

Why Feature Scaling Matters



What is Scaling?

Feature scaling is a technique to adjust the range of features in a dataset. It ensures that all features contribute equally to machine learning models. Without scaling, features with larger values can dominate the model.

Normalization and standardization are two common scaling methods. Normalization scales data to a fixed range, usually between 0 and 1. Standardization transforms data to have a mean of 0 and a standard deviation of 1.



Normalization Basics

Normalization adjusts values to a specific range, typically 0 to 1. It is useful when the data needs to be bounded within a specific interval. This method is commonly used in neural networks.

The formula for min-max normalization is:

$$X_{\text{normalized}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

This ensures all features are on the same scale, improving model performance.



Standardization Basics

Standardization transforms data to have a mean of 0 and a standardization deviation of 1. It is useful when data follows a Gaussian distribution. This method is often used in algorithms like SVM and PCA.

The formula for standardization is:

$$X_standardized = (X - \text{mean}(X)) / \text{standard_deviation}(X)$$
 This method is less affected by outliers compared to normalization.



When to Use?

Use normalization when the data needs to be bounded within a specific range. It is ideal for algorithms sensitive to the magnitude of features, like neural networks. Normalization is also useful for image processing.

Use standardization when the data follows a normal distribution. It is ideal for algorithms like linear regression and logistic regression. Standardization is also useful when features have different units.



Impact on Models

Feature scaling improves the convergence speed of gradient descent. It ensures that all features contribute equally to the model. Without scaling, features with larger values can dominate the model.

Scaling helps in avoiding numerical instability in algorithms. It ensures that the model does not get biased towards features with larger values. This leads to more accurate and reliable predictions.



Practical Example

Consider a dataset with features like age (18–80) and income (20000–200000). Without scaling, income will dominate the model due to its larger range. Normalization can scale both features to a range of 0 to 1.

After normalization, both features contribute equally to the model. This leads to better performance and more accurate predictions. Scaling is a crucial step in preprocessing.



Common Mistakes

Avoid scaling features that are already on the same scale. This includes binary features or features with the same units. Scaling such features can introduce unnecessary complexity.

Do not scale target variables in regression or classification tasks. Scaling is only applied to the input features. Scaling the target variable can lead to incorrect model predictions.



Key Takeaways

Feature scaling is essential for improving model performance. It ensures that all features contribute equally to the model. Normalization and standardization are two common scaling methods.

Choose the right scaling method based on the data and algorithm. Scaling is a crucial step in preprocessing and should not be overlooked. Proper scaling leads to more accurate and reliable models.



Follow for More



Like this post? Repost!

Share your thoughts in the comments!

I share little hacks and tips I actually use in
Data & AI — follow if you wanna see more!



Shoaib Akthar
@theshoaibakthar

Follow for more Data Analytics, Data Science and AI insights