

# Robust Scaling

## ¿Qué es el Escalamiento Robusto?

El escalamiento robusto (Robust Scaling) es una técnica utilizada en el preprocesamiento de datos para normalizar o estandarizar características (features) de conjuntos de datos. A diferencia de otras técnicas de escalamiento que pueden ser sensibles a valores atípicos (outliers), el escalamiento robusto es diseñado para ser menos afectado por estos valores extremos.

Utiliza estadísticas robustas, tales como la mediana y el rango intercuartílico (IQR), para transformar los datos. En lugar de centrar los datos alrededor de la media y escalar según la desviación estándar, el escalamiento robusto centra los datos alrededor de la mediana y los escala según el **IQR**, que es la diferencia entre el **percentil 75** y el **percentil 25** de los datos.

## Fórmula del Escalamiento Robusto

La fórmula para escalar una característica  $X$  utilizando el escalamiento robusto es:

$$X_{\text{escalado}} = \frac{X - \text{Mediana}(X)}{\text{IQR}(X)}$$

donde:

- $\text{Mediana}(X)$  es la mediana de la característica  $X$ .
- $\text{IQR}(X)$  es el rango intercuartílico de  $X$ , definido como  $\text{Percentil}(75) - \text{Percentil}(25)$

## Beneficios

1. **Robustez a Valores Atípicos:** Al utilizar la mediana y el IQR, el escalamiento robusto es menos sensible a valores atípicos que pueden distorsionar los datos.
2. **Mejora en el Rendimiento de Algoritmos de Aprendizaje Automático:** Algoritmos como la regresión logística, las máquinas de vectores de soporte (SVM) y los métodos basados en distancia (como KNN) pueden beneficiarse de un mejor comportamiento cuando los datos están escalados adecuadamente.

3. **Adecuación para Datos con Distribuciones Asimétricas:** El escalamiento robusto puede ser más efectivo que otros métodos cuando se trabaja con datos que no siguen una distribución normal.

## Limitaciones

1. **Interpretación:** Los datos escalados pueden ser más difíciles de interpretar, especialmente cuando se comparan con métodos más tradicionales como la estandarización.
2. **Computacionalmente Intensivo:** Aunque no es significativamente más complejo, calcular percentiles puede ser más costoso computacionalmente que calcular la media y la desviación estándar.

## Otros Métodos de Escalamiento

### Min-Max Scaling

- **Fórmula:**  $X_{\text{escalado}} = \frac{X - \min(X)}{\max(X) - \min(X)}$
- **Ventajas:** Intuitivo y fácil de interpretar.
- **Desventajas:** Muy sensible a valores atípicos

### Standard Scaling (Z-Score Scaling)

- **Fórmula:**  $X_{\text{escalado}} = \frac{X - \text{Media}(X)}{\text{Desviación Estándar}(X)}$
- **Ventajas:** Estandariza los datos en función de la media y la desviación estándar.
- **Desventajas:** Sensible a valores atípicos.

### Escalamiento Robusto

- **Fórmula:**  $X_{\text{escalado}} = \frac{X - \text{Mediana}(X)}{IQR(X)}$
- **Ventajas:** Menos sensible a valores atípicos.
- **Desventajas:** Puede ser más difícil de interpretar.

## Implementación en Python

La biblioteca **scikit-learn** en Python proporciona una implementación sencilla del escalamiento robusto a través de la clase **RobustScaler**.

```
1  from sklearn.preprocessing
2  import RobustScaler
3
4  # Ejemplo de datos
5  data = [[1, 2, 3], [4, 5, 6], [7, 8, 100]]
6
7  # Crear un objeto RobustScaler
8  scaler = RobustScaler()
9
10 # Ajustar el escalador a los datos y transformarlos
11 scaled_data = scaler.fit_transform(data)
12
13 print(scaled_data)
14
```

## Aplicaciones

1. **Análisis de Datos Financieros:** En finanzas, los datos suelen tener valores atípicos debido a eventos extremos del mercado. El escalamiento robusto puede ayudar a manejar estos outliers.
2. **Procesamiento de Señales:** En el procesamiento de señales, especialmente en bioseñales como ECG o EEG, es común encontrar valores atípicos.
3. **Visión por Computador:** En imágenes, las intensidades de píxeles pueden variar ampliamente, y el escalamiento robusto puede ayudar a normalizar estas variaciones.

## Conclusiones

El escalamiento robusto es una buena técnica para normalizar datos, especialmente en presencia de valores atípicos. Al centrarse en la mediana y el IQR, proporciona una manera de reducir la influencia de valores extremos, haciendo que los algoritmos de aprendizaje automático sean más robustos y eficientes en diversas aplicaciones.