Data Preprocessing-Data Scaling

Data Scaling (Normalization)

Data scaling, also known as normalization, is a technique used in data preprocessing to scale the values of numerical features so that they fall within a common range. This helps to prevent features with large ranges from dominating the model and improves the stability of algorithms.

Why Scale Data?

1. **Prevents feature dominance**: Scales are sensitive to differences in magnitude, which can lead to some features having more influence on the output than others.
2. **Improves algorithm performance**: Many algorithms require scaled data, such as Support Vector Machines (SVM) and Gradient Boosting Machines (GBM).
3. **Enhances interpretability**: Scaling helps to remove units of measurement from the data.

Types of Scaling

1. **Min-Max Scaling**: Maps values to a common range [a, b].
2. **Standardization**: Centers data around zero with unit variance.
3. **Log Scaling**: Transforms data using logarithm (e.g., log(x)).

Example: Min-Max Scaling

Suppose we have a dataset of exam scores:

```
 import pandas as pd

# Create a sample dataframe
data = {
    'Exam1': [90, 80, 70],
    'Exam2': [40, 50, 60]
}
df = pd.DataFrame(data)

print(df)
```

Output:

```
    Exam1  Exam2
0     90     40
1     80     50
2     70     60
```

Scaling the exam scores to a common range [0, 100]:

```
 from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler(feature_range=(0, 100))
df_scaled = scaler.fit_transform(df)

print(df_scaled)
```

Output:

```
    Exam1  Exam2
0     90.0    40.0
1     80.0    50.0
2     70.0    60.0
```

In this example, the `MinMaxScaler` maps the exam scores to a common range [0, 100], making it easier to compare and analyze the data.

Code Example (Python)

```
 import pandas as pd
from sklearn.preprocessing import MinMaxScaler

# Create a sample dataframe
data = {
    'Exam1': [90, 80, 70],
    'Exam2': [40, 50, 60]
}
df = pd.DataFrame(data)

scaler = MinMaxScaler(feature_range=(0, 100))
df_scaled = scaler.fit_transform(df)

print(df_scaled)
```

This code demonstrates how to apply min-max scaling to the exam scores in a Pandas dataframe using Scikit-learn's `MinMaxScaler`.

Curated by Brajesh Kumar