

Feature Scaling

Tuesday 14 May 2024 5:41 AM

1) standardization

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

mean centering
scaling

$z \sim (0, 1)$

→ Covariance will be maintained

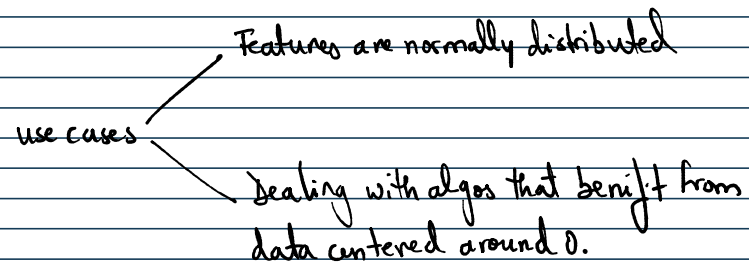
pros - simple and efficient to implement

cons - Doesn't work with algos that assume non-negative values.

StandardScaler [sklearn]

→ Distribution of is maintained [individual feature]

→ Outliers will still be outliers



2) MinMax Scaling

$$\boxed{\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}}$$

[0, 1]

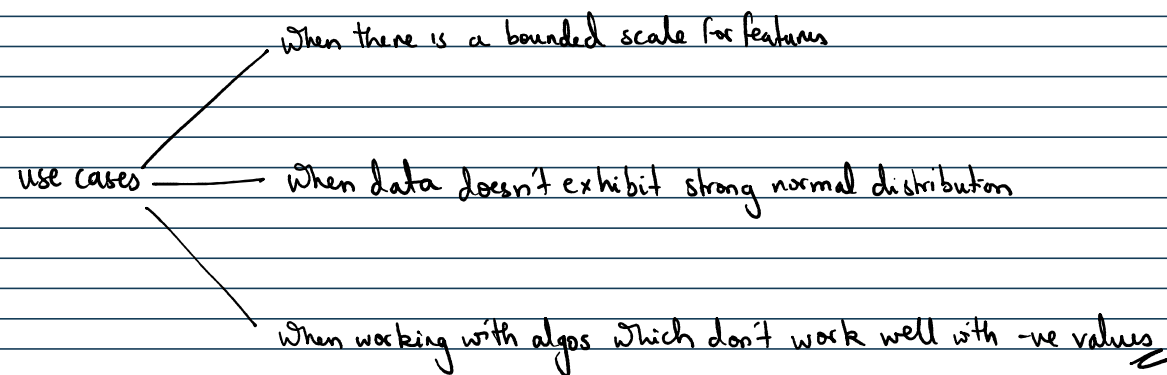
MinMaxScaler [sklearn]

→ Distribution & Relative Distances are maintained

→ sensitive to outliers

→ Zero values transformation

→ Reversibility



3) Robust Scaler

→ Reduces the impact of outlier

→ outlier present & important

$$\boxed{x'_i = \frac{x_i - \text{median}}{IQR}}$$

→ Median centering

→ Distribution is maintained

→ No Fixed Range for scaled data

→ Reversibility

4) Max Absolute Scaler

→ sparse data → lots of 0's (has meaning)

$$\boxed{x'_i = \frac{x_i}{\max(|x|)}}$$

$[-1, 1]$ original data +ve & -ve
 $[0, 1]$ original data +ve
 $[-1, 0]$ original data -ve

→ Preservation of zero

→ sensitive to outliers

→ sign is preserved

→ sensitive to outliers

→ sign is preserved

→ Relative distances maintained

→ No centering

→ Reversible

→ Distribution maintained

5) L2/L1 Normalization [multivariate scaling technique]

→ Row wise scaling

$$x'_i = \frac{x_i}{L2 \text{ Norm of row}}$$

$$x'_i = \frac{x_i}{L1 \text{ Norm of row}}$$

Use cases

- (vectors)
- When Rows are more important than cols
- Text processing → sparse features → L1
- Similarity based algo
- Image processing
- Neural Network embeddings

↑ dimension L1 (manhattan distance)

→ similarity / ranking type problems