GBM (BOOSTED MODELS) TUNING PARAMETERS

Deepanshu Bhalla Add Comment
Machine Learning, R, R Programming

In Stochastic Gradient Boosting Tree models, we need to fine tune several parameters such as n.trees, interaction.depth, shrinkage and n.minobsinnode (R gbm package terms).

Check out : Boosting Tree Explained

The detailed explanation is as follows -

Gradient Boosting Tree Fine Tuning

1. n.trees – Number of trees (the number of gradient boosting iteration) i.e. N. Increasing N reduces the error on training set, but setting it too high may lead to

POPULAR POSTS

- **3 Ways to extract unique values from a range in Excel** Scenario Suppose you have a list of customer names. The list has some duplicate values. You wish to extract unique values from it. Sam...
- **SAS Interview Questions and Answers** The following is a list of frequently asked questions about basic, intermediate and advanced concepts of SAS. 1. Difference between ...
- **Analytics Companies Using SAS in India** SAS (Statistical analysis system), the world's fastest and powerful software for data management, data mining, report

over-fitting.

2. interaction.depth (Maximum nodes per tree) - number of splits it has to perform on a tree (starting from a single node).

> *More than two nodes are required to detect interactions and the default six - node tree appears to do an excellent job*

interaction.depth = 1 : additive model, interaction.depth = 2 : two-way interactions, etc.

As each split increases the total number of nodes by 3 and number of terminal nodes by 2, the total number of nodes in the tree will be 3∗N+1 and the number of terminal nodes 2∗N+1

> ***Salford Default Setting :*** *6 - node tree appears to do an excellent job*

3. Shrinkage (Learning Rate) – It is considered as a learning rate.

Shrinkage is commonly used in ridge regression where it reduces regression coefficients to zero and, thus, reduces the impact of potentially unstable regression coefficients.

In the context of GBMs, shrinkage is used for reducing, or shrinking, the impact of each additional fitted base-learner (tree). It reduces the size of incremental steps and thus penalizes the importance of each consecutive iteration. The intuition behind this technique is that it is

better to improve a model by taking many small steps than by taking fewer large steps. If one of the boosting iterations turns out to be erroneous, its negative impact can be easily corrected in subsequent steps.

> ***Salford Default value =*** *max(0.01, 0.1*min(1, nl/10000))*
> *where nl = number of LEARN records.*

This default uses very slow learn rates for small data sets and **uses 0.1 for all data sets with more than 10,000 records.**

> *High learn rates and especially values close to 1.0 typically result in overfit models with poor performance. Values much smaller than .01 significantly slow down the learning process and might be reserved for overnight runs.*

> *Use a **small shrinkage** (slow learn rate) when growing many trees.*

One typically chooses the shrinkage parameter beforehand and varies the number of iterations (trees) N with respect to the chosen shrinkage.

4. n.minobsinnode - the minimum number of observations in trees' terminal nodes. Set n.minobsinnode = 10. When working with small training samples it may be vital to lower this setting to five or even three.

IMPORTANT LINKS

- 4 Simple VBA Lessons
- Actuarial Science
- Advanced Excel
- Business Analytics
- Charts
- Decision Tree
- Excel
- Excel Macros
- Functions
- Infographics
- Linear Regression
- Machine Learning
- Mathematics Using Excel
- Outlier
- Powerpoint
- R
- R Programming
- random forest
- Resumes
- SAS
- SAS Base Certification Questions and Answers
- SAS For Beginners
- SAS Interview Questions
- SPSS
- SQL
- Statistics
- Statistics Using Excel
- Text Analytics
- Text Mining
- Time Series
- Time Series Forecasting
- VBA
- Web Analytics

5. bag.fraction (Subsampling fraction) - the fraction of the training set observations randomly selected to propose the next tree in the expansion. By default, it is 0.5. That is half of the training sample at each iteration. You can use fraction greater than 0.5 if training sample is small.

6. train.fraction - The first train.fraction * nrows(data) observations are used to fit the gbm and the remainder are used for computing out-of-sample estimates of the loss function (like out of bag error in random forest). By default, it is 1.

Important Note I : You can ignore step 5 and 6 to fine tune the GBM model.

Important Note II : Small shrinkage generally gives a better result, but at the expense of more iterations (number of trees) required.

Examples -

> *distribution = "bernoulli", n.trees = 1000, interaction.depth =6, shrinkage = 0.1 and n.minobsinnode = 10*
> *distribution = "bernoulli", n.trees = 3000, interaction.depth =6, shrinkage = 0.01 and n.minobsinnode = 10*

R Code : TreeNet (Gradient Boosting Tree)

1. Model Build

> *gbm1 = gbm(gb ~ ., data = german_data, distribution = "bernoulli", bag.fraction = 0.5, n.trees = 1000, interaction.depth =6, shrinkage = 0.1, n.minobsinnode = 10)*

Important Point : Make sure the dependent variable is not defined as a factor if the dependent variable is binary. If it is a factor, multinomial is assumed. If the response has only 2 unique values (0/1), **bernoulli** is assumed; otherwise, if the response has class "Surv", **coxph** is assumed; otherwise, **gaussian** is assumed.

2. Variable Importance

> *importance = summary.gbm(gbm1, plotit=TRUE)*

It's Your Turn!

If you want me to keep writing this site, please post your feedback in the comment box below. While I love having friends who agree, I only learn from those who don't!

RELATED POSTS:

- Validate Cluster Analysis
- Cluster Analysis with R
- Ensemble Learning : Stacking / Blending
- GBM (Boosted Models) Tuning Parameters
- Dimensionality Reduction with R
- Predict Functions in R
- Weighting in Conditional Tree and SVM

- R : Apply Function on Rows
- Split a data frame
- R : Convert Data from Wide to Long Format

0 RESPONSE TO "GBM (BOOSTED MODELS) TUNING PARAMETERS"

← PREV NEXT →