02 - Feature engineering Python

File Edit View Run Help Last edit was 51 minutes ago Provide feedback

Run all | cluster_01 | Schedule | Share

Table of contents

Cmd 38

4) Feature Scaling

Cmd 39

Markdown

- **Feature scaling, also known as data normalization:** The process of transforming numerical features in a dataset to a common scale. It is a crucial step in data preprocessing and feature engineering, as it helps to bring the features to a similar range and magnitude. The goal of feature scaling is to ensure that no single feature dominates the learning process or introduces bias due to its larger values

- **There are two common methods for feature scaling:**

1) Standardization (Z-score normalization): In this method, each feature is transformed to have zero mean and unit variance. The formula for standardization is: x_scaled = (x - mean) / standard_deviation. Standardization ensures that the transformed feature has a mean of 0 and a standard deviation of 1.

2) Min-Max scaling: In this method, each feature is scaled to a specific range, typically between 0 and 1. The formula for min-max scaling is: x_scaled = (x - min) / (max - min). Min-max scaling preserves the relative ordering of values and ensures that the transformed feature is bounded within the defined range.

Cmd 40

feature has a mean of 0 and a standard deviation of 1.

2) Min-Max scaling: In this method, each feature is scaled to a specific range, typically between 0 and 1. The formula for min-max scaling is: x_scaled = (x - min) / (max - min). Min-max scaling preserves the relative ordering of values and ensures that the transformed feature is bounded within the defined range.

Cmd 40

Markdown

Feature scaling is important for several reasons:

1. Gradient-based optimization algorithms, such as gradient descent, converge faster when features are on a similar scale. This helps in achieving faster convergence and more efficient training of machine learning models.

2. Features with larger scales can dominate the learning process, leading to biased results. Scaling the features ensures that no single feature has undue influence on the model.

3. Many machine learning algorithms, such as K-nearest neighbors (KNN) and support vector machines (SVM), rely on calculating distances between data points. If features are not on a similar scale, features with larger values can dominate the distance calculations, leading to suboptimal results.

4. Some algorithms, such as principal component analysis (PCA), assume that the data is centered and on a similar scale. Feature scaling is necessary to meet these assumptions and obtain meaningful results.

Cmd 41