Aalto University

Probabilistic machine learning | Intro (B)

Introduction to machine learning

Francesco Corona

Chemical and Metallurgical Engineering
School of Chemical Engineering

Probabilities

Probability theory | Intro (B)

Probability theory

The key concept in modelling is that of accounting for **uncertainty** using probabilities

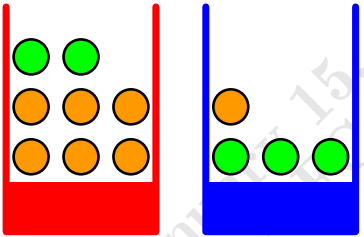
- It gets in the way (we introduce it) through noise on measurable quantities
- It gets in the way (we introduce it) through the finiteness size of datasets

Probability theory provides the framework for quantifying/manipulating uncertainties

- Applied probability theory is central in probabilistic machine learning

Probability theory | Example

Let us suppose that we have two boxes, one of which is red and the other one is blue



- Red box,: 2 apples and 6 oranges
- Blue box: 3 apples and 1 orange

We randomly pick one box and, from it we randomly select one item of fruit

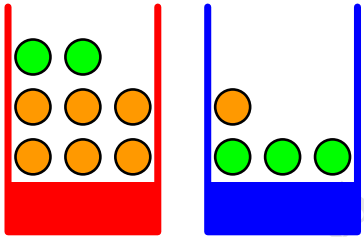
- We firstly check the fruit
- Then we place it back

We repeat this process *many* times

In the experiment, 40% of the time we pick the red box, 60% of the time the blue one

- We are equally likely to select any of the pieces of fruit from the box

Probability theory | Example (cont.)



The **identity of the box**: **Random variable** B

- B can take one of two values
- r (red box) or b (blue box)

The **identity of the fruit Random variable** F

- F can take one of two values:
- a (apple) or o (orange)

The **probability of an event**: Fraction of times the event occurs, out of the number trials

- *In the limit* that the total number of trials goes to infinity

In the experiment

- The probability of selecting the blue box is 6/10
- The probability of selecting the red box is 4/10

We write the probabilities of picking a box

- $p(B = r) = 4/10$
- $p(B = b) = 6/10$

Probability theory | Example (cont.)

We have defined our experiment and we also have the probabilities for certain events

We can start asking (probability) questions about the system

- What is the overall probability that the selection procedure picks an apple?
- Given that we have picked an orange, what is the probability that the box we picked was the blue one?
- ...

We can answer questions such as these and more complex ones arising in machine learning, once we have equipped ourselves with the **two elementary rules of probability**

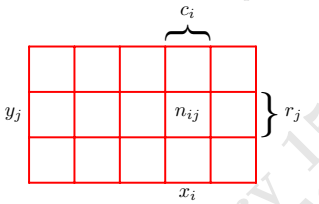
- The **sum rule** and the **product rule**

Rules

Probability theory | Intro (B)

Probability theory | Rules (cont.)

In order to derive the rules of probability, we consider the slightly more general example



Two random variables X and Y

- X can take any value x_i, $i = 1, \ldots, N_i$
- Y can take any value y_j, $j = 1, \ldots, N_j$

Here, $N_i = 5$ and $N_j = 3$

N **trials** in which we sample both X and Y

Let n_{ij} be the number of such trials in which $X = x_i$ and $Y = y_j$

Let n_i (c_i) be the number of trials in which X takes the value x_i

- (irrespective of the value that Y takes)

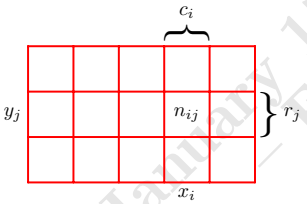
Let n_j (r_j) be the number of trials in which Y takes the value y_j

- (irrespective of the value that X takes)

Probability theory | Rules | Joint

The probability that X takes value x_i and Y takes value y_j is written $p(X = x_i, Y = y_j)$

- This is the **joint probability** of both $X = x_i$ and $Y = y_j$



It is the number n_{ij} of points falling in cell (i, j) as fraction of the number N of points

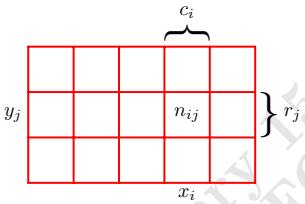
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Implicitly, we are assuming a limit $N \to \infty$

Probability theory | Rules | Marginals

The probability that X takes some value x_i, irrespective of the value of Y, is $p(X = x_i)$

- The fraction of number of points falling in the cells of column i



$$
\begin{aligned}
p(X = x_i) &= n_i / N \\
&= \frac{\sum_{j=1}^{N_j} n_{ij}}{N} \\
&= \sum_{j=1}^{N_j} \underbrace{n_{ij}/N}_{p(X=x_i, Y=y_j)} \\
&= \sum_{j=1}^{N_j} p(X = x_i, Y = y_j)
\end{aligned}
$$

$p(X = x_i)$ is the **marginal probability** because it obtained by the process of marginalisation, or by summing out, the probabilities of all the other variables (in this case, Y)

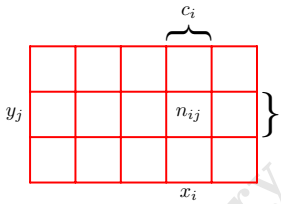
From marginal probability to the **sum rule**

$$
p(X = x_i) = \sum_{j=1}^{N_j} p(X = x_i, Y = y_j)
$$

Probability theory | Rules | Conditionals

The probability that $Y = y_j$, given that X takes some value x_i, is $p(Y = y_j|X = x_i)$

- This is the **conditional probability** that $Y = y_j$, given that $X = x_i$



It is obtained by finding the fraction of points in column i that fall in cell (i, j)

$$p(Y = y_i|X = x_i) = \frac{n_{ij}}{n_i}$$

From the joint and conditional probability to the **product rule**

$$\begin{aligned} p(X = x_i, Y = y_j) &= n_{ij}/N \\ &= \underbrace{\frac{n_{ij}}{n_i}}_{p(Y=y_j|X=x_i)} \underbrace{\frac{n_i}{N}}_{p(X=x_i)} \\ &= p(Y = y_j|X = x_i)p(X = x_i) \end{aligned}$$

Probability theory | Rules (cont.)

The rules of probability

- **Sum rule**

$$p(X) = \sum_Y p(X, Y)$$

$$p(Y) = \sum_X p(X, Y)$$

- **Product rule**

$$\begin{aligned} p(X, Y) &= p(Y|X)p(X) \\ &= p(X|Y)p(Y) \end{aligned}$$

To make the notation compact, $p(\star)$ denotes a distribution over a random variable $\star$

- $p(X, Y)$ is a joint probability, the probability of X and Y
- $p(Y|X)$ is a conditional probability, the probability of Y given X
- $p(X)$ is a marginal probability, the probability of X

Probability theory | Rules | Bayes

From the product rule and the symmetry property $p(X, Y) = p(Y, X)$, we obtain

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

This is a relationship between conditional probabilities known as the **Bayes' rule**

The denominator can be expressed using quantities appearing in the numerator

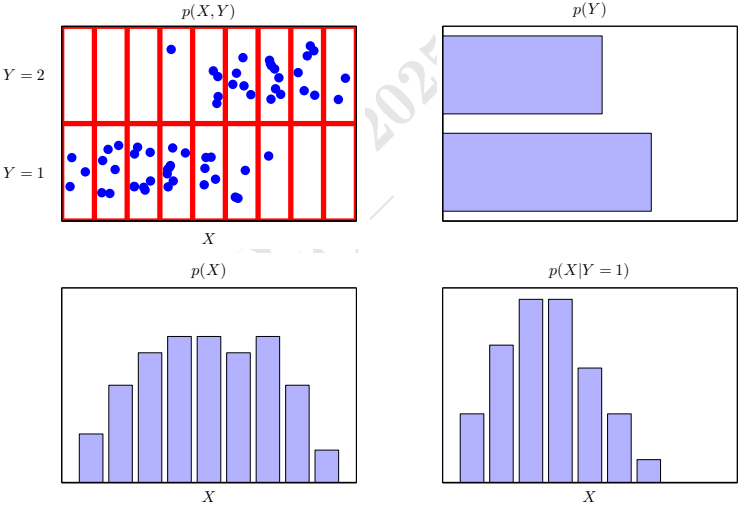
$$p(X) = \sum_Y p(X|Y)p(Y)$$

The denominator is a normalisation constant that ensures that the sum of the conditional probabilities on the left-hand side over all values of Y is one, for all values of X

If the joint distribution of two variables X and Y factorises into the product of the marginals (that is, $p(X, Y) = p(X)p(Y)$), then X and Y are said to be **independent**

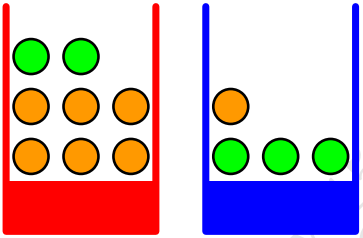
$$p(Y|X) = p(Y)$$
$$p(X|Y) = p(X)$$

Probability theory | Rules |Recap



Probability theory | Example | Rules

Back to the example of fruit boxes, the probability of selecting either red or blue boxes



$$p(B = r) = 4/10$$
$$p(B = b) = 6/10$$

These probabilities satisfy the closure condition

$$\sum_B p(B) = \underbrace{p(B = r)}_{4/10} + \underbrace{p(B = b)}_{6/10}$$
$$= 1$$

Suppose that we pick a box at random (say, the blue box) then the probability of selecting an apple is the fraction of apples in the blue box which is 3/4, thus we have

$$p(F = a|B = b) = 3/4$$

We can write all conditional probabilities of selecting the types of fruit, given the box

$$p(F = a|B = r) = 1/4$$
$$p(F = a|B = b) = 3/4$$
$$p(F = o|B = r) = 3/4$$
$$p(F = o|B = b) = 1/4$$

These probabilities satisfy the closure conditions

$$p(F = a|B = r) + p(F = o|B = r) = 1$$
$$p(F = a|B = b) + p(F = o|B = b) = 1$$

Probability theory | Example | Rules (cont.)

We use sum and product rules to evaluate the overall probability of picking an apple [1]

$$\begin{aligned} p(F=a) &= \sum_B p(F=a,B) \\ &= p(F=a,B=r) + p(F=a,B=b) \\ &= \underbrace{p(F=a|B=r)}_{1/4}\underbrace{p(B=r)}_{4/10} + \underbrace{p(F=a|B=b)}_{3/4}\underbrace{p(B=b)}_{6/10} \\ &= 11/20 \end{aligned}$$

From which (by sum rule), the probability of picking an orange is $p(F=o) = \underbrace{1-11/20}_{9/20}$

[1] $P(F) = \sum_B p(B,F)$, where $p(B,F) = p(F|B)p(B)$ and $p(B,F) = p(B|F)p(F)$

Probability theory | Example | Rules | Bayes

Suppose instead we are told that an item of fruit was selected and it is an orange $F = o$

- We would like to know which box the orange came from, $P(B|F = o)$

We are interested in the probability over boxes conditioned on the identity of the fruit

$$P(B|F)$$

Earlier, we evaluated the probability over fruits conditioned on the identity of the box

$$P(F|B)$$

We solve the problem of reversing the conditional probability, the Bayes' rule

$$p(B = r|F = o) = \frac{\underbrace{p(F = o|B = r)}_{3/4}\underbrace{p(B = r)}_{4/10}}{\underbrace{p(F = o)}_{9/20}}$$

$$= 2/3$$

From which (by sum rule), the probability that the orange is taken from the blue box

$$p(B = b|F = o) = \underbrace{1 - 2/3}_{1/3}$$

Probability theory | Example | Rules | Bayes (cont.)

If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have is the probability

$$p(B)$$

We call this the **prior probability** of B

- It is the probability available before we observe the identity of the fruit

Once we know that the fruit is an orange, we use Bayes' rule to compute the probability

$$p(B|F)$$

We call this the **posterior probability** of B given F

- It is the probability obtained 'after' we observed the identity of the fruit

The prior probability of selecting the red box was 4/10 (the blue box is more probable)

Once we observed that the picked fruit is an orange, the posterior probability of the red box is 2/3 (the red box is more probable to be the one the orange was picked from)

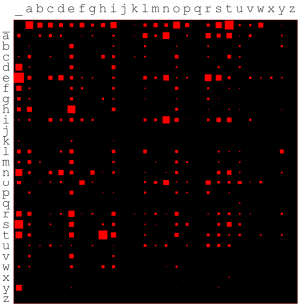
Probability theory | Example

A popular book in the English language (C. R. Darwin: *On the origin of species*, 1859)

$p(x)$

1	0.15754	_
2	0.06715	a
3	0.01420	b
4	0.02950	c
5	0.03120	d
6	0.11100	e
7	0.02335	f
8	0.01536	g
9	0.04191	h
10	0.06259	i
11	0.00060	j
12	0.00310	k
13	0.03530	l
14	0.02115	m
15	0.06032	n
16	0.06091	o
17	0.01601	p
18	0.00077	q
19	0.05287	r
20	0.05785	s
21	0.07597	t
22	0.02158	u
23	0.00997	v
24	0.01347	w
25	0.00209	x
26	0.01387	y
27	0.00039	z

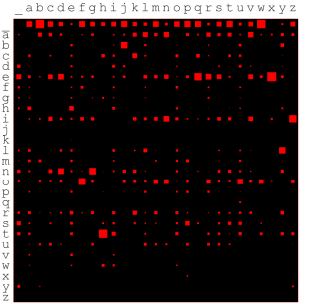
$p(x, y)$



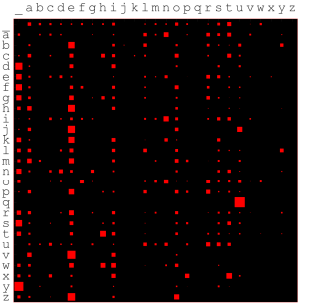
- The probability distribution over the 27 possible letters
- The probability distribution over the 27×27 bigrams

Probability theory | Example (cont.)

$p(x|y)$



$p(y|x)$



- The conditional probability distribution of the first letter, given the second one
- The conditional probability distribution of the second letter, given the first one

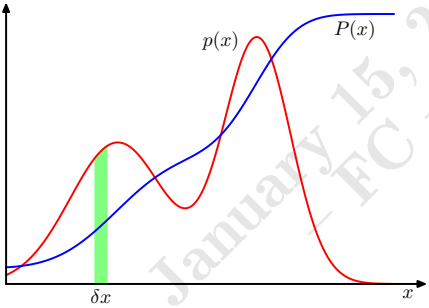
Probability densities

Probability theory | Intro (B)

Probability densities

We also wish to consider probabilities with respect to variables which are continuous

If the probability of a certain real-valued variable X falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for some $\delta x \to 0$, then $p(x)$ is called the **probability density** over x



The probability that x is in (a, b)

$$p(x \in (a, b)) = \int_a^b p(x)\,dx$$

Interpretation of density functions

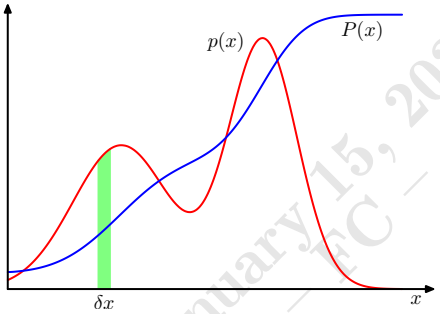
$$p(x \in (a - \frac{\delta x}{2}, a + \frac{\delta x}{2})) = \int_{a-\delta x/2}^{a+\delta x/2} p(x)\,dx \approx p(a)\delta x$$

The probability that x is in a δx-wide interval around a is approximately $p(a)\delta x$

- $p(a)$ is a measure of how likely it is that random variable X is around a

Probability densities (cont.)

Probabilities are nonnegative quantities and, because the value of x must lie somewhere on the real axis, we have that the probability density $p(x)$ must satisfy two conditions



$$p(x) \geq 0$$

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

The probability that x lies in $(-\infty, a)$ is given by the **cumulative distribution function**

$$P(x) = \int_{-\infty}^{a} p(x)\,dx$$

Density $p(x)$ is the derivative of the cumulative distribution function $P(x)$:

$$\frac{d}{dx}P(x) = p(x)$$

Probability densities | Multivariate

Consider a collection of several continuous variables $X_1, \ldots, X_D$ collected in vector X

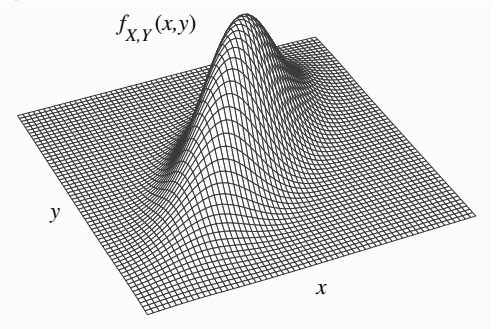
We define a **joint probability density** $p(x) = p(x_1, \ldots, x_D)$ such that the probability of point x falling in an infinitesimal volume δx around x is given by the product $p(x)\delta x$

- Multivariate probability density must satisfy the usual two conditions

For two variables X and Y

$$p(x, y) \geq 0$$

$$\int p(x, y) dx dy = 1$$



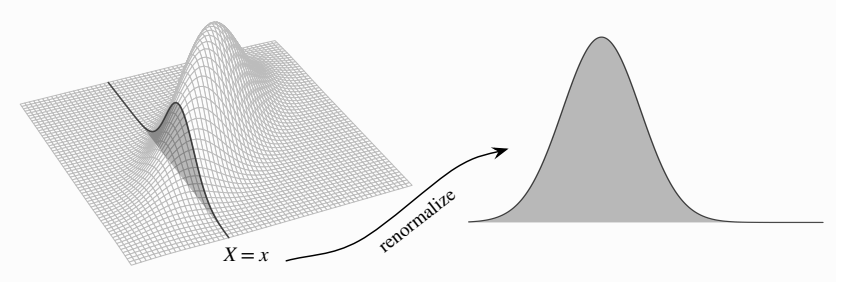
Probability densities | Multivariate (cont.)

Sum and product rules, together with the Bayes' rule, apply to probability densities

If X and Y are two real variables, then the sum and product rules take the form

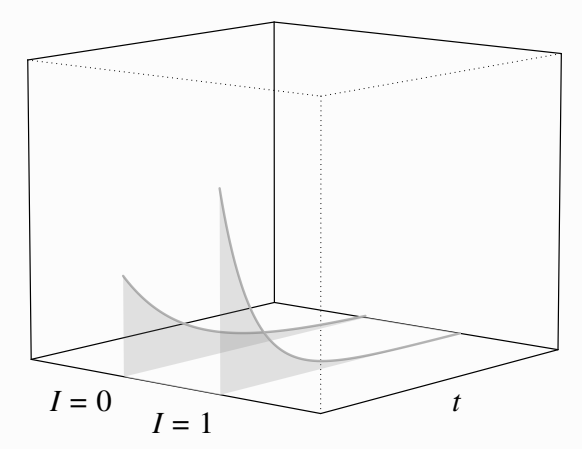
$$p(x) = \int p(x, y)\, dy$$

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



Probability densities | Multivariate (cont.)

We have joint probability distributions over discrete (I) and continuous variables (T)



Expectations

Probability densities

Expectations

A commonly used operation with probabilities is finding weighted averages of a function

- The average value of some function $f(x)$ under a probability distribution $p(x)$
- Such a quantity is called the **expectation** of $f(x)$ and it is denoted by $\mathbb{E}[f]$

For a discrete variable X, the average is weighted by the relative probabilities of x

$$\mathbb{E}[f] = \sum_X p(x) f(x)$$

For a continuous variable X, expectations are expressed in terms of an integration

$$\mathbb{E}[f] = \int p(x) f(x)\, dx$$

If we have a finite number N of points drawn from either a probability distribution or density, then the expectation can be approximated as a finite sum over those N points

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

The empirical approximation becomes exact in the limit $N \to \infty$

Expectations (cont.)

Usually we are interested in determining expectations of functions f of several variables

- We can use a subscript to indicate which variable(s) f is being averaged over

$\mathbb{E}_X[f(x,y)]$ is the average of function $f(x,y)$ with respect to the distribution of X

- $\mathbb{E}_X[f(x,y)] = \sum_X \underbrace{p(x)}_{\sum_Y p(x,y)} f(x,y)$, it is a function of y

$\mathbb{E}_Y[f(x,y)]$ is the average of function $f(x,y)$ with respect to the distribution of Y

- $\mathbb{E}_Y[f(x,y)] = \sum_Y \underbrace{p(y)}_{\sum_X p(x,y)} f(x,y)$, it is a function of x

$\mathbb{E}_{XY}[f(x,y)]$ is the average of $f(x,y)$ with respect to the distribution of X and Y

- $\mathbb{E}_{XY}[f(x,y)] = \sum_X \sum_Y p(x,y) f(x,y)$, it is a number

---

We can be interested **conditional expectation** with respect to a conditional distribution

$$\mathbb{E}_{X|Y}[f(x,y)] = \sum_X p(x|y) f(x,y)$$

Expectations | Variance and covariance

The measure of the variability of f around its expectation $\mathbb{E}[f(x)]$ is the **variance** of f

$$\text{var}[f] = \mathbb{E}\Big[\Big(f(x) - \mathbb{E}[f(x)]\Big)^2\Big]$$

The variance can also be written in terms of the expectations of functions f and f^2

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

The variance of the function $f(x) = x$,

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

Expectations | Variance and covariance (cont.)

For two variables X and Y, the extent to which they vary together is called **covariance**

$$\begin{aligned}\text{cov}[x, y] &= \mathbb{E}_{XY}\Big[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\Big] \\ &= \mathbb{E}_{XY}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

If X and Y are independent variables, then their covariance is zero

For two random vectors X and X, the covariance is a matrix

$$\begin{aligned}\text{cov}[x, y] &= \mathbb{E}_{XY}\Big[\Big(x - \mathbb{E}[x]\Big)\Big(y^T - \mathbb{E}[y^T]\Big)\Big] \\ &= \mathbb{E}_{XY}[xy^T] - \mathbb{E}[x]\mathbb{E}[y^T]\end{aligned}$$

Bayesian probabilities

Probability theory

Bayesian probabilities

We viewed probabilities as frequencies of repeatable random events

- It is the **frequentist** interpretation of probability

In general, we can view probabilities as quantification of uncertainty

- It is the **Bayesian** interpretation of probability

In the example of the boxes of fruit, an observation of the identity of the fruit F yield relevant information that allowed to update the probability of the box B it was from

- Bayes's theorem update a prior probability ($P(B = r) = 4/10$) into a posterior
- This is achieved by incorporating the evidence by the observed data

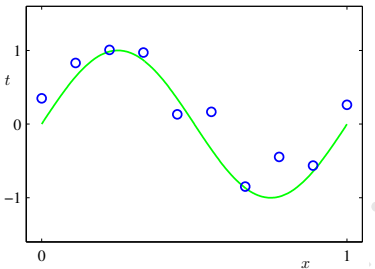
$$p(B = r|F = o) = \underbrace{\frac{p(F = o|B = r)p(B = r)}{p(F = o)}}_{2/3}$$

Bayesian probabilities (cont.)

The approach is general, we can adopt it when making inference about any quantities

- Say, ... the parameters w in the polynomial curve fitting example?

Bayesian probabilities (cont.)



Input data	Target data
x_1	t_1
$\vdots$	$\vdots$
x_n	t_n
$\vdots$	$\vdots$
x_N	t_N

Data

$$\mathcal{D} = \{t_1, \ldots, t_N\}$$

Model

$$y(x|w) = \sum_{n_m=0}^{N_m} w_{n_m} x^{n_m} + \text{noise}$$

Parameters

$$w = \{w_1, \ldots, w_{N_m}\}$$

We encode our assumptions about w, before observing $\mathcal{D}$ as a prior probability $p(w)$

- The effect of the observed $\mathcal{D}$ is expressed as the conditional probability $p(\mathcal{D}|w)$
- We evaluate w, after observing $\mathcal{D}$, as the posterior probability $p(w|\mathcal{D})$

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

Bayesian probabilities (cont.)

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

The quantity $p(\mathcal{D}|w)$ expresses how probable $\mathcal{D}$ is for different values of parameters w

- As such, it is (conditional) probability distribuition

$p(\mathcal{D}|w)$ is evaluated for $\mathcal{D}$, it can be viewed as a function of the parameter vector w

- As such, it is understood as a **likelihood function**

Bayesian probabilities (cont.)

Frequentist setting

w is a fixed parameter, whose value is determined by an *estimator*, and error bars on its estimate $w^\star$ are obtained by considering the distribution of possible data sets $\mathcal{D}$

- A widely used frequentist estimator is **maximum likelihood**
- w^* is the maximiser of the likelihood function $p(\mathcal{D}|w)$
- The w that maximises the probability of the data $\mathcal{D}$

Bayesian setting

There is only a single data set $\mathcal{D}$ (the one that is observed), and the uncertainty in the parameters is expressed through a probability distribution over w, given that data set

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$$
$$\propto \textbf{likelihood} \times \textbf{prior}$$

Integrating both sides of the Bayes' theorem with respect to w, we can express the denominator $p(\mathcal{D})$ in terms of prior distribution $p(w)$ and likelihood function $p(\mathcal{D}|w)$

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)dw$$