Feature Scaling

EMPOWERING High Performance
Technology Teams

OVERVIEW

Overview

- Total Session2 (2 hours)
- Focus majorly on Feature Scaling techniques in Python

Agenda

- Concept of feature scaling
- Feature scaling techniques
- Min-max feature scaling
- Standardization
- Robust scaling

FEATURE SCALING

Why Feature Scaling?

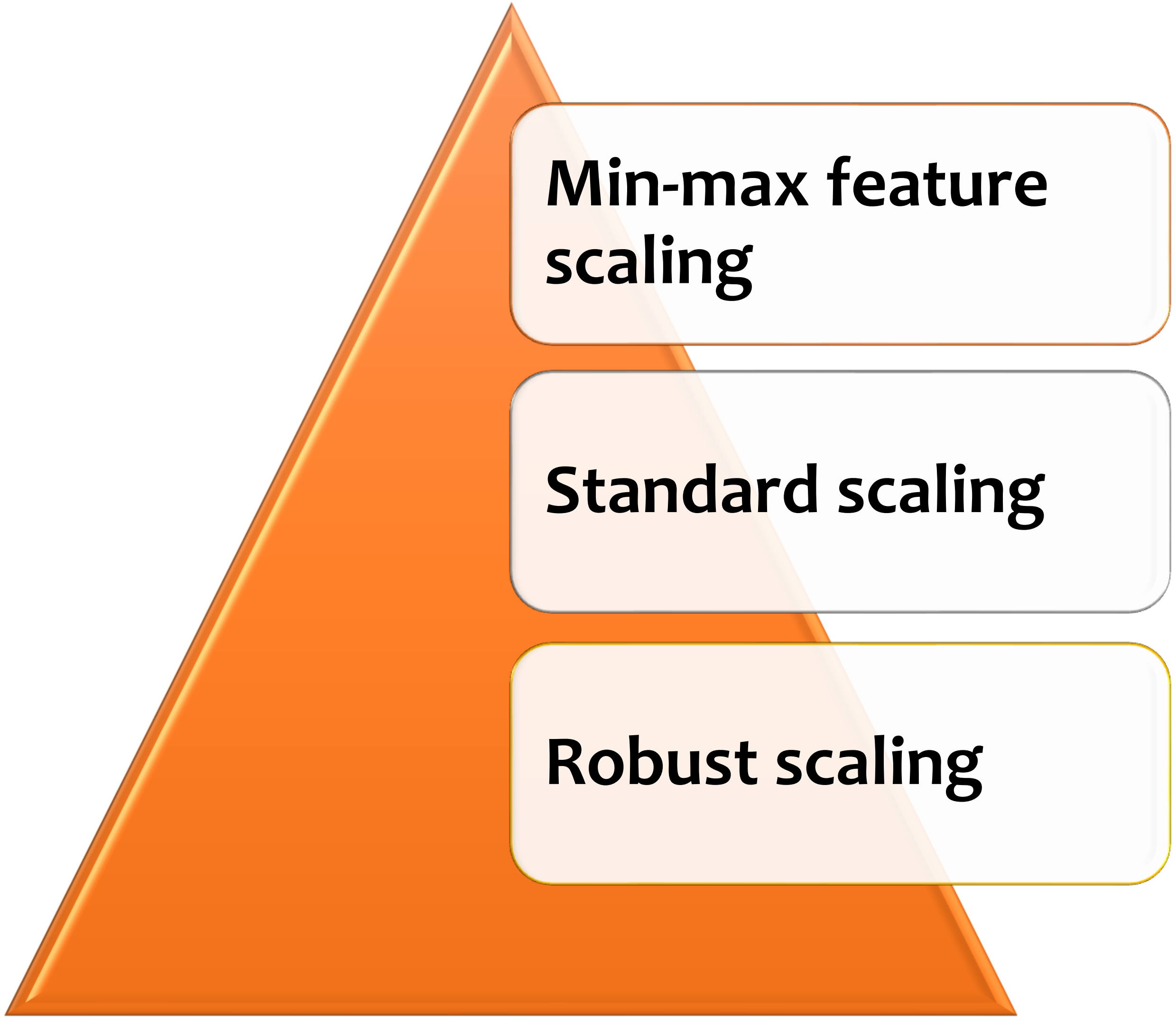
A widely used technique in machine learning to bring numeric columns to a common scale

In machine learning, certain feature values may have a much larger magnitude compared to others, which can lead to dominance of those features in the learning process

However, the magnitude of a feature does not necessarily indicate its importance in predicting the model's outcome

Feature scaling ensures that all variables are transformed to the same scale, mitigating the dominance of high-magnitude features

Feature Scaling Techniques



Min-max feature scaling

Standard scaling

Robust scaling

Min-Max Scaling

- The min-max feature scaling approach rescales the values of a numeric feature to a fixed range, typically between 0 and 1

- The scaling is done by subtracting the minimum value of the feature and then dividing it by the range

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

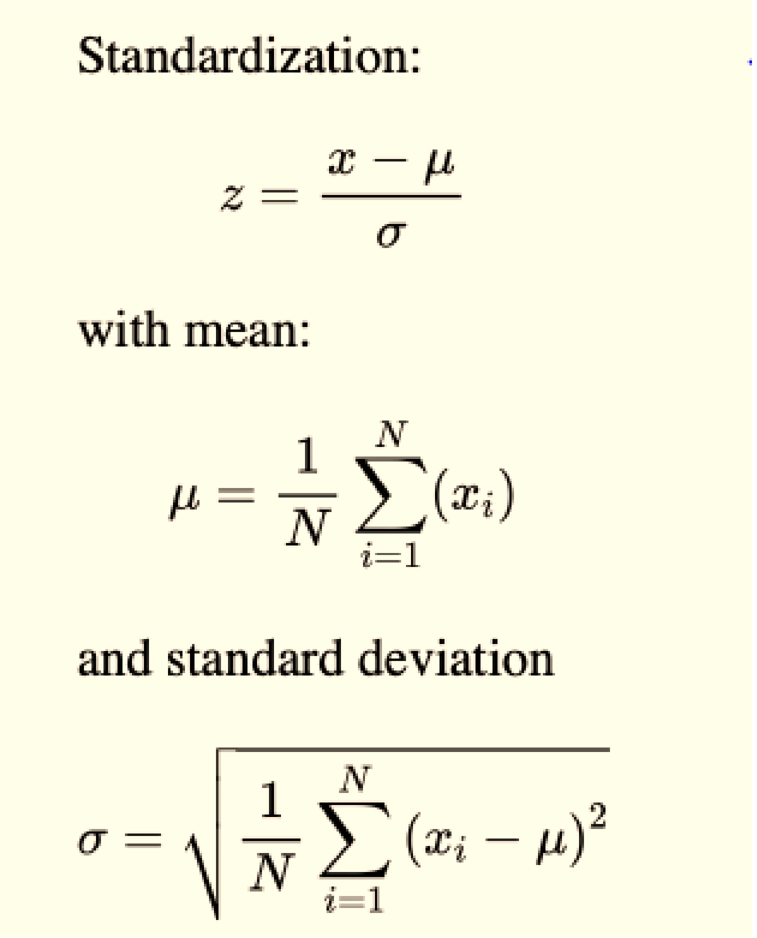
- Min-max scaling is suitable when the exact range of the feature values is known or when there are no extreme outliers

Standardization

- Standardization is the scaling of data to have zero mean and unit standard deviation

- It is preferred when data has Gaussian or normal distribution

- Typically, the z-score ranges from -3.00 to 3.00 – encompassing more than 99% of the data if the input follows a normal distribution



Standardization:

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

with mean:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i)$$

and standard deviation

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

Robust Scaling

- Robust scaling answers a simple question: **How far is each data point from the input’s median?**

- More precisely, it measures this distance in terms of the IQR

$$Scaled\ Value = \frac{Original\ Value\ -\ Input's\ Median}{Input's\ IQR}$$

Robust vs. Standard Scaling

Outliers cause the mean and standard deviation to soar to much higher values; the standard scaler uses these inflated values

When outliers are present, the standard scaler produces a distorted view of the original distribution

Robust scaler resists the pull of outliers

DEMO

Summary

- Feature scaling ensures that all variables are transformed to the same scale, mitigating the dominance of high-magnitude features
- The min-max feature scaling approach rescales the values of a numeric feature to a fixed range, typically between 0 and 1
- Standardization is the scaling of data to have zero mean and unit standard deviation
- Robust scaling uses median and interquartile range instead of mean and standard deviation, which is effective when outliers are present

THANK
YOU !