STATISTICS

(session - 9)

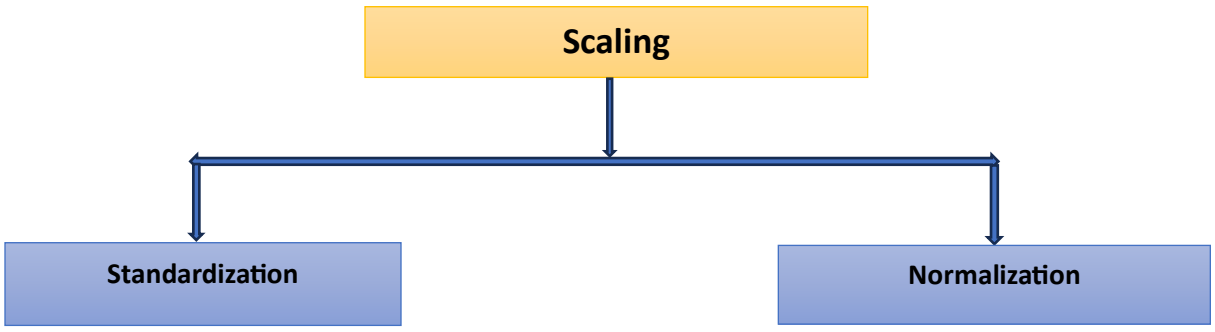
Scaling the Data :

Scaling means Changing data.

- Generally in the data we have different types of columns.
- Different columns has different values.
- Some columns has very less values. Ex : Age 1 to 100 years.
- Some columns has very high values. Ex : K , L , Crs
- Different columns has different scales.
- When we apply a ML Models comparing different scales is not recommended.
- Interpretation problem occurs.
- Some math calculations also takes more time.
- So scale all the column values under one scale , it is easy to work on.
- For Example :
    - Multiply two color images means we are multiplying 255*255
    - If we change image to gray , 255 become 1 , then 1*1
- There are two types of Scaling :

1 . Standardization

2 . Normalization



1 . Standardization :

- Standardization : z – score
- It is also called as z – standardization

Empirical rule : (68 – 95 – 99)

$$u - 1\sigma = x$$
$$-1 = \frac{x - u}{\sigma}$$
$$u + 1\sigma = x$$
$$+1 = \frac{x - u}{\sigma}$$

$$u - 2\sigma = x$$
$$-2 = \frac{x - u}{\sigma}$$
$$u + 2\sigma = x$$
$$+2 = \frac{x - u}{\sigma}$$

$$u - 3\sigma = x$$
$$-3 = \frac{x - u}{\sigma}$$
$$u + 3\sigma = x$$
$$+3 = \frac{x-u}{\sigma}$$

$$\mathbf{z = \frac{x - u}{\sigma}}$$

$z = \frac{x - u}{\sigma}$ values ranges only from -3 to 3

- Original data varies from – inf to + inf scale.
- when we converted to z – value the range become – 3 to + 3.
- Means z – standardization varies from – 3 to + 3.
- In outlier analysis

  UB = Q_3 + 3 * IQR (Huge)

  LB = Q_3 - 3 * IQR (Huge)
- Why 3 why not 3.5 because from z data varies from – 3 to 3 maximum.
- How z range is coming : Empirical rule.

Let us convert

28 , 29 , 30 , 31 , 32 to z – scale

Mean u = 30

x	(x – u)	$(x - u)^2$
28	28 – 30 = -2	4
29	29 – 30 = -1	1
30	30 – 30 = 0	0
31	31 – 30 = 1	1
32	32 – 32 = 2	4
u = 30 , σ = 1.4		10 = root(2) = 1.4

$$\sigma = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$$\sigma = \sqrt{\frac{1}{5} \times 10}$$

$\sigma = \sqrt{2}$ = 1.4

Z value of Data

$$z = \frac{x - u}{\sigma}$$

x	$\mathbf{z = \frac{x - u}{\sigma}}$
28	$\frac{28-30}{1.4} = \frac{-2}{1.4}$ = **- 1.42**
29	$\frac{29-30}{1.4} = \frac{-1}{1.4}$ = **- 0.71**
30	$\frac{30-31}{1.4}$ = 0
31	$\frac{31-30}{1.4} = \frac{1}{1.4}$ = **0.71**
32	$\frac{32-30}{1.4} = \frac{2}{1.4}$ = **1.42**

Normal distribution Formulae :

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2}} = \frac{k}{\sigma} * e^{-\frac{z^2}{2}}$$

Normalization :

- Normalization is also one scaling method.
- Z – standardization varies from – 3 to 3.
- Normalization varies from 0 to 1.
- Formulae :

$$\textbf{Normalize} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Normalization values varies from – 1 to 1 (very rare).
- Normalization is used in Image Operations.
- Color Image to Gray Image : Normalization.