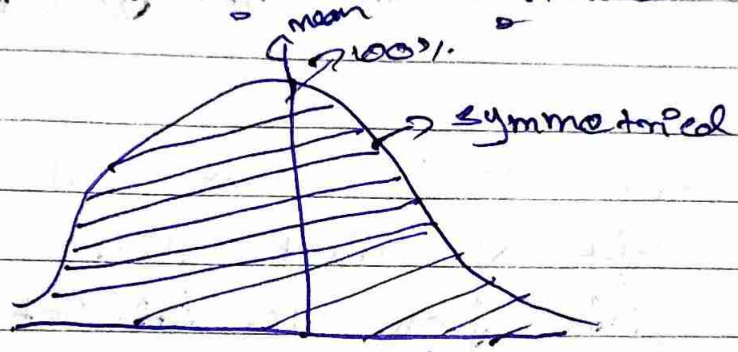
Day - 3

10/9/22

Topics

1) Normal distribution/Gaussian
2) Standard Normal Distribution
3) Z - Score

1) Gaussian or Normal Distribution



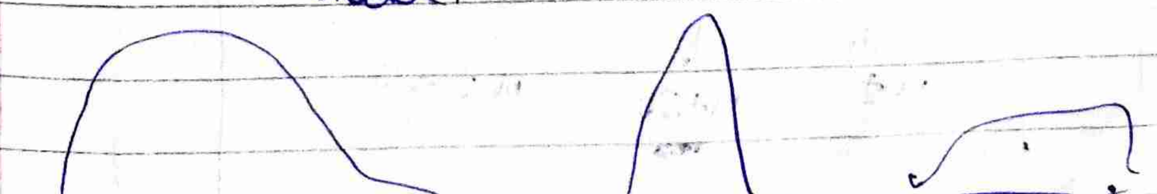
Age, weight, height we can use Normal dist for this types of data

IRIS Dataset

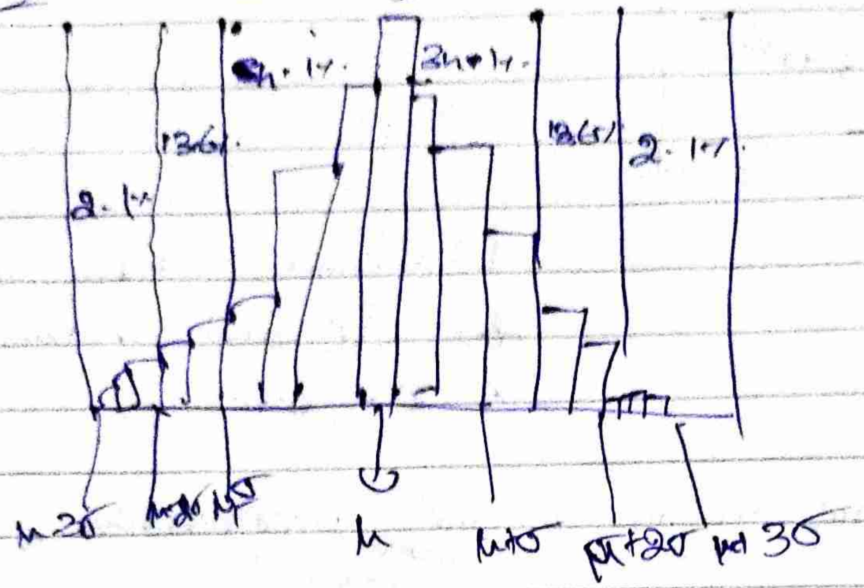
↳ petal length, sepal length, petal width, petal length

↳

These are all the dataset values can use the Normal Distribution to eg. Process the data.



Emperical Rule of Normal / Gaussian Distribution



* within the first standard deviation in left & right that's maps 68%. of data from the data set

* within the second sd from the right & left there is 95% of data from the data set.

* within the third sd to the left & right there is 99.7 data from the data set.

∴ Emperical formula is

$$\boxed{\begin{array}{ccc} 68 - & 95 - & 99.7\% \\ \downarrow & \downarrow & \downarrow \\ \mu\pm\sigma & \mu\pm2\sigma & \mu\pm3\sigma \end{array}}$$

* Remaining data comes out of 3rd distribution.

* How to find or determine it's normal distribution by using Q-Q-Plot,

Q-Q-Plot => Distribution is Gaussian or Not?

2) Standard Normal Distribution

$X \approx$ Gaussian Distribution(μ, σ) -> Random variable

1) Convert the x to y

$y \approx$ Standard $(\mu = 0, \sigma = 1)$ Normal Distribution

$x = \{1, 2, 3, 4, 5\}$ $\{\mu = 3$ and $\sigma =$

we can convert the x to y by using z score formula.

$$z = \frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\boxed{n = 1}$

-> this is called Standard error (Inferential Stat)

$$z = \frac{x_i - \mu}{\frac{\sigma}{\sqrt{1}}}$$

$$\boxed{z = \frac{x_i - \mu}{\sigma}}$$

$\mu = 3$

$\sigma = 1.414$

$\{1, 2, 3, 4, 5\}$

Let's apply

$$y = \frac{1 - 3}{1.414} = \frac{-2}{1.414} = -1.414$$

$\boxed{y = -1.414}$

$x=2$

$$y = \frac{2-3}{1.414} = \frac{-1}{1.414}$$

$$y = -0.7072$$

$x=3$

$$y = \frac{3-3}{1.414} = \frac{0}{1.414} = 0$$

$x=4$

$$y = \frac{4-3}{1.414} = \frac{1}{1.414} = 0.7072$$

$x=5$

$$y = \frac{5-3}{1.414} = \frac{2}{1.414} = +1.414$$

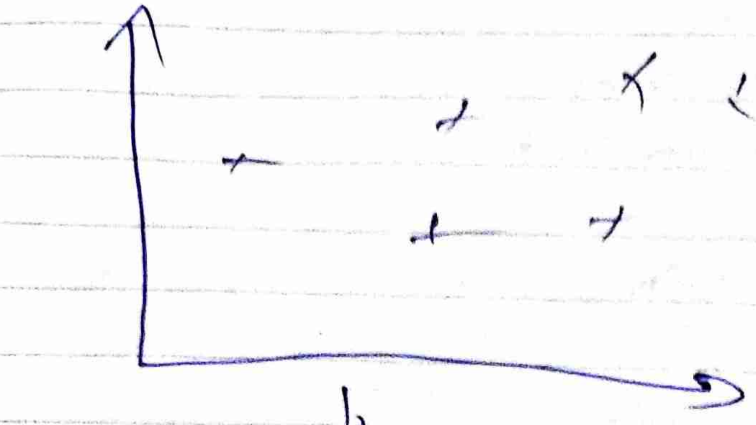
$$\therefore y = \{-1.414, -0.7072, 0, 0.7072, 1.414\}$$

∴ So we converted ND to SND
Main moto is mean should be zero in SND.

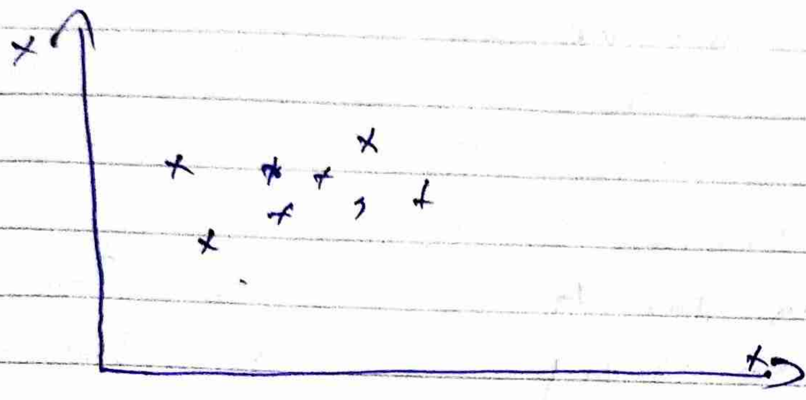
Why we changed? (Example)

(years) Age	(kg) weight	(cm) Height
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180
29	80	175
avg: 29.4285	82.28	164.285

→ by using this differe[nt] scale values in normal D[istribution] it will take time so to reduce the time calculation we have to use SND by the z score why ?.



↓



↓

for this we can get small scalling values so for this we have use z score the Purpose is to salling the small scale and after that we can revalue back the Normal scalling.

* It is called Standardization
* Here $\mu = 0$ & $\sigma = 1$ $(-3 \leftrightarrow +3)$

Normalization vs Standardization (z score) $\mu=0, \sigma=1$

* Here we have to give the range how much range will our data
  ↓ formula is
  $$\frac{x - x_{min}}{x_{max} - x_{min}}$$

* Here the range is default. It is mostly -3 to $+3$

Lower scale → Higher   Higher → lower scale

Min max scalar [0-1]   [-3 +3]

is used to
convert the lower
scale to Higher scale

Min max scalar

$$x_{scaled} = \frac{x - x_{min}}{x_{min} - x_{max}}$$

Ex:

$x = 1, 2, 3, 4, 5$

apply formula
to change it to y

So here we
normalized the
data is same
scale it is [0-1]

$$y = \frac{1-1}{5-1} = \frac{0}{4} = 0$$

$$y = \frac{2-1}{5-1} = \frac{1}{4} = 0.25$$

$$y = \frac{3-1}{5-1} = \frac{2}{4} = 0.25$$

$$y = \frac{4-1}{5-1} = \frac{3}{4} = 0.75$$

$$y = \frac{5-1}{5-1} = \frac{4}{4} = 1$$

$$y = \{0, 0.5, 0.25, 0.75, 1\}$$

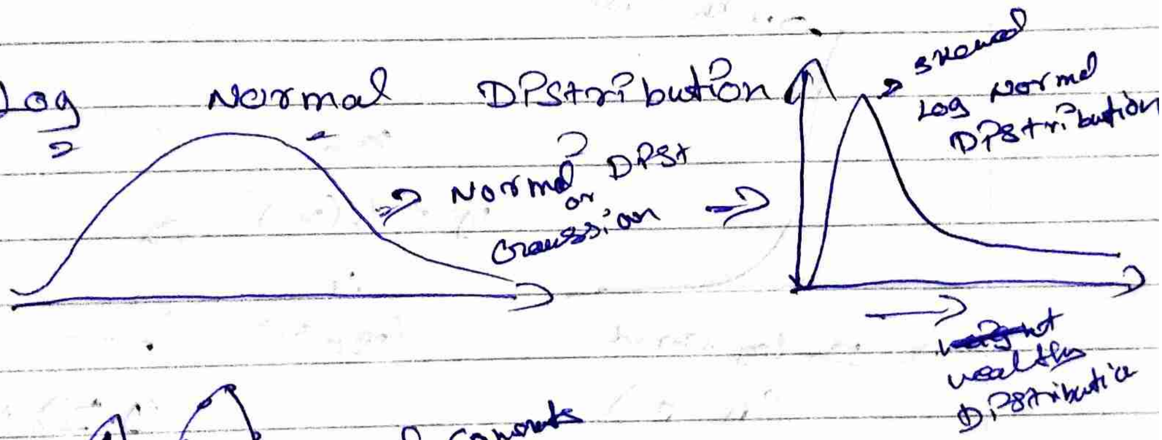
* why we are normalization?
In the image we can see by the use of pixels, so the pixels are mostly (0, 255)
* scale down the values from 0,255 to (0,1)

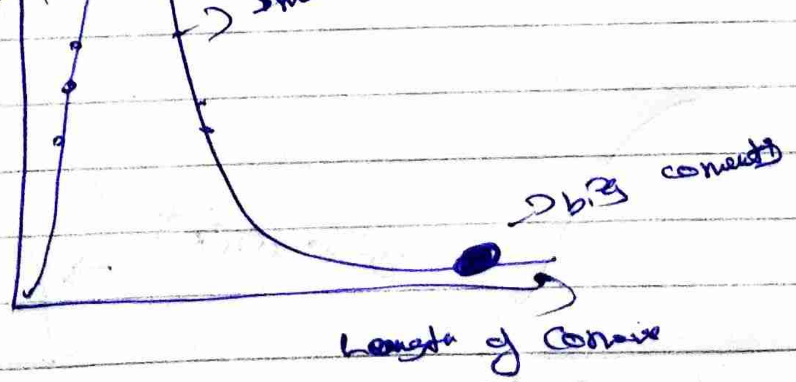
Feature scaling

1.) Normalization
2.) Standardization.

* To scale down the values in same scale.

3) Log Normal Distribution





* In website most of the peoples are comment small that will neglected in (·) small dot and big comments will more (·) for understanding the distribution

$x \rightarrow y = \ln(x)$ ∴ ln = natural log.

∴ log to the base e

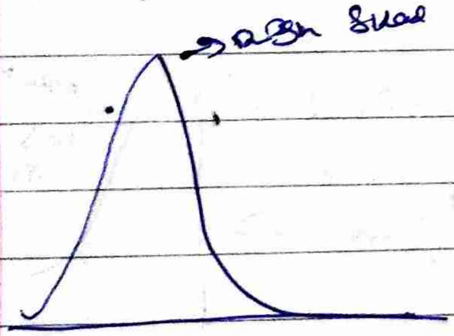
$x \Rightarrow$ log normal distribution

$y = \ln(x)$ normal distribution

* In normal distribution(x) by get the exp(x) we will get log normal Distribution.

* In log normal distribution (x) by using $y = \ln(x)$ we will get normal distribution.

Ex:



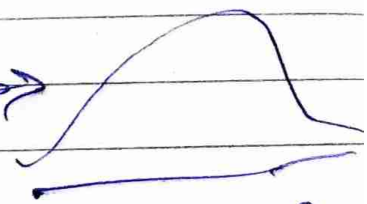
$x \Rightarrow$ log normal (μ, σ)

log x

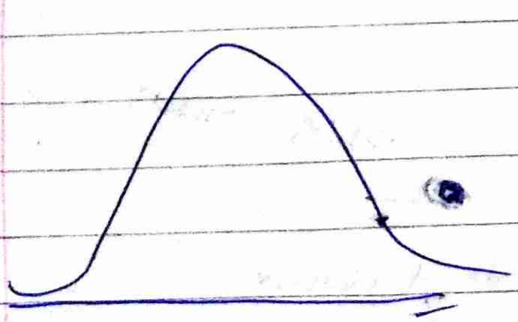
$y = \ln(x)$ $\Rightarrow$

↓

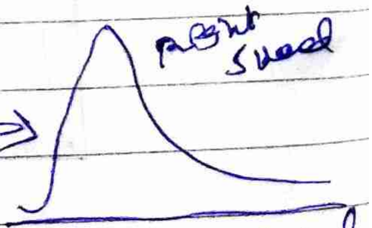
$\log_e(x)$



$y \Rightarrow$ Normal distribution



$y = \exp(x)$ $\Rightarrow$



$x \Rightarrow$ Normal dist (μ, σ)

$y =$ log normal Distribution

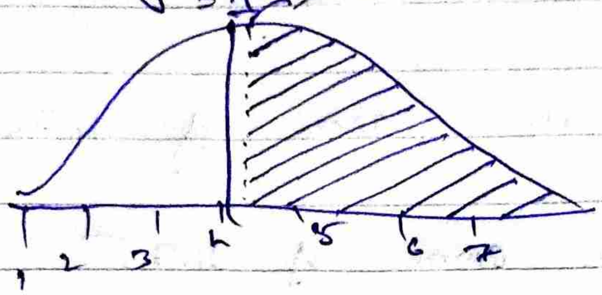
Problem

$x = \{1, 2, 3, 4, 5, 6, 7\}$

let's assume

$\mu = 4$

$\sigma = 1$ $x = 4.25$



Question:

1.) What is the percentage of score that fall's above 4.25

1.) z score $= \dfrac{4.25 - 4}{1} = \dfrac{x_i - \mu}{\sigma} = 0.25$

4.25 is

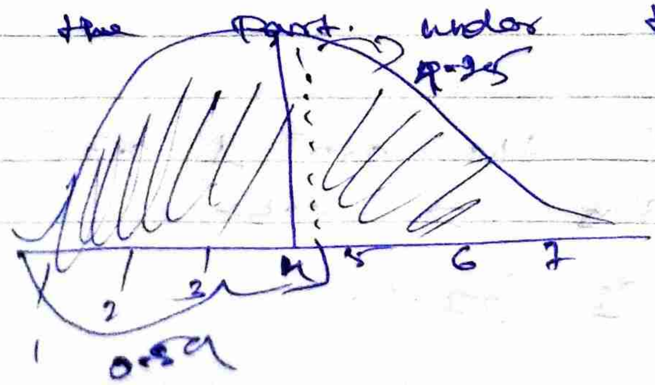
* 0.25 standard deviation away from the mean

2.) z - table.

* To find the area under the curve

* Negative z score table
* possitive z score table.

∴ If our z is 0.25 an in z Table in y axis take the value of 0.2 & in x axis take the 0.5 in z table it will give the area of the part under the curve



So: $1 - 0.59 = 0.41$

0.41 is the area of curve