Data Preprocessing-Data Splitting

Data Splitting

Data splitting is a crucial step in data preprocessing that involves dividing the dataset into two or more subsets, typically for the following purposes:

1. **Training**: A subset of the data used to train a machine learning model.
2. **Validation**: Another subset used to evaluate the performance of the trained model and make adjustments as needed.
3. **Testing**: The final subset used to assess the performance of the optimized model.

The goal of data splitting is to ensure that the model's performance is evaluated on unseen data, providing a more accurate representation of its real-world effectiveness.

Types of Data Splitting

1. **Random Split**: A random division of the dataset into subsets.
2. **Stratified Split**: Ensures that each subset has a similar distribution of classes or labels.
3. **Time-Based Split**: Divides data based on a specific time period, e.g., historical vs. current data.

Example:

Suppose we have a dataset containing information about customer transactions:

Customer ID	Transaction Date	Amount
1	2022-01-01	100
1	2022-01-15	200
2	2022-02-01	50
...	...	...

We want to train a model to predict the likelihood of a customer making a purchase based on their transaction history. We split our dataset into:

- **Training Set** (60%): Used to train the model.
    - Contains transactions from January and February 2022.
- **Validation Set** (20%): Used to evaluate the performance of the trained model.
    - Contains transactions from March 2022.
- **Testing Set** (20%): Used to assess the final performance of the optimized model.
    - Contains transactions from April 2022.

In this example, we've applied a random split with stratification by date, ensuring that each subset has a similar distribution of classes or labels. The training set is used to train the model, while the validation and testing sets are used for evaluation and final assessment, respectively.

By splitting our dataset in this way, we can develop a more accurate and reliable machine learning model that generalizes well to unseen data.

Curated by Brajesh Kumar