Course Project Proposal

Jennifer Zhuang

Enhancing Machine Learning Model Evaluation Using Bootstrapping Techniques

Introduction:

In this project, I aim to apply **bootstrapping**, a powerful resampling method, to improve the evaluation and selection process of machine learning models. Traditional evaluation methods, such as a single train-test split, often suffer from variability due to the random nature of data partitioning. Bootstrapping offers a statistically grounded solution by repeatedly resampling the dataset, allowing for more comprehensive model performance analysis.

Statistical Foundation: Understanding Bootstrapping

Bootstrapping is a computational statistics method used to estimate the distribution of a sample statistic by resampling the dataset with replacement. This technique enables us to make inferences about the population without relying on strong assumptions about its underlying distribution.

The paper will include an in depth statistical explanation of boostrapping with necessary and relevant mathemtical theories / equations.

- Explanation of resampling procedure
- Boostrap estimate of a statistics
- Distribution and confidence interval of statistic
- Bias and Variance estimate of statistic

Project Objectives

1. Apply bootstrapping techniques to evaluate the **variability and robustness** of machine learning models.
2. Use bootstrapped estimates to construct confidence intervals for performance metrics, providing more reliable insights for model comparison.
3. Analyze and compare the stability of different machine learning models using these bootstrapped statistics.

Dataset:

For this project, I will use the **Boston Housing Prices Dataset**, a classic and widely-used benchmark dataset for regression algorithms. This dataset is sourced from the **StatLib library** maintained at Carnegie Mellon University. It contains the following features

- **CRIM:** Per capita crime rate by town
- **RM:** Average number of rooms per dwelling
- **LSTAT:** Percentage of lower status of the population
- **MEDV:** Median value of owner-occupied homes (response variable)
- The dataset will provide a standard context for evaluating performance and for comparison across different machine learning models.

Methodology:

1. Data Preprocessing:
    - Handle missing values, normalize features, and encode categorical variables.
    - Split the dataset into training and testing sets.
2. Implement ML Models:
    - **NOTE**: We will implement the regression models using **Python** (not in R code if possible), due to its extensive ecosystem of machine learning libraries such as Scikit-Learn, TensorFlow, and PyTorch, which provide robust implementations of various algorithms.
    - Additionally, the StatLib library, where the Boston Housing Prices Dataset originates, is implemented in Python.
    - Python's dominance in the field of machine learning makes it the preferred choice for this project s.t. we can leverage well-established tools.
    - We consider implementing the following ML models:
        - Linear Regression
        - Random Forest
        - Support Vector Machine
        - Gradient Boosting Regressor
        - Neural Network
3. Bootstrapping Process:
    - Generate 1,000 bootstrap samples from the training data.
    - Train each model on each bootstrap sample and evaluate it using metrics like **mean squared error (MSE)**, **R-squared score**, and **mean absolute error (MAE)**.
    - Calculate the **mean, variance, bias**, and **confidence intervals** for each metric using the bootstrapped estimates.
4. Model Evaluation and Comparison:
    - Compare the models based on the bootstrapped statistics, focusing on the mean, variance, and width of confidence intervals.
    - Visualize the metric distributions using **box plots** and **histograms**.
    - Determine the most stable and robust model by analyzing the variability of the performance metrics.

Expected Results:

- The bootstrapping approach is expected to reveal differences in metric stability
- Confidence intervals will provide a clearer understanding of each model's performance, allowing for more informed model selection beyond simple point estimates of MSE or R-squared score.

References:

1. Increasing Transparency in Machine Learning Through Bootstrap Simulation and Shapley Additive Explanations Link to paper

2. Estimating Neural Network's Performance with Bootstrap Link to paper