Data Preprocessing-Data Sampling

Data Sampling: A Crucial Step in Data Preprocessing

Data sampling is a technique used to select a subset of data from a larger dataset. The goal of data sampling is to create a representative sample that accurately reflects the characteristics and patterns of the original dataset, while also reducing the computational cost and storage requirements.

Why is Data Sampling Important?

1. **Reducing Computational Cost**: Large datasets can be computationally expensive to process, especially for complex algorithms.
2. **Improving Model Accuracy**: A representative sample can improve the accuracy of machine learning models by reducing overfitting.
3. **Decreasing Storage Requirements**: Smaller datasets require less storage space.

Types of Data Sampling:

1. **Random Sampling**: Every data point has an equal chance of being selected.
2. **Stratified Sampling**: The dataset is divided into subgroups (strata) based on specific variables, and a random sample is taken from each subgroup.
3. **Cluster Sampling**: A subset of clusters or groups is randomly selected from the population.
4. **Systematic Sampling**: Data points are selected at regular intervals.

Example:

Suppose we have a dataset of customers with 10,000 rows, including demographic information (age, gender, income), purchase history, and other relevant variables. We want to train a machine learning model to predict customer churn.

- **Initial Dataset:** 10,000 rows
- **Desired Sample Size:** 1,000 rows
- **Sampling Method:** Stratified Sampling (based on age)
- **Subgroups:**
    - Age < 30: 2,000 customers
    - Age between 30-50: 4,000 customers
    - Age > 50: 4,000 customers

We randomly select 1,000 rows from each subgroup to create a representative sample.

Advantages of Data Sampling in this Example:

- Reduced computational cost and storage requirements
- Improved model accuracy due to the representative sample
- Faster model training time

However, data sampling can introduce biases if not done properly. It's essential to carefully select the sampling method and ensure that the sample is representative of the original dataset.

In summary, data sampling is a crucial step in data preprocessing that helps reduce computational cost, improve model accuracy, and decrease storage requirements while creating a representative subset of the larger dataset.

Curated by Brajesh Kumar