Artificial Intellignece & Machine Learning

Kushagra lakhwani (2021UCI8036)

November 15, 2023

Assignment 2

Normalization **vs** 0-1 Normalization (Min-Max Scaling) **vs** Standardization

Data normalization and ***standardization*** are two essential data preprocessing techniques commonly used in artificial intelligence and machine learning. Both methods aim to transform the data into a more suitable format for machine learning algorithms. However, they differ in their approach and have different advantages and disadvantages.

Normalization

Normalization, also known as min-max scaling, rescales the values of a feature to a specified range, typically between 0 and 1. This is done by subtracting the minimum value from each data point and then dividing by the range (maximum value minus minimum value).

$$x_{\text{normalized}} = \frac{x - \min(X)}{\max(X) - \min(X)}$$

Consider a dataset with a 'Salary' feature ranging from 20,000 to 100,000. To normalize the 'Salary' feature, we would apply the following formula:

```
salary_normalized = (salary - 20000) / (100000 - 20000)
```

This would transform the 'Salary' feature to a range between 0 and 1.

Standardization

Standardization, also known as Z-score normalization, centers the data around a mean of 0 and scales it to a standard deviation of 1. This is done by subtracting the mean from each data point and then dividing by the standard deviation.

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma}$$

Consider the same dataset with a ‘Salary’ feature ranging from 20,000 to 100,000. To standardize the ‘Salary’ feature, we would apply the following formula:

```
salary_standardized = (salary - mean) / standard_deviation
```

This would transform the ‘Salary’ feature to a mean of 0 and a standard deviation of 1.

Code

We can perform normalization and standardization using the scikit-learn library in Python:

```python
from sklearn.preprocessing import MinMaxScaler, StandardScaler

data = ...

# Normalization
scaler = MinMaxScaler()
data_normalized = scaler.fit_transform(data)

# Standardization
scaler = StandardScaler()
data_standardized = scaler.fit_transform(data)
```

Normalization vs Standardization

Normalization (Min-Max Scaling)	Standardization (Z-Score Scaling)
Transforms the data to a specified range, typically between 0 and 1.	Centers the data around a mean of 0 and scales it to a standard deviation of 1.
`x_normalized = (x - min(X)) / (max(X) - min(X))`	`x_standardized = (x - μ) / σ`
Not sensitive to the absolute values of the features.	Sensitive to the relative values of the features.
Uniform distribution, algorithms sensitive to absolute values.	Gaussian distribution, algorithms sensitive to relative values.
Salary is normalized to the range 0 to 1.	Salary is centered around a mean of 0 and a standard deviation of 1.
Preserves the original range of the data.	Easy to interpret, makes features comparable.
Sensitive to outliers.	Assumes Gaussian distribution.

- **Normalization** is generally preferred when the data is assumed to have a uniform distribution or when the algorithm is sensitive to the absolute values of the features.
- **Standardization** is generally preferred when the data is assumed to have a Gaussian distribution or when the algorithm is sensitive to the relative values of the features.

Conclusion

Normalization and standardization are essential data preprocessing techniques that can improve the performance of machine learning models. The choice between the two methods depends on the specific algorithm and the characteristics of the data. Understanding the strengths and weaknesses of each method is crucial for selecting the most appropriate approach for a given task.