Regression Trees

Jiaqi Zeng 15220152202409

April 22, 2019

Abstract

This paper will introduce Regression Trees about its process and goal. Then we will talk about how to build a regression tree and trade off. Also, the Strengths and weaknesses of regression trees will be pointed out.

1 Introduction

Basic regression trees partition a data set into smaller groups and then fit a simple model (constant) for each subgroup.

There are many methodologies for constructing regression trees but one of the oldest is known as the classification and regression tree (CART) approach developed by Breiman et al. (1984).

This tutorial focuses on the regression part of CART. Basic regression trees partition a data set into smaller subgroups and then fit a simple constant for each observation in the subgroup. The partitioning is achieved by successive binary partitions based on the different predictors. The constant to predict is based on the average response values for all observations that fall in that subgroup.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the training data. A regression tree of M leaves can be thought of as dividing the predictor space into M regions $R_1, \ldots, R_M$, each corresponding to a leaf of the tree. For each R_M , m = 1, . . . , M, let

$$\overline{y}_m = \frac{1}{n_m} \sum_{x_i \in R_m} y_i$$

where

$$n_m$$

is the number of observations in R_M.

The estimated regression function is:

$$\widehat{f}(x) = \sum_{m=1}^{M} \overline{y}_m \cdot \mathcal{I}\{x \in R_m\}$$

2 How to build a regression tree

In order to build a regression tree, we should know about the goal: find $R_1, \ldots, R_M$ to minimize E_{in}.In the regression setting, using the L2 loss, we minimize the RSS:

$$\text{RSS} = \sum_{m=1}^{M} \sum_{x_i \in R_m} (y_i - \overline{y}_m)^2$$

There is often a balance to be achieved in the depth and complexity of the tree to optimize predictive performance on some unseen data. To find this balance, we typically grow a very large tree and then prune it back to find an optimal subtree.

We find the optimal subtree by using a cost complexity parameter (α) that penalizes our objective function for the number of terminal nodes of the tree (T) as follow:

$$minimize\{SSE + \alpha|T|\}$$

For a given value of α, we find the smallest pruned tree that has the lowest penalized error.As with these regularization methods, smaller penalties tend to produce more complex models, which result in larger trees. Whereas larger penalties result in much smaller trees.

Consequently, as a tree grows larger, the reduction in the SSE must be greater than the cost complexity penalty. Typically, we evaluate multiple models across a spectrum of α and use cross-validation to identify the optimal α and, therefore, the optimal subtree.

3 Advantages and disadvantages

As for regression trees, they have some advantages comparing to other methods:

1. They're very interpretable and fast to make predictions (no complicated calculations, just looking up constants in the tree).

2. The model provides a non-linear jagged response, so it can work when the true regression surface is not smooth. If it is smooth, though, the piecewise-constant surface can approximate it arbitrarily closely (with enough leaves).

3. There are fast, reliable algorithms to learn these trees.

But there are also some significant weaknesses:

1. Single regression trees have high variance, resulting in unstable predictions (an alternative subsample of training data can significantly change the terminal nodes).

2. Due to the high variance single regression trees have poor predictive accuracy.

4 Bagging

Single tree models suffer from high variance. Although pruning the tree helps reduce this variance, there are alternative methods that actually exploite the variability of single trees in a way that can significantly improve performance over and above that of single trees.

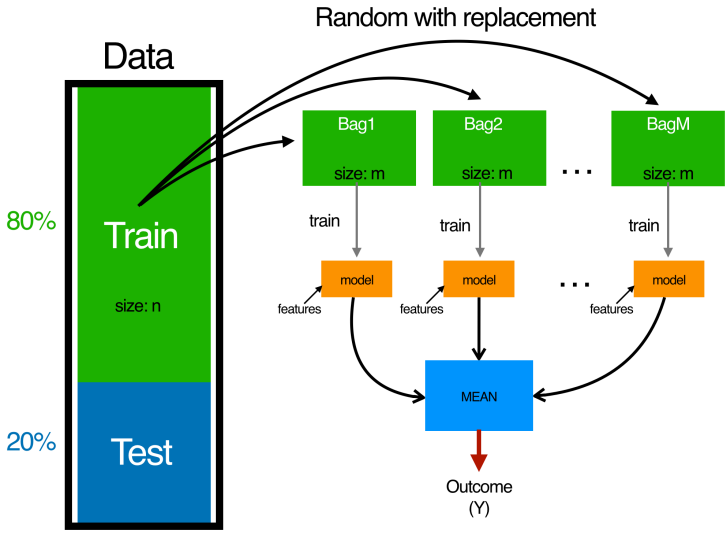
Bootstrap aggregating (bagging) is one such approach (originally proposed by Breiman, 1996).

Bagging combines and averages multiple models. Averaging across multiple trees reduces the variability of any one tree and reduces overfitting, which improves predictive performance. Bagging follows three simple steps:

1. Create m bootstrap samples from the training data. Bootstrapped samples allow us to create many slightly different data sets but with the same distribution as the overall training set.

2. For each bootstrap sample train a single, unpruned regression tree.

3. Average individual predictions from each tree to create an overall average predicted value.



This process can actually be applied to any regression or classification model; however, it provides the greatest improvement for models that have high variance. For example, more stable parametric models such as linear regression and multi-adaptive regression splines tend to experience less improvement in predictive performance.

5 Conclusion

Regression tree is an important method in machanism learning. It's easy to understand and can be used in non-linear estimation, but it needs many data and has high variance. So when we choose methods to estimate, we will consider about its advantages and disadvantages.

References

[1] UC Business Analytics R Programming Guide

[2] Classification and Regression Trees by L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone.