ML Lab Assignment 2: Performance Estimation using Bootstrap Resampling

25MCMI20, Sachin A

30/01/2026

1 Introduction

In this experiment, I implemented the bootstrap resampling technique to estimate the performance of a machine learning classifier. Unlike a single train–test split, bootstrapping allows us to understand how model performance varies when the training data changes. A Decision Tree classifier was used on the Iris dataset, and the model performance was evaluated using accuracy as the metric.

2 Dataset Description

The Iris dataset was used in this experiment. It is a well-known multi-class classification dataset containing 150 samples belonging to three different species of iris flowers. Each sample consists of four numerical features:

- Sepal length
- Sepal width
- Petal length
- Petal width

The target variable represents the species class of the flower.

3 Libraries Used

The following Python libraries were used:

- **NumPy**: Used for numerical computations and statistical calculations.
- **Matplotlib**: Used for visualizing the distribution of accuracy values.
- **Scikit-learn**: Used for dataset loading, model training, bootstrap resampling, and performance evaluation.

4 Train–Test Split Baseline

As a baseline, the dataset was first divided into training and testing sets using a 70–30 split. A Decision Tree classifier was trained on the training set and evaluated on the test set. The accuracy obtained from this single train–test split serves as a reference point for comparison with the bootstrap-based performance estimation.

5 Bootstrap Resampling Method

Bootstrap resampling was applied only to the training data. In each iteration, a new training set was created by sampling the original training data with replacement. A Decision Tree classifier was trained on this bootstrapped dataset and evaluated on the same fixed test set.

This process was repeated for 100 bootstrap iterations. For each iteration, the test accuracy was recorded. The mean accuracy and standard deviation across all bootstrap samples were then computed. This approach provides an estimate of the average model performance as well as the variability caused by different training data samples. The bootstrap procedure was performed using 100 resampled training sets, and model performance was evaluated on a fixed test set using accuracy as the evaluation metric.

6 Results

Method	Accuracy
Train–Test Split	0.9333
Bootstrap Mean	0.9197
Bootstrap Standard Deviation	0.0324

Table 1: Comparison of Decision Tree performance using train–test split and bootstrap resampling

6.1 Baseline Accuracy

The Decision Tree model trained using a single train–test split achieved a baseline accuracy on the test set.

6.2 Bootstrapped Accuracy Statistics

From the bootstrap resampling process, the following statistics were obtained:

- Mean bootstrap accuracy
- Standard deviation of bootstrap accuracy

The mean bootstrap accuracy was observed to be close to the baseline train–test accuracy. The standard deviation reflects the variability in model performance due to changes in the training data.

6.3 Accuracy Distribution

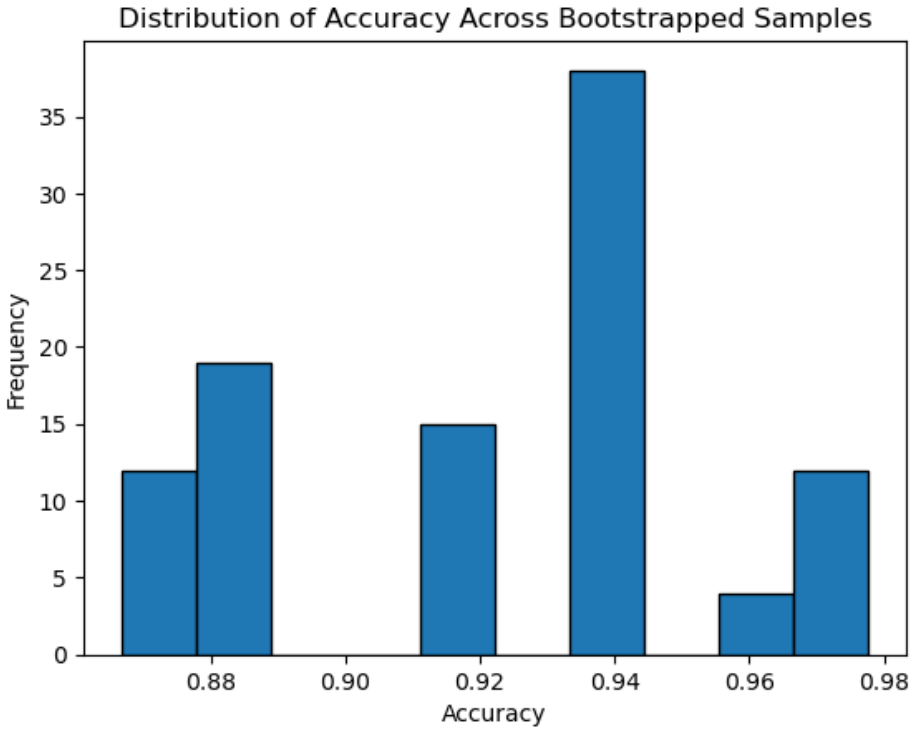

Figure 1: Distribution of Decision Tree accuracy across bootstrap resampled training sets

A histogram was plotted to visualize the distribution of accuracy values obtained across the bootstrap samples. The spread of the histogram indicates that the Decision Tree model exhibits some sensitivity to variations in the training data, which is expected given the high-variance nature of decision trees.

7 Comparison with Train–Test Split

The train–test split provides a single point estimate of accuracy, whereas bootstrap resampling provides a distribution of accuracy values. While the mean bootstrap accuracy does not necessarily improve over the baseline accuracy, bootstrap resampling offers additional insight into the stability and reliability of the model by quantifying performance variability.

8 Conclusion

In this experiment, bootstrap resampling was used to estimate the performance of a Decision Tree classifier on the Iris dataset. The results show that while the average accuracy

obtained through bootstrap resampling is comparable to the baseline train–test accuracy, the standard deviation provides valuable information about the variability of the model. This demonstrates that bootstrap resampling is useful for performance estimation rather than performance improvement.