### Introduction

Ensemble Methods (EM) combine multiple DTs in a statistical ensemble in order to improve the forecasting accuracy. EM can be used in DTs as well as in regression analysis and classification machine learning methods.

### EM Methods

The most popular EM methods are the following:

* Bagging
* Boosting
* Random forests

Random forests and bagging use a frequentist approach technique (bootstrapping) in order to generate samples from a population without taking additional training data. Bootstrap or bootstrapping technique will be explained in the following paragraph.

### Bootstrapping

This method is similar to Monte Carlo simulations.

|  |  |
| --- | --- |
| **Monte Carlo experiment** | **Bootstrapping** |
| Generate random variables from a known distribution (generally the normal distribution) | Random variables are drawn from their observed distribution (plug-in principle) |

In machine learning, bootstrapping is used for creating multiple training sets. This allows meta-learner to reduce the variance of predictions.

Bootstrap sample – random sample of size T drawn with replacement from the observed data putting a probability of 1/T on each of the observed values.

Example:

Suppose we have 10 values of

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  | 0.8 | 3.5 | 0.5 | 1.7 | 7.0 | 0.6 | 1.3 | 2.0 | 1.8 | -0.5 |

Sample mean = 1.87

Standard deviation = 2.098

Probability of each element of the sample = 1/10

We generate bellow 3 different bootstrap samples. The values are drawn with replacement. The table is presented bellow:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |  |
|  | 3.5 | 1.7 | -0.5 | 0.5 | 1.8 | 2.0 | 1.7 | 0.6 | 0.6 | 7.0 | 1.89 |
|  | -0.5 | 0.6 | 0.6 | 0.8 | 1.7 | 7.0 | 1.8 | 3.5 | 1.8 | 0.8 | 1.81 |
|  | 0.5 | 0.6 | 7.0 | 1.3 | 1.3 | 7.0 | 1.3 | 1.8 | 3.5 | 0.6 | 2.49 |

- denotes bootstrap sample

- sample mean

0.6 and 1.7 appear twice in the first bootstrap sample

0.6, 0.8 and 1.8 appear twice in the second sample

1.3 appears three times in the third bootstrap sample

If the number of samples , then the moments of the bootstrap samples converge to the population moments according to Efron (1979) the pioneer of bootstrapping.

### Bootstrapping regression coefficients

Consider the following linear regression between two time series variables:

### – time series variables

The statistical properties of and can be checked using Monte Carlo simulations. In the using residuals generated from a normal distribution, we can use actual regression residuals. This approach is called bootstrapped residuals.

Steps for bootstrapping residuals:

STEP 1: Estimate regression coefficients using OLS and calculate residuals as following:

STEP 2: Generate a bootstrap sample of the error terms containing the elements of . The bootstrap sample is used to calculate series (bootstrap series).

and -> fixed values.

-> fixed

STEP 3: New values for and are estimated using bootstrap sample based on the resulting values of and .

STEP 4: STEPS 2 and 3 are repeated many times and the sample statistics for of and are calculated. They should be distributed in the same way as and .

Example:

95 % confidence interval for -> the interval between the lowest 2.5 % and highest 97.5 % of values.