Data Normalization

What is Normalization

- Normalization is a specific form of feature scaling that transforms the range of features to a standard scale.
- Normalization and, for that matter, any data scaling technique is required only when your dataset has features of varying ranges.
- Normalization encompasses diverse techniques tailored to different data distributions and model requirements.

Why Normalize Data?

- Normalized data enhances model performance and improves the accuracy of a model. It aids algorithms that rely on distance metrics, such as **k-nearest neighbors** or **support vector machines**, by preventing features with larger scales from dominating the learning process.
- Normalization fosters stability in the optimization process, promoting faster convergence during gradient-based training. It mitigates issues related to vanishing or exploding gradients, allowing models to reach optimal solutions more efficiently.

Data Normalization or Scaling

There are mainly four techniques to do Data Normalization or Scaling

1. Min-Max Normalization
2. Z-Score normalization using mean and standard deviation
3. Z-Score using mean and mean absolute deviation
4. Normalization by decimal scaling

Min- Max Normalization or Scaling

Min-Max Normalization

$$V = \frac{x - min}{max - min}$$

min = 200 and Max= 1000

$$V = \frac{200 - 200}{1000 - 200} = 0$$

$$V = \frac{300 - 200}{1000 - 200} = 0.125$$

$$V = \frac{400 - 200}{1000 - 200} = 0.25$$

$$V = \frac{600 - 200}{1000 - 200} = 0.5$$

$$V = \frac{1000 - 200}{1000 - 200} = 1$$

Data(v)	Normalized Data(v)
200	0
300	0.125
400	0.25
600	0.5
1000	1

Z-Score Normalization or Scaling

Z-Score Normalization

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

$\mu =$ Mean
$\sigma =$ Standard Deviation

$$\textit{Mean} = \frac{(200 + 300 + 400 + 600 + 1000)}{5} = 500$$

$$\textit{Standard Deviation} = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

$$= \sqrt{\frac{(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2}{5}}$$

$$= 282.8$$

Data(v)
200
300
400
600
1000

Z-Score Normalization or Scaling

Z-Score Normalization

$$z = \frac{(x - \mu)}{\sigma}$$

$$V = \frac{200 - 500}{282.8} = -1.06$$

$$V = \frac{300 - 500}{282.8} = -0.707$$

$$V = \frac{400 - 500}{282.8} = -0.354$$

$$V = \frac{600 - 500}{282.8} = 0.354$$

$$V = \frac{1000 - 500}{282.8} = 1.77$$

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

$\mu =$ Mean

$\sigma =$ Standard Deviation

***Mean* = 500**

***Standard Deviation* = 282.8**

Data(v)
200
300
400
600
1000

Normalized Data(v)
-1.06
-0.707
-0.354
0.354
1.77

Z-Score Normalization – Mean Absolute Deviation

Z-Score Normalization

$$z = \frac{x - \mu}{A}$$

$\mu =$ Mean

$A =$ Mean Absolute Deviation

$$Mean = \frac{(200 + 300 + 400 + 600 + 1000)}{5} = 500$$

$$Mean\ Absolute\ Deviation = A = \frac{|200 - 500| + |300 - 500| + \ldots + |1000 - 500|}{5} = 240$$

Data(v)
200
300
400
600
1000

Z-Score Normalization – Mean Absolute Deviation

Z-Score Normalization

$$z = \frac{(x - \mu)}{A}$$

$$V = \frac{200 - 500}{240} = -1.25$$

$$V = \frac{300 - 500}{240} = -0.833$$

$$V = \frac{400 - 500}{240} = -0.417$$

$$V = \frac{600 - 500}{240} = 0.417$$

$$V = \frac{1000 - 500}{240} = 2.08$$

$$z = \frac{x - \mu}{A}$$

$\mu =$ Mean

$A =$ Mean Absolute Deviation

$\boldsymbol{Mean}$ **= 500**

$\boldsymbol{Mean\ Absolute\ Deviation = 240}$

Data(v)
200
300
400
600
1000

Normalized Data(v)
-1.25
-0.833
-0.417
0.4117
2.08

Normalization using Decimal Scaling

Normalization using Decimal Scaling

- **Find Value of j,**
- The smallest integer *j* such that $Max\left(\frac{v_i}{10^j}\right) \leq 1$
- $\frac{200}{10^3} = 0.2$
- $\frac{300}{10^3} = 0.3$
- $\frac{400}{10^3} = 0.4$
- $\frac{600}{10^3} = 0.6$
- $\frac{1000}{10^3} = 1$

Data(v)
200
300
400
600
1000