Bootstrapping

- Bootstrapping generates the training set using sampling with replacement when insufficient datasets
- N random training samples are drawn from the full dataset of N records with replacement (Training).
- The unselected samples become the test set.

$$e_{boot} = \frac{1}{b}\sum_{i}^{b}(0.632e_i + 0.368e_s)$$

where e_s is the training error of the whole set. as $N \to \infty$ the probability of any given object being in the bootstrap sample is ≈ 0.632.

Powerful statistical method used for estimating the distribution of a statistic (like the mean or variance) from a set of data. It's widely used because it's relatively simple and doesn't make strong assumptions about the distribution of the data, unlike many traditional statistical methods.
Here's a basic overview of how bootstrapping works:

- **Resampling**: From the original dataset (of size N), repeatedly draw samples, typically with replacement, to create many "bootstrap samples." The size of each bootstrap sample is usually the same as the original dataset.
- **Calculate Statistic on Each Sample**: For each bootstrap sample, calculate the statistic of interest (e.g., mean, median, standard deviation).
- **Estimate the Distribution**: By repeating the resampling and calculation steps many times (often thousands or more), you can build up a distribution of the statistic.
- **Analyze the Distribution**: This bootstrap distribution of the statistic can be used to estimate its standard error, confidence intervals, and other properties.

Some key points about bootstrapping include:

- **Non-Parametric**: Bootstrapping is a non-parametric method, meaning it doesn't assume the data follows a particular distribution (like normal distribution).
- **Versatility**: It can be applied to many different types of statistics and is especially useful when the theoretical distribution of a statistic is complex or unknown.
- **Simplicity and Power**: The method is computationally simple yet powerful, relying on random sampling with replacement.
- **Limitations**: While bootstrapping is widely applicable, it may not perform well with very small sample sizes or with data that are not representative of the population. Additionally, it can be computationally intensive.

In summary, bootstrapping is a useful technique for statistical inference,

especially in situations where traditional methods are difficult to apply or when the sample size is not large enough to rely on asymptotic properties. It's used extensively in fields like biostatistics, econometrics, and machine learning.