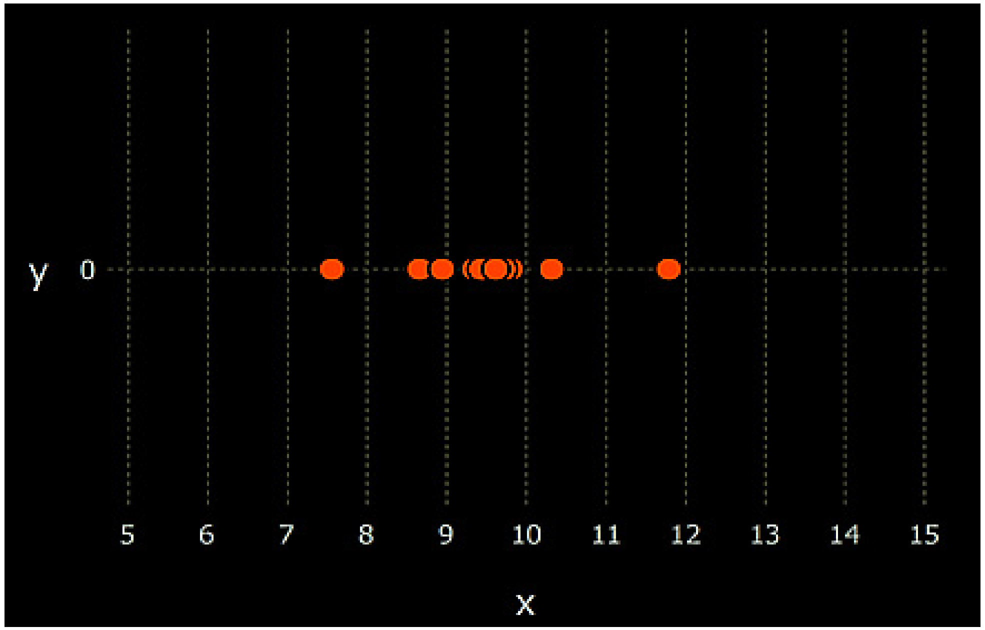
Maximum Likelihood Estimation

(**MLE** – PRML)

Maximum Likelihood Estimation (MLE)

Basic Idea:
Maximum likelihood estimation is a method that *determines values for the parameters of a model*. The parameter values are found such that *they maximise the likelihood* that the process described by the model produced the data that were actually observed.

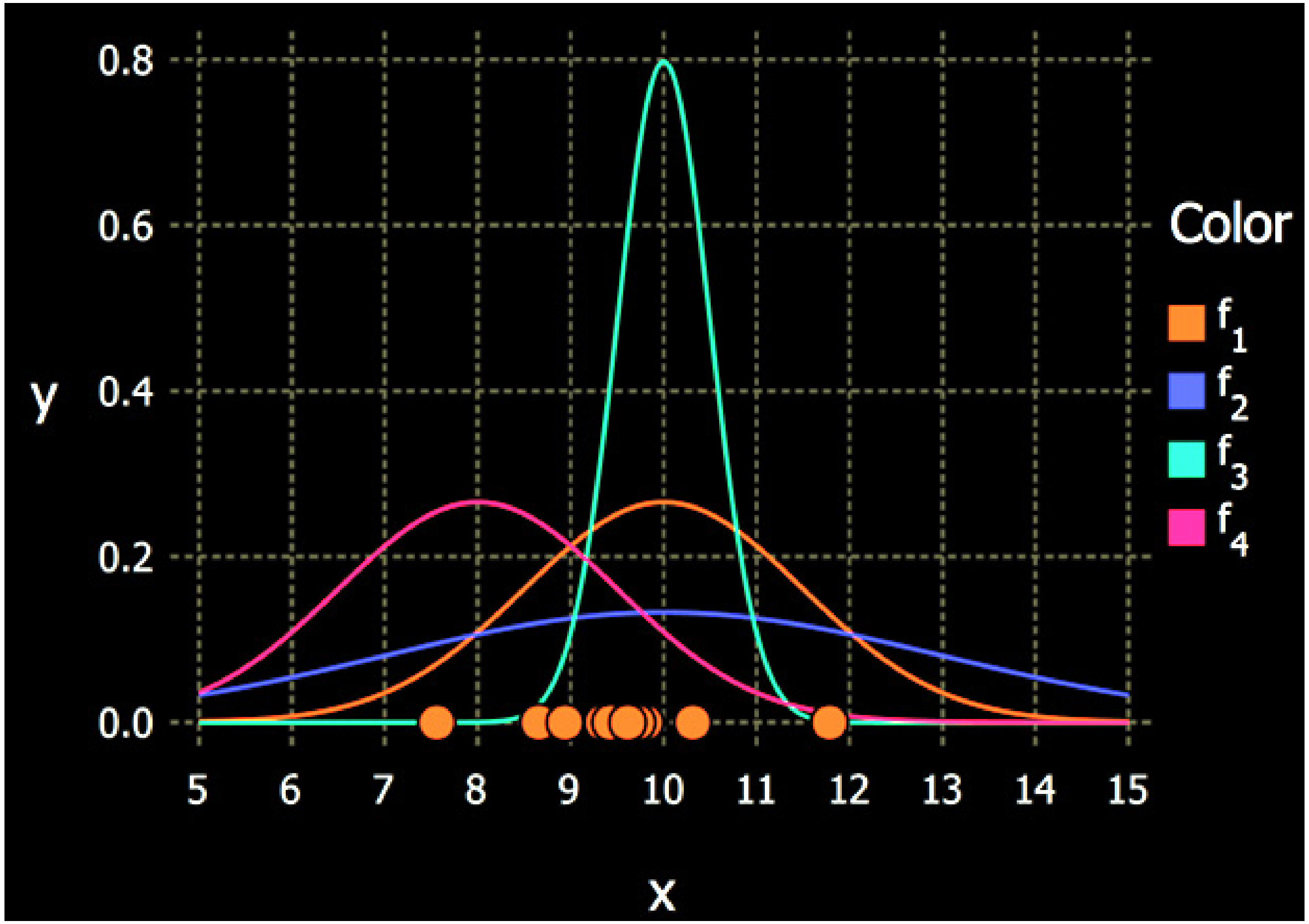
Example: Suppose there are 10 data points. For example, each data point could represent the length of time in seconds that it takes a student to answer a specific exam question. These 10 data points are shown in the figure below:



First task: which model best describes the observed data?

For given data, the data generation process can be adequately described by a **Gaussian (normal) distribution**

Gaussian distribution parameters: **mean μ and standard deviation σ.** Different values of these parameters result in different curves.



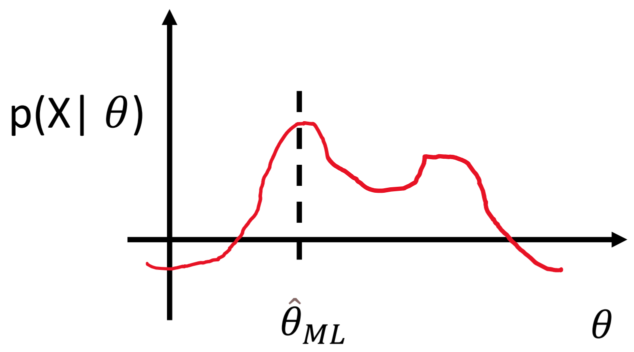
The true distribution from which the data were generated was
f1 ~ N(10, 2.25), which is the **Orange** (f1) curve in the figure above.

Intuitive definition: Maximum likelihood estimation is a method that will find the values of μ and σ that result in the curve that best fits the data.

Definition of the MLE

The ML estimate of a parameter is that value which, when substituted into the probability distribution (or density), produces that distribution for which the probability of obtaining the entire observed set of samples is maximized.

i.e., $\hat{\theta}_{ML}$ is the value of parameter θ that **maximizes** the "Likelihood Function" p(X|θ) for the **specific measured** data X



$\hat{\theta}_{ML}$ maximizes the likelihood function

Note: Because ln(z) is a monotonically increasing function…

$\hat{\theta}_{ML}$ maximizes the **log** likelihood function ln{p(X|θ)}

General Analytical Procedure to Find the MLE

1. Find log-likelihood function: ln p(X|θ)

2. Differentiate w.r.t θ and set to 0: $\partial \ln p(X|\theta)/\partial \theta = 0$

3. Solve for θ value that satisfies the equation.

Some examples of model and their unknown parameters:

S.No.	Distribution	Unknown parameters (θs)
1	Binomial distribution	**n, p**
2	Poisson distribution	$\boldsymbol{\lambda}$
3	Geometric distribution	**p**
4	Normal distribution	$\boldsymbol{\mu, \sigma^2}$

MLE for Gaussian Distribution

For the case of a single real-valued variable x, the Gaussian distribution is defined as:

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Suppose we have a data set of observations **x** = $(x_1, \cdots\cdots x_N)^T$, representing N observations of the scalar variable x.

Data set **x** is i.i.d.,

Thus, we can write the probability of the data set, given µ and σ^2, in the form:

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N}\mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

Likelihood function for the Gaussian

Note: Data points that are drawn independently from the same distribution are said to be independent and identically distributed (i.i.d.)

The log likelihood function can be written in the form:

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi).$$

Maximizing above equation with respect to **μ**, we obtain the maximum likelihood solution given by:

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Similarly, maximizing with respect to $\boldsymbol{\sigma^2}$, we obtain the maximum likelihood solution for the variance in the form

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Example : Suppose that X is a discrete random variable with the following probability mass function: where $0 \le \theta \le 1$ is a parameter. The following 10 independent observations

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ.

Solution: Since the sample is (3,0,2,1,3,2,1,0,2,1), the likelihood is

$$\begin{aligned} L(\theta) = \quad & P(X=3)P(X=0)P(X=2)P(X=1)P(X=3) \\ \times \quad & P(X=2)P(X=1)P(X=0)P(X=2)P(X=1) \end{aligned}$$

Substituting from the probability distribution given above, we have

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

Clearly, the likelihood function $L(\theta)$ is not easy to maximize.

Let us look at the log likelihood function

$$
\begin{aligned}
l(\theta) &= \log L(\theta) = \sum_{i=1}^{n} \log P(X_i|\theta) \\
&= 2\left(\log\frac{2}{3} + \log\theta\right) + 3\left(\log\frac{1}{3} + \log\theta\right) + 3\left(\log\frac{2}{3} + \log(1-\theta)\right) + 2\left(\log\frac{1}{3} + \log(1-\theta)\right) \\
&= C + 5\log\theta + 5\log(1-\theta)
\end{aligned}
$$

where C is a constant which does not depend on θ. It can be seen that the log likelihood function is easier to maximize compared to the likelihood function.

Let the derivative of $l(\theta)$ with respect to θ be zero:

$$
\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0
$$

and the solution gives us the MLE, which is $\hat{\theta} = 0.5$.

Example : Suppose $X_1, X_2, \cdots, X_n$ are i.i.d. random variables with density function $f(x|\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|x|}{\sigma}\right)$, please find the maximum likelihood estimate of σ.

Solution: The log-likelihood function is

$$l(\sigma) = \sum_{i=1}^{n}\left[-\log 2 - \log\sigma - \frac{|X_i|}{\sigma}\right]$$

Let the derivative with respect to θ be zero:

$$l'(\sigma) = \sum_{i=1}^{n}\left[-\frac{1}{\sigma} + \frac{|X_i|}{\sigma^2}\right] = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}|X_i|}{\sigma^2} = 0$$

and this gives us the MLE for σ as

$$\hat{\sigma} = \frac{\sum_{i=1}^{n}|X_i|}{n}$$

References:

1. Christopher M. Bishop, Pattern recognition and machine learning, Springer, 2006
2. https://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/MLE.pdf
3. https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1