Unit – IV

Feature Scaling

Syllabus
Significance of Feature Scaling, Related terms in feature scaling, Normalization, Standardization, difference between normalization and standardization, Types of Scalers - Max Abs scaler, Robust scaler, Quantile Transformer scaler, Power Transformer scaler.

4.1. Definition: Feature scaling is a data preprocessing technique used to transform numerical features into a common scale without distorting their relationships.

It ensures that all features contribute proportionately during analysis and modeling, especially when their original ranges differ significantly.

4.2. Significance of feature scaling:

1. **Ensures Equal Contribution of Features**
   When features have different ranges, large-scale features dominate smaller ones. Scaling brings them to a comparable level.
2. **Improves Performance of Distance-Based Algorithms**
   Models like **KNN, K-means, Hierarchical Clustering** rely on distance calculations. Scaling prevents large values from skewing distance measurements.
3. **Speeds Up Convergence in Gradient-Based Algorithms**
   Algorithms like **Linear Regression, Logistic Regression, Neural Networks** use gradient descent. Scaling helps gradients move smoothly and converge faster.
4. **Enhances Visualization During EDA**
   Heatmaps, scatter plots, and cluster plots become more interpretable when all features are on similar scales.
5. **Reduces Impact of Outliers (with Robust Scaling)**
   Scaling techniques like robust scaler minimize influence of extreme values.
6. **Improves Model Accuracy and Stability**
   Many machine learning models behave more consistently and produce better predictions when input features are scaled properly.
7. **Prepares Data for Standard Assumptions**
   Some algorithms assume normally distributed or standardized data; scaling helps meet these assumptions.

4.3. Related Terms in Feature Scaling

A. Normalization
B. Standardization
C. Robust Scaling
D. Min–Max Scaling

4.4. Normalization

Normalization is a feature scaling technique that transforms data into a specific range, usually 0 to 1.
It is useful when the dataset does not follow a normal distribution or when preserving the proportion between minimum and maximum values is important.

Formula (Min–Max Normalization):

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Example: Consider data set

User	Steps
A	**2,000**
B	**5,000**
C	**7,000**
D	**10,000**

Step 1: Identify min and max

- **Min = 2,000**
- **Max = 10,000**

Step 2: Apply Min–Max Formula

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Let's normalize **7,000**:

$$7000' = \frac{7000 - 2000}{10000 - 2000} = \frac{5000}{8000} = 0.625$$

Normalized Dataset:

User	Steps	Normalized Steps (0–1)
A	2,000	0.000
B	5,000	0.375
C	7,000	0.625
D	10,000	1.000

Advantageous of Normalization:

1.Brings all features to the same scale

- Converts data to a fixed range (usually 0–1).
- Prevents large-value features from dominating small-value ones.

2. Improves performance of distance-based models

Especially useful for:

- **KNN**
- **K-means**

- **Hierarchical clustering**
  These rely on distance; normalization avoids bias.

3. Faster convergence in gradient-based algorithms

- Helps neural networks, logistic regression, etc., learn faster and more smoothly.

4. Makes visualization easier

- Heatmaps, scatter plots, and cluster plots become more interpretable.

5. Maintains the proportion of data values

- Good when relative differences (min/max relationships) matter.

Disadvantages of Normalization

1. Highly sensitive to outliers

- A single large outlier can distort the scale.
- Causes most values to shrink very close to 0.

2. Not suitable for data that needs normal distribution

- PCA, SVM, and linear regression often work better with **standardization**, not normalization.

3. Range is bounded

- If new data contains values outside the original min–max range, the model may behave unpredictably.
- Requires recalculating min/max each time.

4. Changes the original meaning of values

- Harder to interpret (e.g., “salary = 0.78” doesn’t convey real units).

Applications of Normalization

1. Distance-Based Machine Learning Algorithms

Normalization is essential when algorithms rely on distance calculations.
Examples:

- **K-Nearest Neighbors (KNN)**
- **K-means clustering**
- **Hierarchical clustering**

These algorithms treat all features equally only when they are on a similar scale.

2. Neural Networks and Deep Learning

- Normalized inputs help networks train faster.
- Prevents saturation in activation functions like sigmoid/tanh.

- Improves gradient flow and convergence.

3. Image Processing

- Pixel values are normalized to **0–1** before feeding into ML models.
- Ensures consistent brightness and contrast scaling.

4. Optimization Algorithms (Gradient Descent)

- Normalized features lead to smoother and faster convergence.
- Reduces oscillations during weight updates.

5. Time-Series and Financial Data

- Used to compare variables like prices, stock returns, or volatility.
- Allows combining different financial indicators in common models.

6. Data Visualization

- Normalization helps compare variables with different units and ranges.
- Useful in heatmaps, scatter plots, radar charts, etc.

7. Feature Engineering & Preprocessing Pipelines

- Ensures consistent scaling when combining multiple datasets.
- Helps build reliable preprocessing workflows for ML pipelines.

8. Medical Data & Sensor Data

- Vital signs, ECG signals, sensor readings often differ in range.
- Normalization allows fair input to models or diagnostic algorithms.

4.5. Standardization:

Standardization is a feature scaling technique that transforms data so that it has a mean of 0 and a standard deviation of 1.

It rescales data using the Z-score formula:

$$X' = \frac{X - \mu}{\sigma}$$

Example: Consider student test scores:

Student	Score
A	50
B	60
C	80

- Mean (μ) = 63.33
- Standard deviation (σ) ≈ 15.27

Standardize Score = 80

$$80' = \frac{80 - 63.33}{15.27} \approx 1.09$$

Standardized Data

Original Score	Standardized Score
50	-0.87
60	-0.22
80	1.09

Advantages of Standardization

1. Less Sensitive to Outliers

- Uses mean and standard deviation, so extreme values affect scaling less compared to normalization.

2. Preferred for Algorithms Assuming Normal Distribution

Useful in:

- Linear Regression
- Logistic Regression
- SVM
- PCA
- LDA

These models work better when data is standardized.

3. No Fixed Range Limitations

- Values can be negative or positive, providing more flexibility than 0–1 scaling.

4. Improves Gradient Descent Convergence

- Leads to faster and smoother training.

5. Useful for High-Dimensional Data

- Essential in PCA where variance-based feature extraction is required.

Disadvantages of Standardization

1. Harder to Interpret Scaled Values

- A value like **1.2** or **–0.8** does not directly represent the original meaning.

2. Assumes Distribution is Gaussian

- If data is heavily skewed, standardization may not perform well.

3. Does Not Bound Values

- Unlike normalization, standardized values can go far beyond –3 to +3.
- Some models expecting bounded input may not work well.

4. Requires Mean and Std of Training Data

- When new data arrives, the same μ and σ must be used.
- Wrong scaling can degrade model performance.

Applications of Standardization

1. Regression Models (Linear & Logistic)

- Ensures all variables contribute equally.
- Prevents coefficients from being biased toward larger-scaled features.

2. Support Vector Machines (SVM)

- SVM is sensitive to feature magnitude; standardization is crucial.

3. Principal Component Analysis (PCA)

- PCA maximizes variance.
- Without standardization, high-magnitude features dominate variance.

4. Clustering Algorithms

(When data is assumed to be normally distributed)

- Gaussian Mixture Models (GMM)
- Some hierarchical clustering variants

5. Neural Networks

- Standardized data helps faster training and better stability.

6. Statistical Analysis

- Z-scores are widely used in hypothesis testing, anomaly detection, and standard score comparison.

7. High-Dimensional Scientific and Medical Data

Used in:

- ECG/EEG signal analysis
- Genomics
- Experimental measurements

Difference Between Normalization and Standardization

Basis	Normalization	Standardization
Definition	Scales data to a fixed range, usually 0 to 1.	Converts data to have mean = 0 and standard deviation = 1.
Formula	$(X - X_{\min})/(X_{\max} - X_{\min})$	$(X - \mu)/\sigma$
Range	Bounded (0–1 or −1–1).	Unbounded (values can be negative or positive).
Effect of Outliers	Very sensitive to outliers.	Less sensitive compared to normalization.

When to Use	Distance-based algorithms like KNN, K-means, Neural Networks.	Algorithms assuming normal distribution: Regression, SVM, PCA.
Preserves	Proportion between minimum and maximum values.	Z-score relative distance from mean.
Interpretation	Values represent a relative position in a fixed scale.	Values represent how many standard deviations a point is from the mean.