Author: Madhurima Rawat

Data Sampling and Stratification

Implement data sampling techniques to generate representative subsets of large datasets, and stratify the data based on specific criteria for balanced sampling.

Explanation of the Code

1. **Dataset**: The Iris dataset is loaded using `load_iris()` from `scikit-learn`. This dataset contains 150 samples from three species of iris flowers, with four features for each sample (sepal length, sepal width, petal length, petal width), and a target variable indicating the species.
2. **Random Sampling**: We first split the dataset into a training set (80%) and testing set (20%). Random resampling is applied to the training set to generate a balanced subset (100 samples), and we visualize the class distribution before and after sampling.
3. **Stratified Sampling**: We apply stratified sampling using `train_test_split` with the `stratify` parameter set to the target variable (`y_train`). This ensures that each class is represented proportionally in the training and testing sets. We also visualize the class distribution before and after stratified sampling.
4. **Statistical Measures**: We calculate and print various statistical measures:
    - **Mean**: The average of each feature.
    - **Standard Deviation**: The spread of the feature values around the mean.
    - **Variance**: The squared spread of the feature values around the mean.
5. **Visualizations**: We plot bar charts of the class distribution before and after sampling to visually compare how the sampling methods affect the balance of the dataset.

4. Output

Class Distribution Before and After Random Sampling:

- **Before**: The class distribution in the original dataset.
- **After**: The class distribution after random sampling to balance the dataset.

Class Distribution Before and After Stratified Sampling:

- **Before**: The class distribution before stratified sampling.
- **After**: The class distribution after applying stratified sampling to ensure proportional class representation.

Statistical Measures:

- **Original Data**: The mean, standard deviation, and variance of the original dataset.
- **Resampled Data**: The mean, standard deviation, and variance of the resampled dataset.

Explanation of the Concepts

1. Introduction

Data sampling is a key technique used in data science to select subsets from large datasets for various purposes, such as model training, validation, and analysis. Stratification ensures that subsets are representative and balanced based on specific criteria, ensuring fairness and reducing bias in model performance.

2. Random Sampling

Random sampling involves selecting a random subset from the data. It is the simplest method and is useful when the dataset is relatively balanced and no specific subgroup prioritization is required.

3. Stratified Sampling

Stratified sampling divides the population into subgroups (strata) based on certain criteria and ensures that each subgroup is proportionally represented in the sample. This method is especially important when the dataset is imbalanced or has skewed distributions, as it helps to maintain diversity in the sample.

4. Statistical Measures

We calculated the following statistical measures to understand the effect of sampling:

- **Mean**: Measures the central tendency of each feature.
- **Standard Deviation**: Indicates the spread of the data.
- **Variance**: A squared measure of data spread, showing how far individual data points are from the mean.

Conclusion

By using random and stratified sampling techniques, we can generate balanced subsets for model training and testing. Stratified sampling is particularly useful in ensuring that minority classes are well-represented, preventing bias in machine learning models.