EDA-Data Standardization

Data standardization is a technique used in data preprocessing to scale the numerical features of a dataset to have similar ranges, usually between 0 and 1. This process helps to:

1. **Improve model performance**: By scaling features, you can prevent some features from dominating the model's decision-making process.
2. **Enhance comparability**: Standardized features are easier to compare across different datasets or models.
3. **Reduce noise**: Standardization can help reduce the impact of noisy or outlier values.

Example

Suppose we have a dataset with two features: `Age` and `Salary`. The raw data looks like this:

Age	Salary
25	$40000
30	$60000
35	$80000
...	...

Let's say we want to standardize these features so that they have a range of [0, 1].

Step 1: Calculate the mean and standard deviation

For `Age`, we calculate the mean (μ) and standard deviation (σ):

μ = (25 + 30 + 35 + ...) / n ≈ 45 σ ≈ 5

Similarly, for `Salary`, we get:

μ ≈ $60000 σ ≈ $12000

Step 2: Scale the data

We use the following formula to scale each feature:

x' = (x - μ) / σ

Applying this formula to our example data:

For `Age`: 25 - 45 ≈ -20 (Age - mean) (-20) / 5 ≈ -4 (standardized age)

30 - 45 ≈ -15 (-15) / 5 ≈ -3

35 - 45 ≈ -10 (-10) / 5 ≈ -2

...and so on.

For `Salary`: $40000 - $60000 ≈ -$20000 (Salary - mean) ($-20000) / $12000 ≈ -1.67 (standardized salary)

$60000 - $60000 ≈ $0 ($0) / $12000 ≈ 0

...

Result

After standardizing the features, our data looks like this:

Standardized Age	Standardized Salary
-4	-1.67
-3	-0.33
-2	0
...	...

In this example, we've scaled the `Age` and `Salary` features to have similar ranges, making it easier to compare and analyze them together.

Code

Here's a Python code snippet using scikit-learn library to perform data standardization:

```
 from sklearn.preprocessing import StandardScaler

# assume 'data' is your DataFrame with Age and Salary columns
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data[['Age', 'Salary']])
```

Note that `StandardScaler` is the default standardization method in scikit-learn.

Curated by Brajesh Kumar