Data Preprocessing and Scaling

⚠ **Feature scaling is critical: Distance-based algorithms are sensitive to scale**

Scaling Methods

StandardScaler

$$z = (x - \mu) / \sigma$$

- ✓ Zero mean
- ✓ Unit variance
- ✓ Assumes Gaussian

MinMaxScaler

$$x' = (x - min) / (max - min)$$

- ✓ Scale to [0, 1]
- ✓ Preserves zero
- ✓ Bounded range

RobustScaler

$$x' = (x - median) / IQR$$

- ✓ Uses median
- ✓ Uses IQR
- ✓ Robust to outliers

Preprocessing Pipeline

1. **Missing Values**
   Imputation before clustering
2. **Outlier Treatment**
   Remove or cap extremes
3. **Feature Selection**
   Remove irrelevant features
4. **Dimensionality**
   Handle curse of dimensions

Calculation Examples

StandardScaler Example

Data: [10, 20, 30, 40, 50]

μ = 30
σ = 14.14
For x = 10:
z = (10 - 30) / 14.14
z = -1.41

Result: [-1.41, -0.71, 0, 0.71, 1.41]

MinMaxScaler Example

Data: [10, 20, 30, 40, 50]

min = 10, max = 50
range = 50 - 10 = 40
For x = 20:
x' = (20 - 10) / 40
x' = 0.25

Result: [0, 0.25, 0.5, 0.75, 1.0]

RobustScaler Example

Data: [10, 20, 30, 40, 100]

median = 30
Q1 = 20, Q3 = 40
IQR = 40 - 20 = 20
For x = 100:
x' = (100 - 30) / 20 = 3.5

Result: [-1.0, -0.5, 0, 0.5, 3.5]