Step 06

DATA PREPROCESSING

- Standardization
- Scaling
- Transformation
- Data splitting

Data preprocessing

- Data preprocessing is a critical step in building a successful machine learning model.
- The main goal of data preprocessing is to enhance the quality of the data and prepare it for further analysis, modeling, and prediction.
- This involves a series of techniques and procedures such as data cleaning, data transformation, data normalization, feature selection, and splitting the data into training and testing sets.

Let's discuss some of the important ones. This includes,

- Standardization
- Scaling
- Transformation
- Data splitting

Standardization

- It is the process of transforming data so that it has a mean of zero and a standard deviation of one.
- By standardizing the data, we can ensure that variables with larger values do not dominate the analysis.

Importance :

It ensures that each feature has the same scale. So that the model can learn from each feature equally.

Example :

- Suppose we have a dataset with two features, age and income. Age is measured in years and income is measured in rupees.
- We can standardize the data to remove the different units by calculating the z-score for each data point.
- This will give us data that is unitless and easier to work with.

Here are the common techniques used for standardization:

- Z-Score Standardization
- Scaling to Unit Length

Scaling

- Scaling is the process of transforming data so that it falls within a specific range, often between 0 and 1.
- This is done by subtracting the minimum value from each data point and then dividing by the range (maximum value - minimum value).

Importance :

Scaling can ensures that the model can learn from the features with small values as well as features with large values.

Example :

- Suppose we have a dataset with two features, age and height.
- Age is measured in years and height is measured in centimeters.
- We can scale the data to ensure that both features are within the same range, say between 0 and 1.
- This will help to remove any differences in the scale of the data and make it easier to compare the features.

Here are the common techniques used for scaling:

- Min-Max Scaling (Normalization)
- Robust Scaling
- Max Absolute Scaling

Transformation

- Transforming is the process of applying mathematical functions to the data to transform it into a new format.
- For example, logarithmic, exponential, and power transformations can be applied to the data to transform it into a more normal distribution.

Importance :

Transforming is important because it can improve the accuracy of the model by reducing the impact of outliers and skewed distributions.

Example :

- Suppose we have a dataset with a feature that has a skewed distribution, such as income.
- We can transform the data using a logarithmic function, which can help to make the distribution more normal.

Transformation can help to improve the performance of some machine learning algorithms, which may be sensitive to the distribution of the data.

Here are the some common techniques used for transformation,

- Log Transform
- Power Transform
- Box-Cox Transform

Data splitting

- Data splitting is the process of dividing the dataset into two or more subsets.
- The most common way of splitting data is into training and testing sets.
- The training set is used to train the machine learning model, while the testing set is used to evaluate the model's performance

- The purpose of data splitting is to assess the model's performance on unseen data.
- If we do not split the data, the model may simply memorize the training data and perform poorly on new data.

Common techniques used for splitting the data are :

- Random Split
- Stratified Split
- Time-Series Split

Here are some best practices for applying standardizing, scaling, and transforming in data preprocessing:

- Always visualize the data before and after applying these techniques to ensure that they have been applied correctly.

- Consider the requirements of the model and the nature of the data before applying standardizing, scaling, or transforming.
- Apply standardizing, scaling, or transforming only to the training set and not the test set, to prevent data leakage and overfitting.
- Experiment with different techniques and parameters to see which method works best for the data and the model.

In summary, scaling, standardizing, transforming, and data splitting are all important steps in data preprocessing.

Each method has its own use case and can help to improve the performance of a machine learning model.

Happy learning!😊

THANK YOU

@ Gangadhar Neelam