Random Forest Method I

Learn How Ensemble Learning Can be Used for Predictive Modeling

Contents

1. Introduction to Bootstrapping
2. Understanding Bagging
3. Quick Re-look at Decision Trees
4. Introduction to Random Forest

Bootstrapping

Bootstrapping is a method for estimating the sampling distribution of an estimator by resampling with replacement from the original sample

- The method is especially useful in situations where sampling distribution of estimator is not standard distribution.
- The method can be used in any statistical inference problem.
- The use of the term 'bootstrap' comes from the phrase “To pull oneself up by one's bootstraps ” - generally interpreted as succeeding in spite of limited resources.

Bootstrapping

- Many conventional statistical methods of analysis make assumptions about normality, including correlation, regression, t tests, and analysis of variance. When these assumptions are violated, such methods may fail.

- Bootstrapping, a data-based simulation method, is steadily becoming more popular as a statistical methodology. It is intended to simplify the calculation of statistical inferences, sometimes in situations where no analytical answer can be obtained.

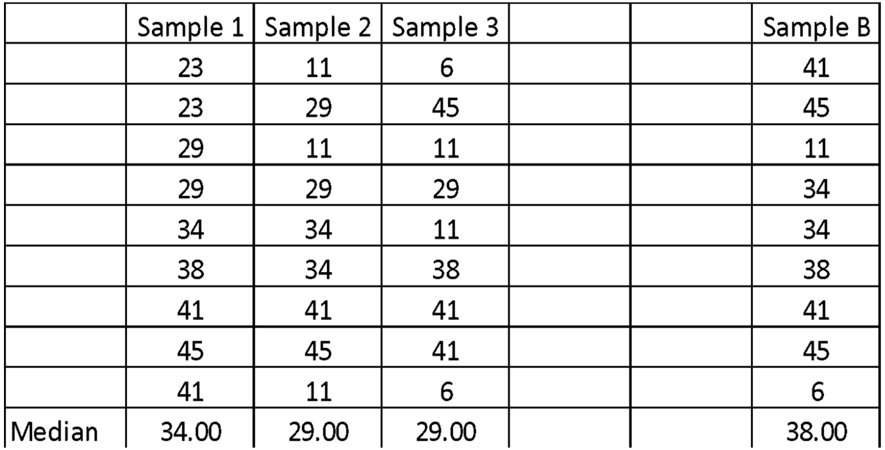
- As computer processors become faster and more powerful, the time and effort required for bootstrapping decreases to levels where it becomes a viable alternative to standard parametric techniques.

Bootstrapping

Suppose the original sample is 12, 23, 11, 29, 34, 38, 41, 45, 6

Median = 29.00

We now draw multiple random samples using Bootstrapping



	Sample 1	Sample 2	Sample 3			Sample B
	23	11	6			41
	23	29	45			45
	29	11	11			11
	29	29	29			34
	34	34	11			34
	38	34	38			38
	41	41	41			41
	45	45	41			45
	41	11	6			6
Median	34.00	29.00	29.00			38.00

Sampling distribution of sample median is generated

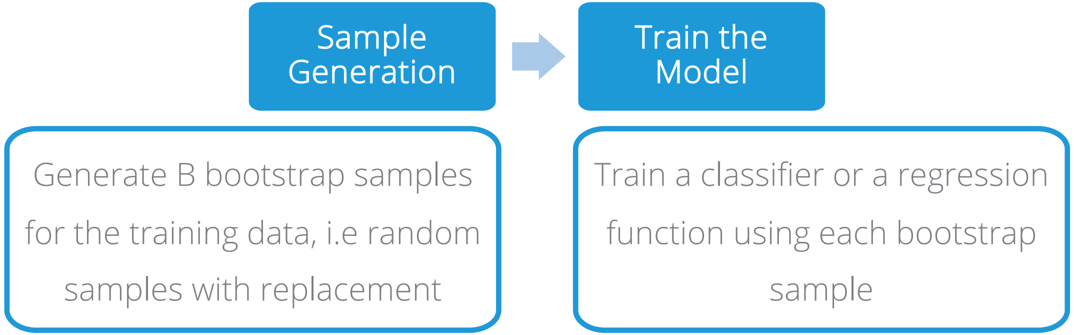
Assuming B=1000, **25th** value and **975th** value will provide 95% confidence interval for median

Bagging

- The term “Bagging” was Introduced by Breiman (1996).
- “Bagging” stands for “Bootstrap Aggregating”.
- It is an ensemble method: a method of combining results from multiple resamples.
- Ensemble method can also be applied by using different classifiers for a given sample.

Bagging Method Framework

Bagging has two basic steps:



Sample Generation

Train the Model

Generate B bootstrap samples for the training data, i.e random samples with replacement

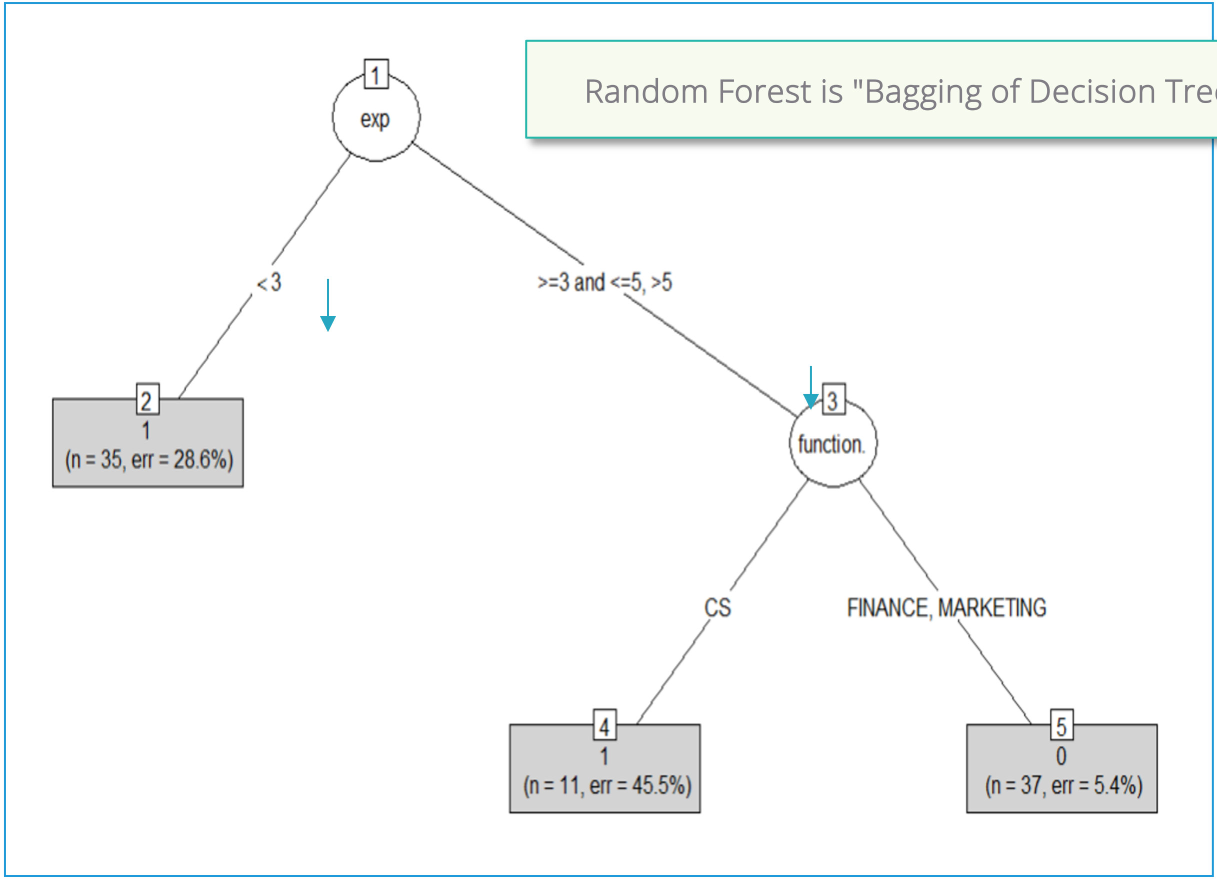
Train a classifier or a regression function using each bootstrap sample

Model for **classification** Majority vote on the classification

Model for **regression** Average of the predicted value

Bagging improves performance for unstable classifiers which vary significantly with small changes in the data set

Re-look at the CHAID Decision Tree



Random Forest is "Bagging of Decision Trees"

Random Forest Classifier

Random Forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.

- The term came from random decision forests that was first proposed by Tin Kam Ho of Bell Labs in 1995.
- The method combines Breiman's “Bagging" idea and the random selection of features.

Random Forest Algorithm

1. Grow a forest of many trees (R default is 500)

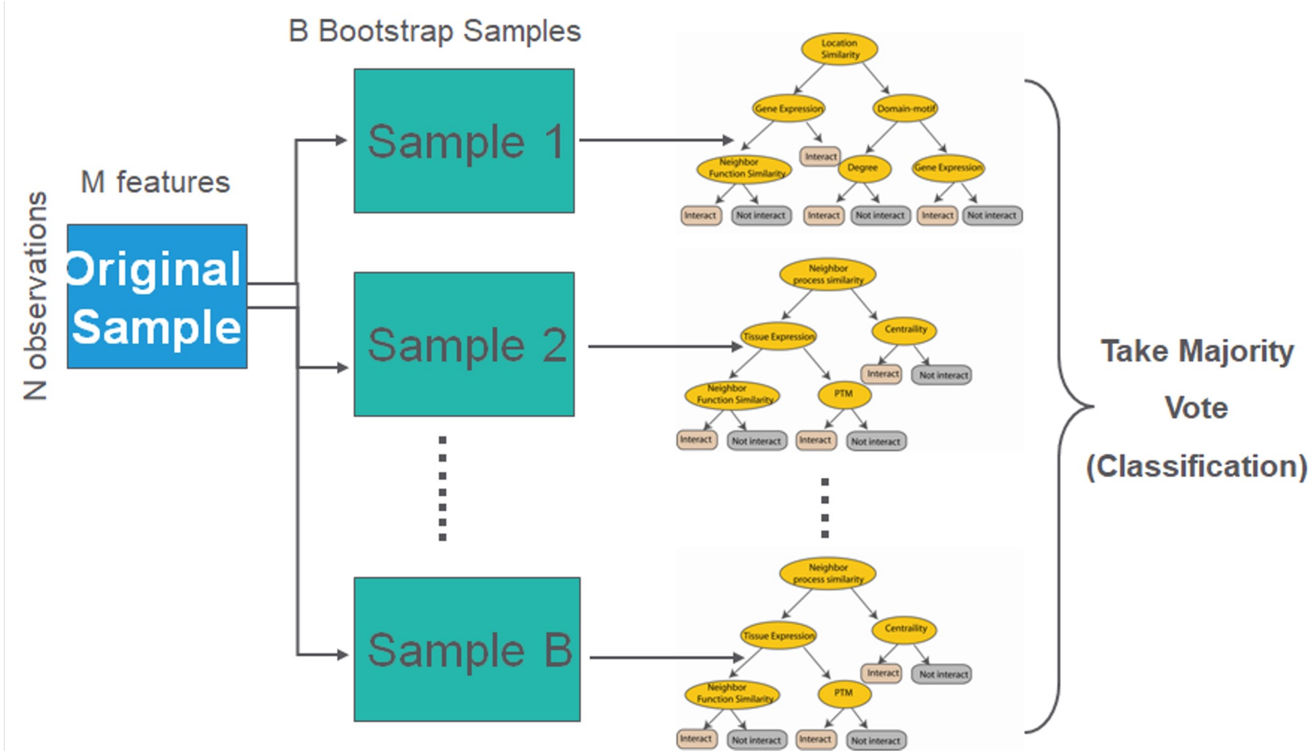
2. Grow each tree on an independent **bootstrap sample*** from the training data

3. At each node:
   - Select *m* variables **at random** out of all *M* possible variables (independently for each node)
   - Find the best split on the selected *m* variables

4. Grow the trees to maximum depth

5. Vote/average the trees to get predictions for new data

Random Forest Algorithm



B Bootstrap Samples

M features

N observations

Original Sample

Sample 1

Sample 2

Sample B

Take Majority Vote (Classification)

Random Forest Algorithm

- In bootstrap samples of data, approx. 2/3 of original values get included in each sample.
- Fit a tree to its greatest depth determining the split at each node through minimizing the loss function considering a random sample of covariates (size is user specified).
- For each tree,
  - Predict classification of the leftover 1/3 using the tree, and calculate the misclassification rate = *Out of Bag (OOB) Error Rate*
  - For each variable in the tree, permute the variables values and compute the OOB error, compare to the original OOB error, the increase is a indication of the variable's importance .

Random Forest Algorithm

- Aggregate OOB error and importance measures from all trees to determine overall OOB error rate and Variable Importance measure

OOB Error Rate

Calculate the overall percentage of misclassification

Variable Importance

Average increase in OOB error over all trees

Quick Recap

Bootstrapping

- Method for estimating the sampling distribution of an estimator by resampling with replacement from the original sample

Bagging

- “Bagging” stands for “Bootstrap Aggregating”
- It is an ensemble method: a method of combining results from multiple resamples

Random Forest Method

- Its an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees
- Random forests also work for regression problems
- The method combines Breiman's “Bagging" idea and the random selection of features