# Data Science and Artificial Intelligence

## Machine Learning

**Bayesian learning**

**Lecture No. 1**


By- SIDDHARTH SABHARWAL SIR

# Recap of Previous Lecture

Topic MLE.

Topic

Topic

Topic

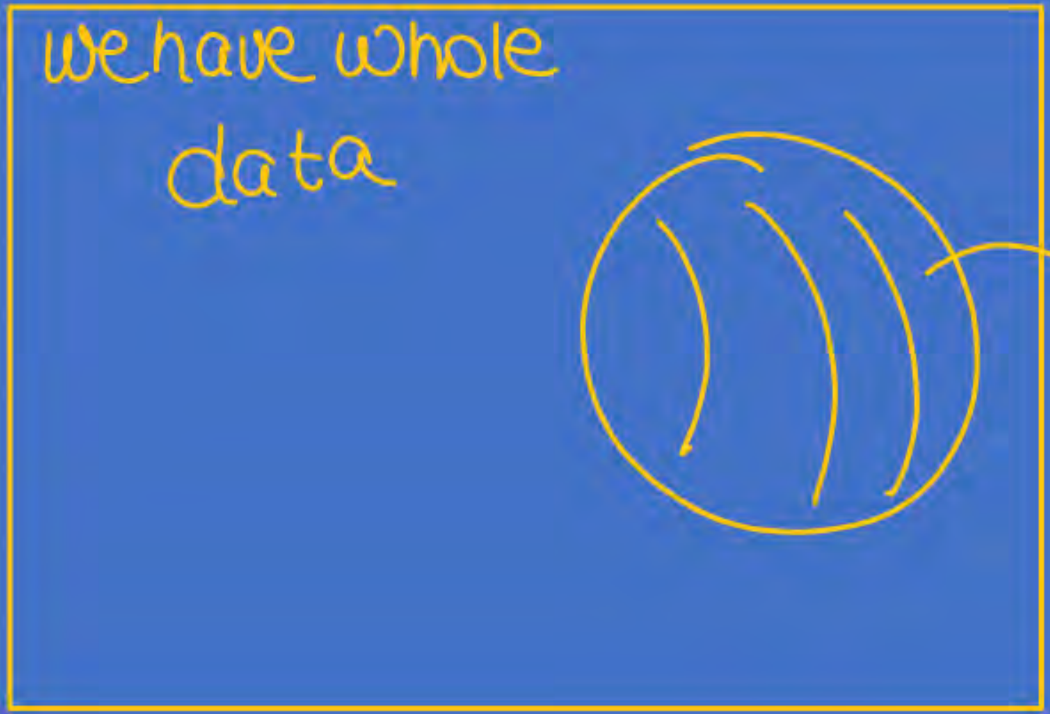
Topic

# Topics to be Covered

- **Topic** MLE
- **Topic** MAP
- **Topic** Bayesianlearning
- **Topic**
- **Topic**

STOP DOUBTING YOURSELF.
WORK HARD AND MAKE IT HAPPEN.

MLE




Using Sample we want to predict the PDF/distribution of whole data

## MLE

data $(x_1, x_2, \ldots x_n)$

$\rightarrow P(\text{data}) = \underbrace{P_{x_1} \cdot P_{x_2} \cdot P_{x_3} \ldots P_{x_n}}$ $\Rightarrow$ Gaussian

$= \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2} \ldots \ldots \right) \Rightarrow$ Likelihood

$\rightarrow \frac{d}{d\mu} \Rightarrow 0, \frac{d}{d\sigma} = 0 \Rightarrow \mu, \sigma$

REVISION

MLE

done

$$\left\{ \begin{aligned} \mu &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ \sigma &= \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 \end{aligned} \right\}$$

# Maximum likelihood Estimation

## What is MLE (lets see an example)

Sample of data $\Rightarrow$ $\left\{ \begin{array}{l} \overset{x_1}{1}, \overset{x_2}{0}, \overset{x_3}{0}, \overset{x_4}{1}, \overset{x_5}{0}, \overset{x_6}{1} \text{ - - - - - } 1 \\ K \text{ are } 1,\ N\text{-}K \text{ are } 0 \end{array} \right\}$ N values are given

$\rightarrow$ I have Sample of data

K is known.

- all points or values are Independent from each other, $1 \rightarrow P$, $0 \rightarrow 1-P$

$\rightarrow P(\text{Sample of data}) = P_{x_1} P_{x_2} P_{x_3} \text{ - - - } P_{x_N}$

$P(1-P)(1-P) \text{ - - - } \Rightarrow \left(P^K (1-P)^{N-K}\right)$

So 'P' is variable

So we have to maximize likelihood

$$L \Rightarrow \left(P^K \cdot (1-P)^{N-K}\right)$$

$$\log L \Rightarrow \log P^K + \log (1-P)^{N-K}$$

$$\Rightarrow K \log P + (N-K) \log (1-P)$$

$$\frac{d \log L}{dP} \Rightarrow \frac{K}{P} + \frac{(N-K)}{(1-P)}(-1) = 0$$

$$\left(\log f(x) \xrightarrow{d/dx} \frac{1}{f(x)} \cdot f'(x)\right)$$

$$\frac{K}{P} = \frac{N-K}{1-P}$$

$$K - \cancel{Kp} = Np - \cancel{Kp}$$

$$\boxed{\boxed{P = \frac{K}{N}}}$$

- So using sample of data we predict probab of 1 in whole data
- This is done by maximizing probab of sample of data.

# Maximum likelihood Estimation

## What is MLE (Logistic Regression)

In logistic Regression $\left(P = \dfrac{1}{1+e^{-X\beta}}\right)$

and in logistic Reg$^n$ we find $\beta$ such that

If we have N points

and

- If point has class 1 → $y=1$
  we max P for that point
- If point has class 0 → $y=0$
  we max $(1-P)$ for that point

So we Max Product of $\prod_{i=1}^{N} (P)^{y_i} (1-P)^{1-y_i}$

$y_i = 1$ So $(P)^{y_i} (1-P)^{1-y_i} = P$

$y_i = 0$ So $(P)^{y_i} (1-P)^{1-y_i} = 1-P$

$$\left\{ \max \prod_{i=1}^{N} (P)^{y_i} (1-P)^{1-y_i} \right\}$$

$\Rightarrow$ (In logistic Regression we use MLE)

# Maximum likelihood Estimation

## What is MLE (Linear Regression)

So in Regression

Sample of data $\{(x_1, y_1)\ (x_2, y_2)\ \cdots\ (x_N, y_N)\}$

So y and x must be linearly related

$$\Rightarrow \left(y = \underbrace{h(x)}_{\text{some fxn of } x} + \varepsilon_i\right)$$

gaussian noise zeromean

noise in Samples.

** In Linear Regression we use OLS ⇒ MLE

$$\rightarrow \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

So Probability of Sample of data

$\Rightarrow$ Probab of getting $y_1$ when $x_1$ is given

X " " " $y_2$ " $x_2$ " "

X " " " $y_3$ " $x_3$ " "

X " " " $y_4$ " $x_4$ " "

- $x$ axis fix ✓
- Y axis $\approx$ $h(x)$ but $\varepsilon$ is not fixed

fix

$y_1 = h(x_1) + \varepsilon_1$

$y_2 = h(x_2) + \varepsilon_2$

$y_3 = h(x_3) + \varepsilon_3$ $\rightarrow$ Randomness

Not Random

So $\Rightarrow \left( \underbrace{P(\varepsilon = \varepsilon_1)\, P(\varepsilon = \varepsilon_2)\, P(\varepsilon = \varepsilon_3)\, P(\varepsilon = \varepsilon_4)} \text{----} \right)$

Likelihood of Sample of data $\Rightarrow \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\varepsilon_1^2/2\sigma^2} \; \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\varepsilon_2^2/2\sigma^2} \text{----} \right\}$

$\Rightarrow \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^N e^{-\sum_{i=1}^{N} \varepsilon_i^2/2\sigma^2}$

- Noise has PDF
- Noise is Zero mean Gaussian
- PDF $= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\varepsilon_i^2/2\sigma^2}$

So if we run L.R on data then $\hat{y} = h(x)$

- Likelihood $\Rightarrow \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\sum_{i=1}^{N}(y_i - h(x_i))^2/2\sigma^2}$

$\Rightarrow$ Likelihood $\Rightarrow \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$

Log(likelihood) $\Rightarrow N \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ max

So $MLE \approx \left\{ \min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \right\}$.

Q. Sample of data $(x_1, x_2, x_3, x_4 \ldots x_{10})$

data distribution $\Rightarrow$ $\lambda e^{-\lambda x} u(x)$ exponential dist.

Find $\lambda$ to maximize the likelihood of data

Likelihood $\Rightarrow$ $\left(\lambda e^{-\lambda x_1} \, \lambda e^{-\lambda x_2} \ldots \lambda e^{-\lambda x_{10}}\right)$

$$\Rightarrow \lambda^{10} e^{-\lambda \sum_{i=1}^{10} x_i}$$

log likelihood $\Rightarrow$ $10 \log \lambda - \lambda \sum_{i=1}^{10} x_i$

$$\frac{d}{d\lambda}\, 10 \log \lambda - \lambda \sum_{i=1}^{10} x_i = 0$$

$$10/\lambda = \sum_{i=1}^{10} x_i, \quad \lambda = 10 \Big/ \sum_{i=1}^{10} x_i$$

# Maximum likelihood Estimation

## Probability Density Estimation & Maximum Likelihood Estimation

### So what is Probability Density Estimation

- **Probability Density:** Assume a random variable x that has a probability distribution p(x). The relationship between the outcomes of a random variable and its probability is referred to as the probability density.
- The problem is that we don't always know the full probability distribution for a random variable. This is because we only use a small subset of observations to derive the outcome. This problem is referred to as Probability Density Estimation as we use only a random sample of observations to find the general density of the whole sample space.

# Maximum likelihood Estimation

## Probability Density Estimation & Maximum Likelihood Estimation

### So what is Probability Density Estimation

- **Density Estimation:** It is the process of finding out the density of the whole population by examining a random sample of data from that population.

done

# Maximum likelihood Estimation

## Probability Density Estimation & Maximum Likelihood Estimation

### Definition

- Maximum Likelihood Estimation

- our primary job is to analyse the data that we have been presented with.
- First thing would be to identify the distribution from which we have obtained our data.
- Next, we need to use our data to find the parameters of our distribution.
- Normal distributions, as we know, have mean (μ) & variance (σ2)
- Binomial distributions have the n and p.
- Exponential distributions have the inverse mean (λ).

## What is Maximum Likelihood Estimation (MLE)

- we want to do now is obtain the parameter set $\theta$ that maximises the joint density function of the data vector; the so-called Likelihood function $L(\theta)$.
- This likelihood function can also be expressed as $P(X|\theta)$, which can be read as the conditional probability of X given the parameter set $\theta$.

$$L(\theta) = p(X \mid \theta) = p(X(1), X(2), \ldots X(n) \mid \theta)$$

X is the data matrix, and X(1) up to X(n) are each of the data points, and $\theta$ is the given parameter set for the distribution.

## What is Maximum Likelihood Estimation (MLE)

- To obtain this optimal parameter set, we take derivatives with respect to θ in the likelihood function and search for the maximum: this maximum represents the values of the parameters that make observing the available data as likely as possible.

$$\frac{\partial}{\partial \theta} p(X|\theta) = 0$$

Taking derivatives with respect to θ

# Maximum likelihood Estimation

## What is Maximum Likelihood Estimation (MLE)

- if the data points of X are independent of each other, the likelihood function can be expressed as the product of the individual probabilities of each data point given the parameter set:

$$L(\theta) = p(X \mid \theta) = \prod p(X(j) \mid \theta)$$

Taking the derivatives with respect to this equation for each parameter (mean, variance, etc...) keeping the others constant, gives us the **relationship between the value of the data points, the number of data points, and each parameter.**

## What is Maximum Likelihood Estimation (MLE)

From the likelihood function we take log likelihood function

# Maximum likelihood Estimation

## What is Maximum Likelihood Estimation (MLE)

The goal of MLE is to infer $\Theta$ in the likelihood function $p(X|\Theta)$.

$$
\begin{aligned}
&\theta_{MLE} \\
&= arg\, max\; p(X|\theta) \quad \leftarrow \text{data values} \\
&= arg\, max \prod_i p(x_i|\theta) \quad \leftarrow \text{Independent} \\
&= arg\, max\; log \prod_i p(x_i|\theta) \\
&= arg\, max \sum_i log\, p(x_i|\theta)
\end{aligned}
$$

Lets see some examples of MLE

Bayes theorem

$\Rightarrow P(A/B) = \dfrac{P(B/A)\,P(A)}{P(B)}$

$\Rightarrow$ $P(X/\theta) \Rightarrow$ likelihood $\rightarrow$ Predict $\theta$ from sample of data by maximizing $P(X|\theta)$.

$\Rightarrow$ If we have some Prior knowledge of $\theta$

then $\underbrace{P(X/\theta)}_{\text{likelihood from samples}} \cdot \underbrace{P(\theta)}_{\text{Prior knowledge}} \Rightarrow$ Aposteriori Probability

$$\text{Aposteriori Probab} = \underbrace{\text{Likelihood}}_{\text{Sample of data}} \times \underbrace{\text{Prior Knowledge}}_{\text{expert.}}$$

# Bayesian learning

# Maximum Aposterori Probability Rule

## What is Maximum Aposteriori Probability Rule(MAP)

Here we maximize ...

- MAP stands for Maximum A Posteriori probability. It is a method for estimating the parameters of a statistical model, given a dataset and some prior knowledge about the model. The goal of MAP is to find the parameter values that maximize the posterior probability of the data, given the model and the prior knowledge. This is done by choosing the values of the parameters that make the observed data most probable, given the prior knowledge.

# Maximum Aposterori Probability Rule

## What is Maximum Aposteriori Probability Rule(MAP)

Here we maximize ...

- The goal of MAP is to find the parameter values that maximize the posterior probability of the data, given the model and the prior knowledge.
- MAP is similar to MLE (Maximum Likelihood Estimation), but it incorporates prior knowledge about the model into the estimation process. This can be useful in cases where the data is limited or noisy, or where there is a need to incorporate domain-specific knowledge into the model.

# Maximum Aposterori Probability Rule

## What is Maximum Aposteriori Probability Rule(MAP)

Here we maximize ...

- MAP is similar to MLE (Maximum Likelihood Estimation), but it incorporates prior knowledge about the model into the estimation process. This can be useful in cases where the data is limited or noisy, or where there is a need to incorporate domain-specific knowledge into the model.

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \propto p(X|\theta)p(\theta)$$

$$\begin{aligned} &\theta_{MAP} \\ &= arg\ max\ p(X|\theta)p(\theta) \\ &= arg\ max\ log[p(X|\theta)] + log(p(\theta)) \\ &= arg\ max\ log \prod_i p(x_i|\theta) + log(p(\theta)) \\ &= arg\ max \sum_i log\ p(x_i|\theta) + log(p(\theta)) \end{aligned}$$

Comparing the equation of MAP with MLE, we can see that the only difference is that MAP includes prior in the formula, which means that the likelihood is weighted by the prior in MAP.

## Classification – Bayesian Perspective

Sample from data $\rightsquigarrow$ $\boxed{\text{Study data}}$ $\rightarrow$ Classifier Build

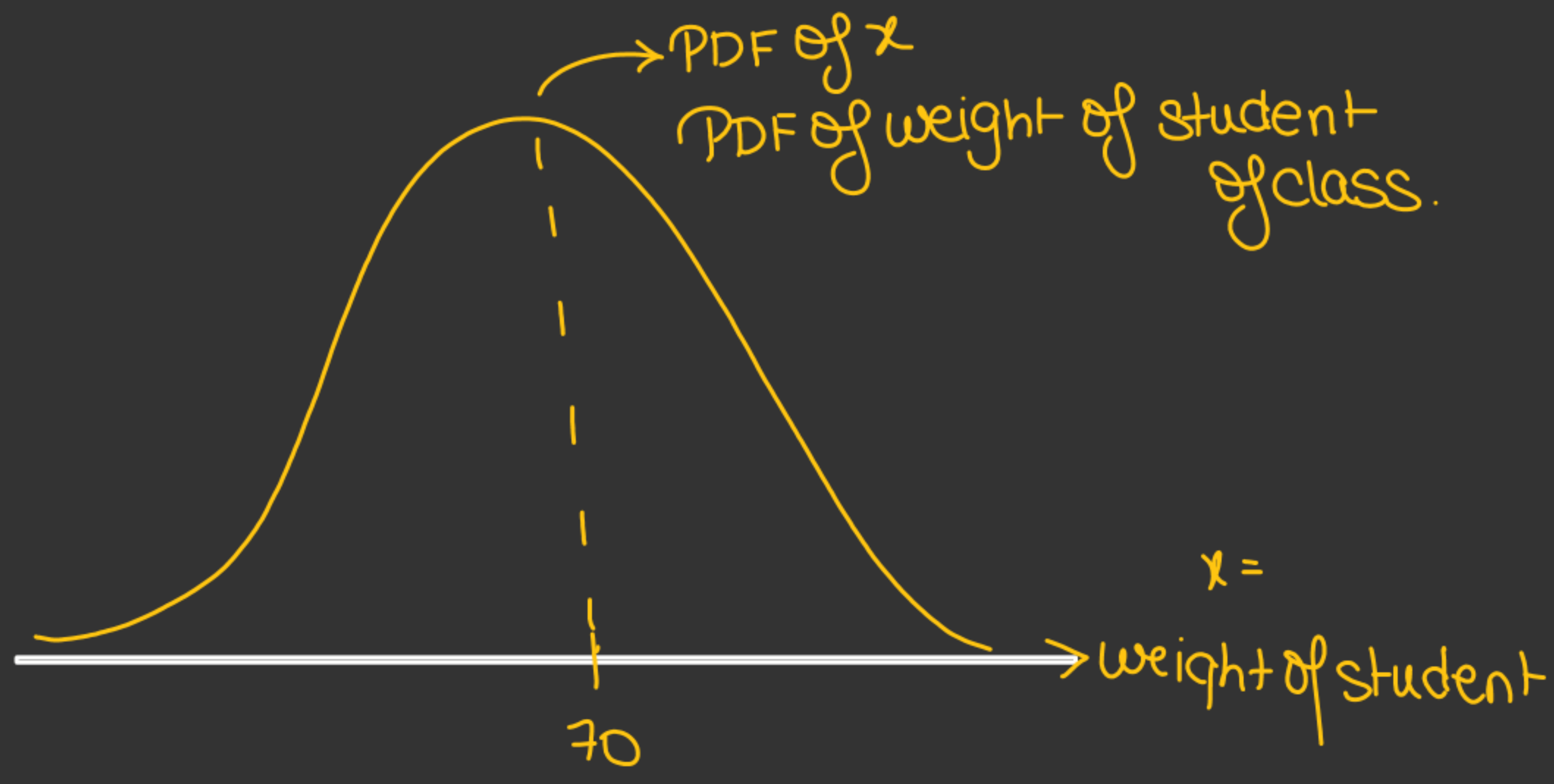
- We have to build a classifier using the data...

## Classification – Bayesian Perspective

What is bayes theorem

## Classification – Bayesian Perspective

Approach 1 : Using prior knowledge


PDF of x
PDF of weight of student of class.
x =
weight of student
70

| Fever($x$) | Covid |
|---|---|
| 105 | +ve |
| 104 | -ve |
| 103.5 | |
| 106 | |
| 102 | |

(+ve points)

| $x$ |
|---|
| 104 |
| 103 |
| 102 |
| 105 |
| 106 |
| 105.5 |

N points

So $P(x|1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2}$

gaussian

$\mu = \bar{x}$ ✓

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$ ✓

- - - - - -

-ve points

| $x$ |
|---|
| 98 |
| 97 |
| 97.5 |
| ⋮ |

$P(x|0)$

$\rightarrow \mu$

$\rightarrow \sigma^2$

## Classification – Bayesian Perspective

**Approach 2 : Using Class conditioned PDF**

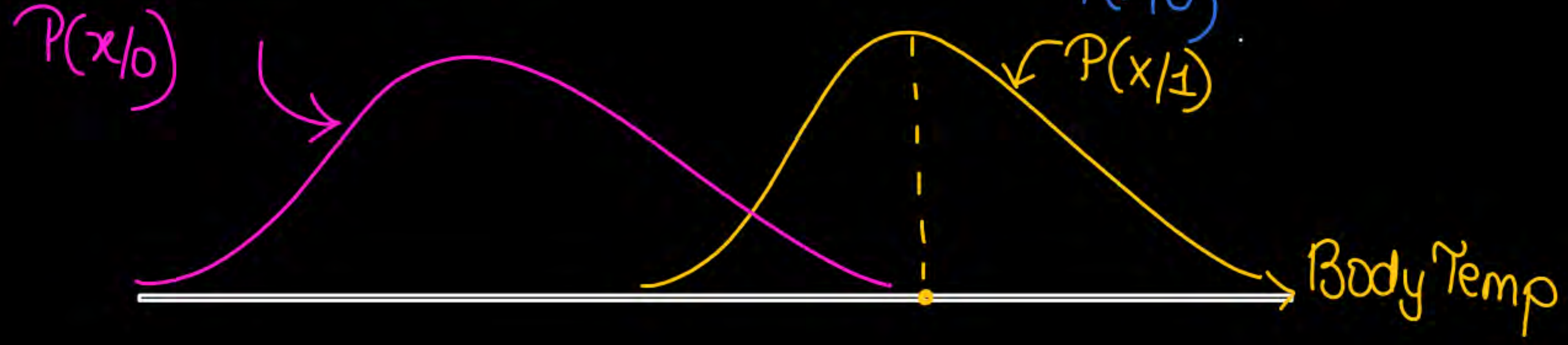
Covid Patients data

- $x \Rightarrow$ Fever
- Class 1 → Covid +ve
- Class 2 → Covid -ve

Study

→ PDF of x given Covid +ve $P(x/1)$.

→ PDF of x given Covid -ve $P(x/0)$



These are called class Conditioned PDF

These PDF are generated using MLE

So in training we generate Class Conditioned PDF

Testing ⇒


The class Conditioned which has high value is assigned to new Test Point

Steps ⇒

data
Categorical
nature

data
→ class 0 $\xrightarrow{\text{MLE}}$ $P(X|0)$ ✓ → $f(x)$

data
→ class 1 $\xrightarrow{\text{MLE}}$ $P(X|1)$ ✓ → $G(x)$

data
→ Class 2 $\xrightarrow{\text{MLE}}$ $P(X|2)$ ✓ → $w(x)$

Class Conditioned PDF

For any new point newx, we find value of $P(X|0)$, $P(X|1)$, $P(X|2)$

So whichever give max value ⇒ Class is assigned

## Classification – Bayesian Perspective

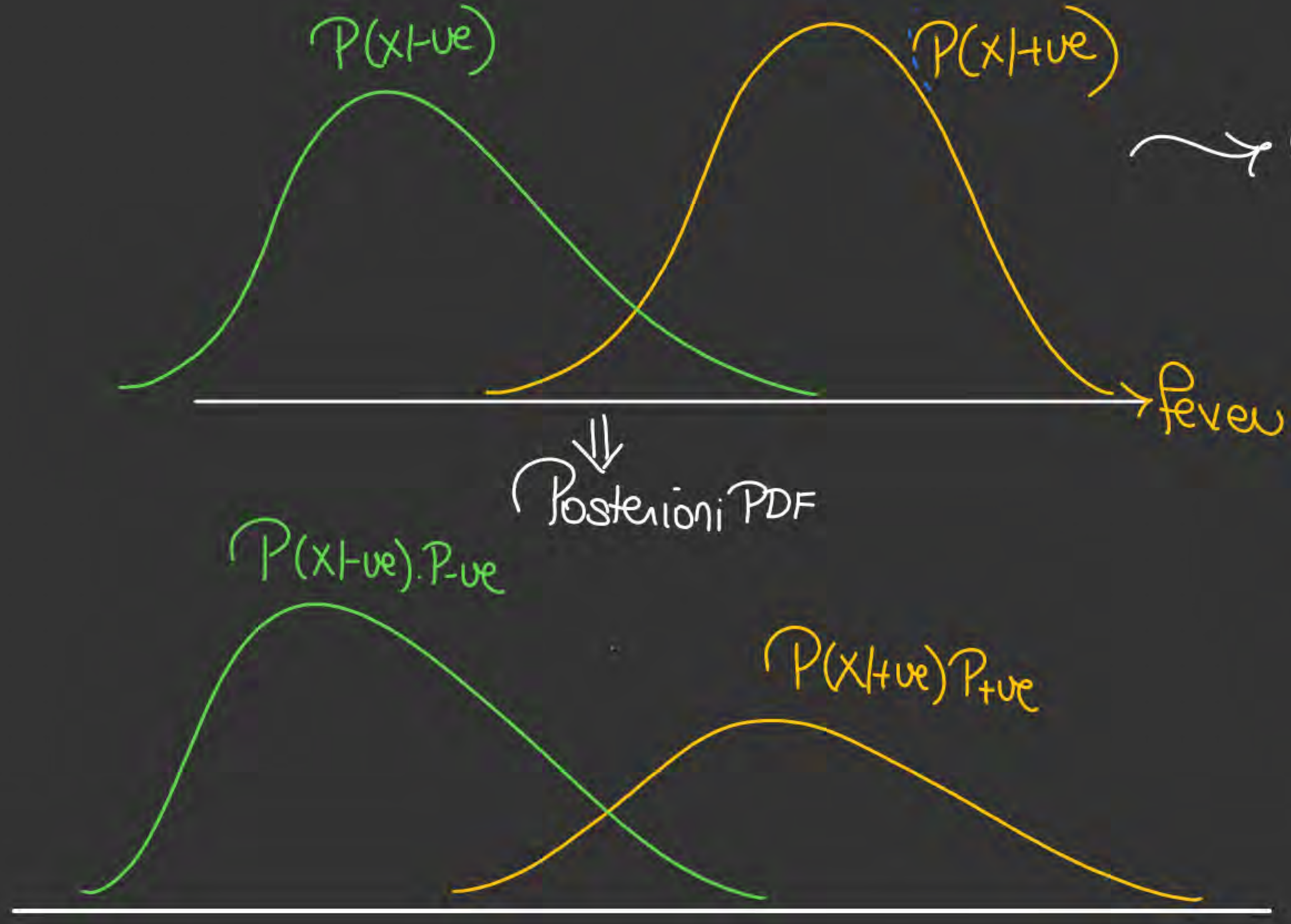
**Approach 3 : Using Posterior PDF**

Sample of data

$\rightarrow +ve \xrightarrow{MLE} P(x/+ve)$

$\rightarrow -ve \xrightarrow{MLE} P(x|-ve)$

Posteriori PDF $\Rightarrow \underbrace{P(x/+ve)}_{\text{Class Conditioned PDF}} \cdot \underbrace{P(+ve)}_{\text{Prior Know.}}$




Class Conditioned PDF → MLE

only data

doctor prior knowledge

$P_{+ve} \Rightarrow .1$

$P_{-ve} \Rightarrow .9$

So

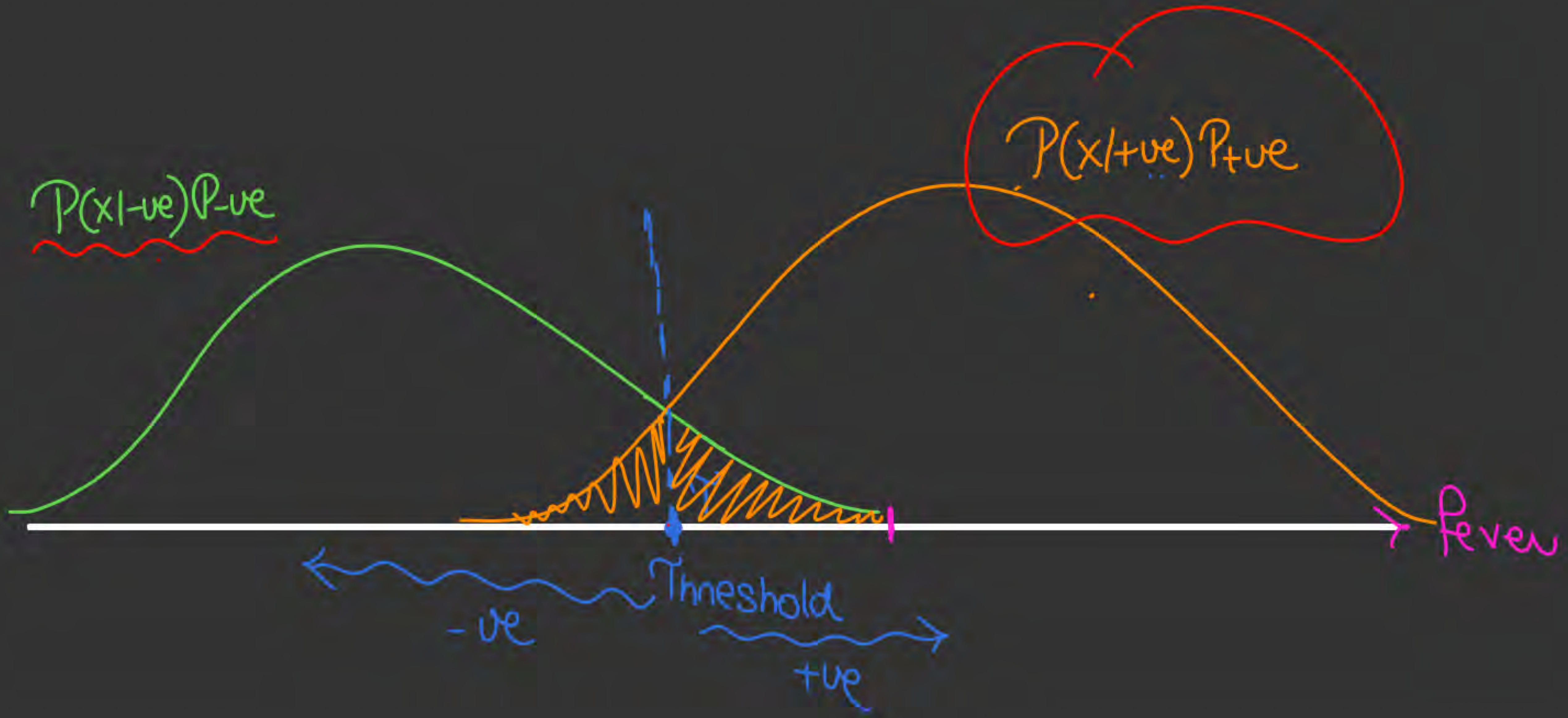
data

class 0 $\xrightarrow{\text{MLE}} P(X|0) \rightsquigarrow P(X|0)P_0 \Rightarrow f(x)$

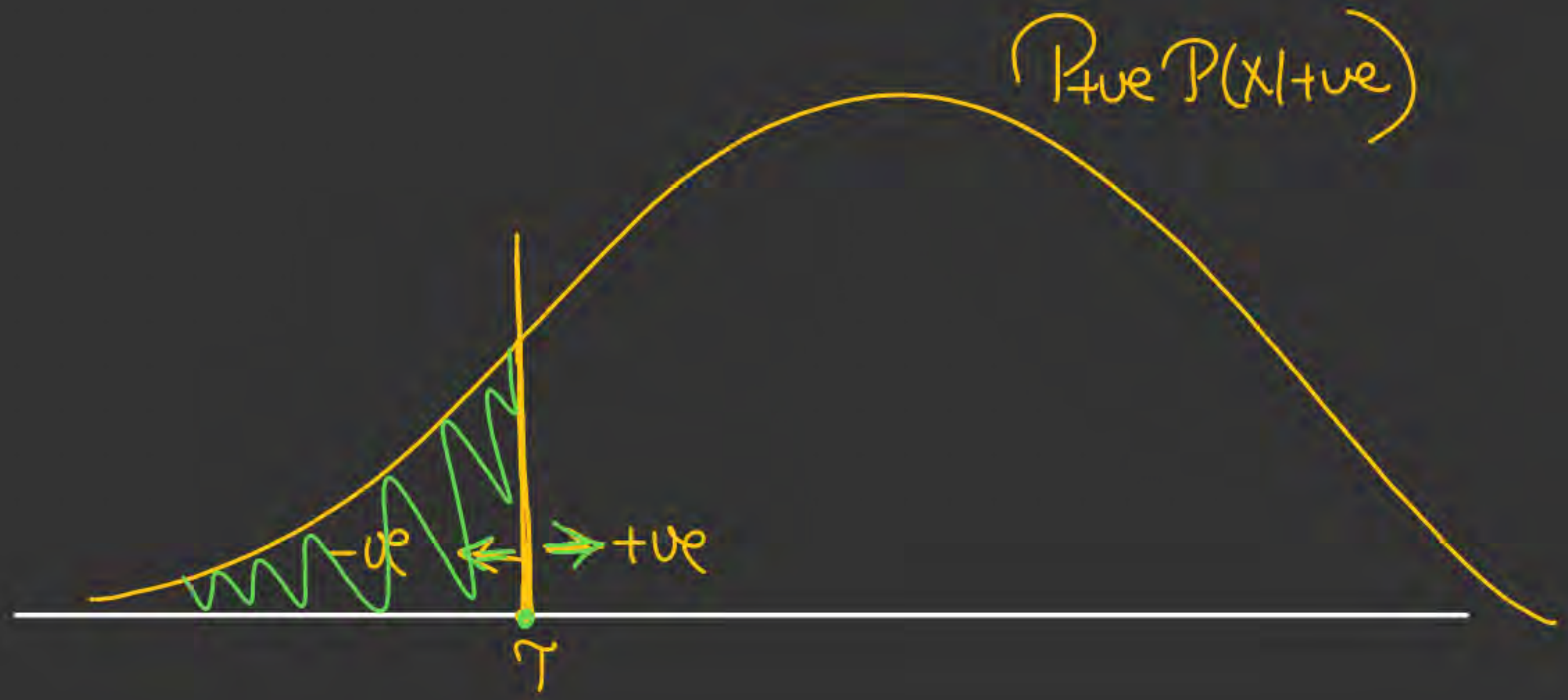
Class 1 $\xrightarrow{\text{MLE}} P(X|1) \rightsquigarrow P(X|1)P_1 \Rightarrow w(x)$

Class 2 $\xrightarrow{\text{MLE}} P(X|2) \rightsquigarrow P(X|2)P_2 \Rightarrow g(x)$

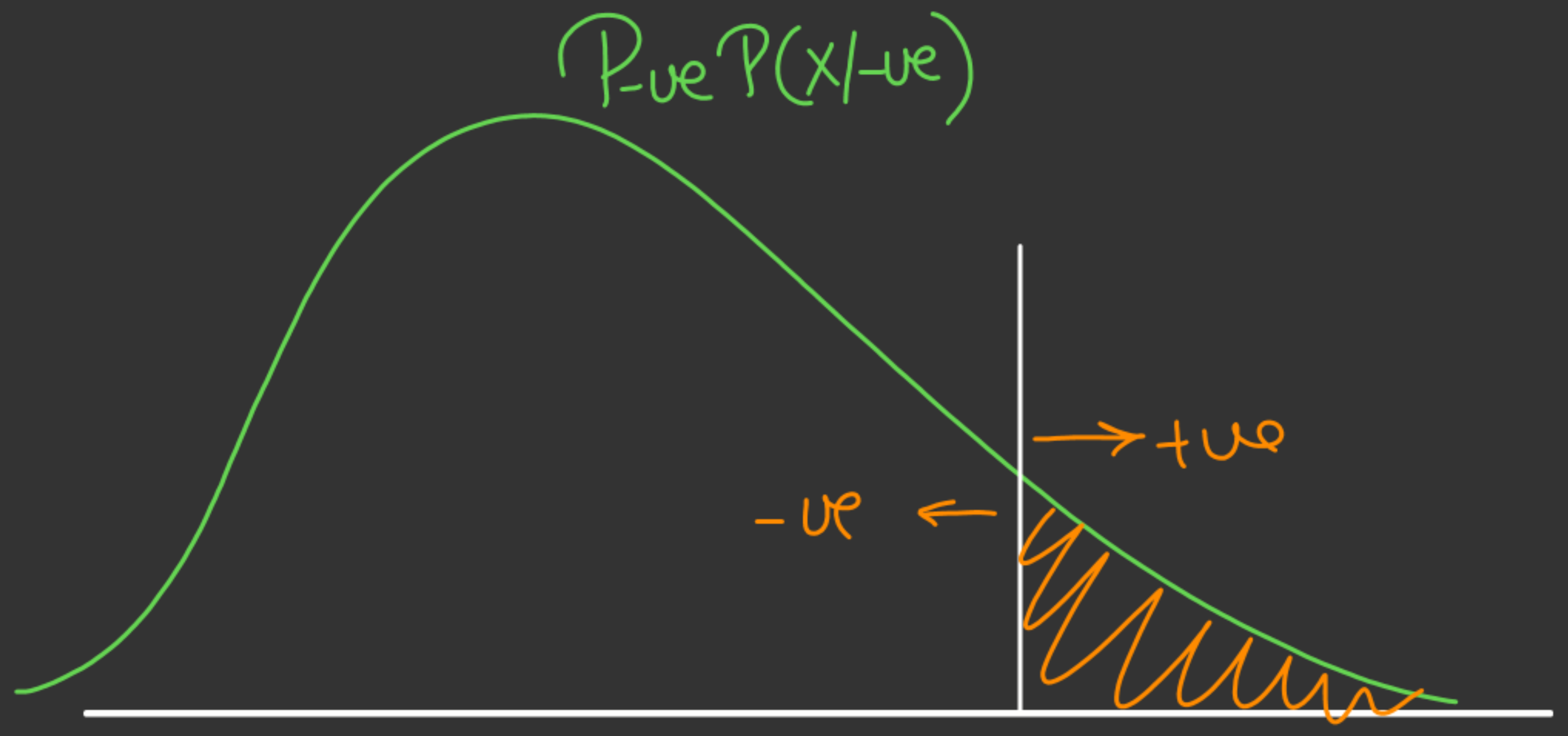
now for any new point we find $f(x)$, $g(x)$, $w(x)$ and assign class to max value.

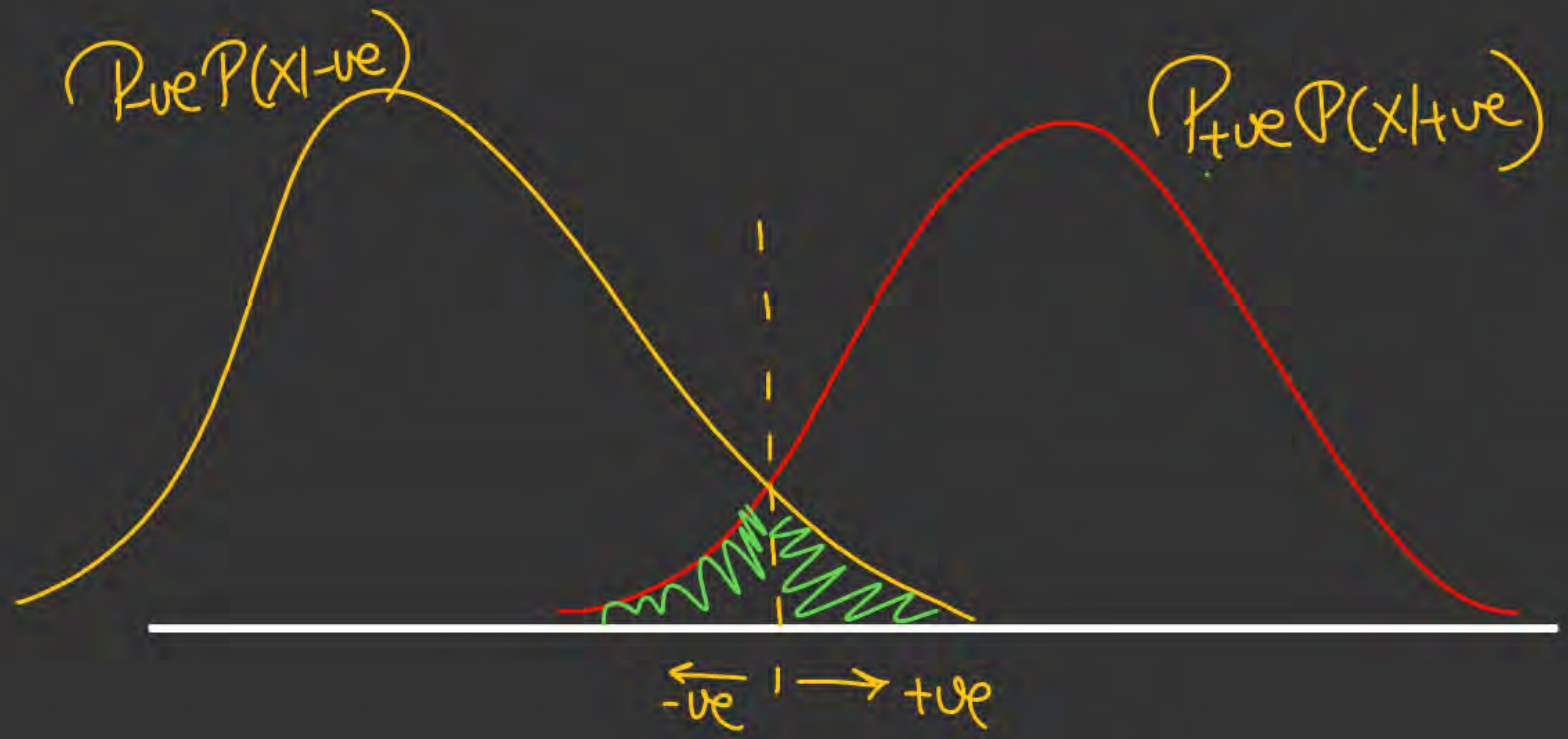


P(x|-ve)P-ve
P(x|+ve)P+ve
Fever
Threshold
-ve
+ve



P+ve P(x|+ve)
-ve
+ve
T


P-ve P(x/-ve)
→ +ve
-ve ←



$(P_{-ve} P(x|-ve))$
$(P_{+ve} P(x|+ve))$
-ve
+ve

## Classification – Bayesian Perspective

**Approach 3 : Error region**

# THANK - YOU