = 0.1
\end{aligned}$$

The victim lies dead in the room and the inspector quickly finds the murder weapon, a Knife (K). What is the probability that the Butler is the murderer? (Remember that it might be that neither is the murderer).

Inspector Clouseau

Using b for the two states of B and m for the two states of M,

$$p(B|K) = \sum_m p(B, m|K) = \sum_m \frac{p(B, m, K)}{p(K)} = \frac{p(B)\sum_m p(K|B, m)p(m)}{\sum_b p(b)\sum_m p(K|b, m)p(m)}$$

Plugging in the values we have

$$\begin{aligned} p(B = \text{murderer}|\text{knife used}) &= \frac{\frac{6}{10}\left(\frac{2}{10}\times\frac{1}{10} + \frac{8}{10}\times\frac{6}{10}\right)}{\frac{6}{10}\left(\frac{2}{10}\times\frac{1}{10} + \frac{8}{10}\times\frac{6}{10}\right) + \frac{4}{10}\left(\frac{2}{10}\times\frac{2}{10} + \frac{8}{10}\times\frac{3}{10}\right)} \\ &= \frac{300}{412} \approx 0.73 \end{aligned}$$

Hence knowing that the knife was the murder weapon strengthens our belief that the butler did it.

Exercise: compute the probability that the Butler and not the Maid is the murderer.

Inspector Clouseau

The role of $p(\text{knife used})$ in the Inspector Clouseau example can cause some confusion. In the above,

$$p(\text{knife used}) = \sum_b p(b) \sum_m p(\text{knife used}|b, m)p(m)$$

is computed to be 0.456. But surely, $p(\text{knife used}) = 1$, since this is given in the question!

Note that the quantity $p(\text{knife used})$ relates to the *prior* probability the model assigns to the knife being used (in the absence of any other information). If we know that the knife is used, then the *posterior* is

$$p(\text{knife used}|\text{knife used}) = \frac{p(\text{knife used}, \text{knife used})}{p(\text{knife used})} = \frac{p(\text{knife used})}{p(\text{knife used})} = 1$$

which, naturally, must be the case.

Inspector Clouseau

Assume that Inspector Clouseau has found another evidence, a male jacket button T, and he knows that

$$\begin{aligned} p(T|B = \text{not murderer}) &= 0.6 \\ p(T|B = \text{murderer}) &= 0.2 \end{aligned}$$

What is the probability that the Butler is the murderer?

Inspector Clouseau

We aim to compute:

$$p(B = \text{murderer}|K, T) = \frac{p(K, T|B = \text{murderer})p(B = \text{murderer})}{p(K, T)}$$

we can assume that conditional independence

$$p(K, T|B = \text{murderer}) = p(K|B = \text{murderer})p(T|B = \text{murderer})$$

Do we need to recompute everything from scratch? No

$$p(B = \text{murderer}|K, b) = p(T|B = \text{murderer})\frac{p(B = \text{murderer}|K)p(K)}{p(K, T)}$$

Also note that we can always rewrite $p(K, b) = p(b|K)p(K)$ and, therefore

$$p(B = \text{murderer}|K, T) = p(T|B = \text{murderer})\frac{p(B = \text{murderer}|K)}{p(T|K)}$$

Inspector Clouseau

$$p(B = \text{murderer}|K, T) = p(T|B = \text{murderer})\frac{p(B = \text{murderer}|K)}{p(T|K)}$$

Note that

$$p(T|K) = \sum_b p(T|B = b)p(B = b|K) = 0.6 \cdot 0.27 + 0.2 \cdot 0.73 = 0.308$$

and so

$$p(B = \text{murderer}|K, T) = \frac{0.2 \cdot 0.73}{0.308} = 0.474$$

Bayes' rule

Bayes' rule is an "iterative" procedure

$$p(\theta|D_1, D_2) = \frac{p(D_2|\theta)p(D_1|\theta)p(\theta)}{p(D_2, D_1)}$$

we can always write it as

$$p(\theta|D_1, D_2) = \frac{p(D_2|\theta)p(\theta|D_1)}{p(D_2|D_1)}$$

where $p(\theta|D_1)$ is the posterior given the data D_1, and so on iteratively:

$$p(\theta|D_1, D_2, D_3) = \frac{p(D_3|\theta)p(\theta|D_1, D_2)}{p(D_3|D_1, D_2)}$$