Machine Learning via Gradient Boosting

Shaurya Jauhari (Email: shauryajauhari@gzhmu.edu.cn)

2019-11-01

Contents

Introduction

In this module, we're going to explore the gradient boosting algorithm. *Gradient boosting could be construed as a unison of boosting and gradient descent techniques*, i.e. boosting carried out on the gradient descent of the cost function. The gradient descent approach, in loose terms, is the protocol to reach the minima of the function that maps the error terms. Our objective is to minimize the error (error between the true values and the predicted values), thereby finding appropriate coefficients that fit the objective function's equation in a way that best approximations are engendered.

Boosting is a way to derive random sub-samples of the same size as the original one.

This aspect of gradient boosting differentiates it from randome forests that aim to aggregate outputs from multiple single-decision trees, implementing basically a consensus-based scheme. Gradient Boosting would contrarily penalize the leafs from the same decision tree until the gradient descent strategy's optimal parameters are met.

Package Installation

The package in question is "gbm".

```
install.packages("gbm", dependencies = TRUE, verbose = TRUE,
                 repos = "https://mirrors.tuna.tsinghua.edu.cn/CRAN/")
```

```
## Installing package into '/Users/mei/Library/R/3.6/library'
## (as 'lib' is unspecified)
```

```
##
## The downloaded binary packages are in
##   /var/folders/hm/c3_fjypn62v5xh5b5ygv267m0000gn/T//RtmpbIBqAc/downloaded_packages
```

```
library(gbm)
```

```
## Loaded gbm 2.1.5
```

Dataset

We'll be employing the enhancer prediction dataset to exemplify an application of gradient boosting algorithm. The details for the dataset could be reached here.

```
epdata <- readRDS("../Machine_Learning_Deep_Learning/data/ep_data.rds")
rownames(epdata) <- c()
```

```r
set.seed(001)
data_partition <- sample(2, nrow(epdata), replace = TRUE, prob = c(0.8,0.2))
train <- epdata[data_partition==1,]
test <- epdata[data_partition==2,]
```