technische universität dortmund

Data Science Center

SFB 876 Providing Information by Resource-Constrained Data Analysis

Variable Importance Measures for Functional Gradient Descent Boosting Algorithm

Master Thesis at the University of Göttingen
Zeyu Ding, Faculty of Statistics

Mathematical Statistics with Applications in Biometrics | 17.06.2021

Introduction

Challenges in statistics as variables increase
High-dimensional Data

- Number of variables p is much higher than the number of samples n

Introduction

Challenges in statistics as variables increase

High-dimensional Data

- Number of variables *p* is much higher than the number of samples *n*

Overly complex models

- High performance, low interpretability

Introduction

Challenges in statistics as variables increase
High-dimensional Data

- Number of variables p is much higher than the number of samples n

Overly complex models

- High performance, low interpretability

Overfitting

- Model performs well in the training phase and the prediction accuracy is however weak

Introduction

Solutions to these problems
Model Selection

- AIC/BIC based model selection methods

Introduction

Solutions to these problems

Model Selection

- AIC/BIC based model selection methods

Sparse Regression

- Lasso and Ridge based regression methods

Introduction

Solutions to these problems

Model Selection

- AIC/BIC based model selection methods

Sparse Regression

- Lasso and Ridge based regression methods

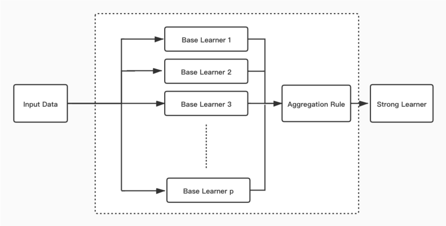
Variable Importance Measures

- Usually used in ensemble algorithm, i.e., Random Forest, Gradient Boosting

Methodology

Functional Gradient Descent Boosting Algorithm Statistical Boosting

- Gradient boosting algorithm can be viewed as a statistical model of the generalized additive model class.



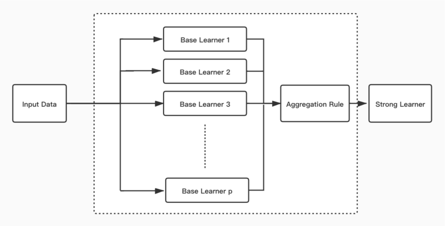
$$f(\boldsymbol{x}) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

Methodology

Functional Gradient Descent Boosting Algorithm

Statistical Boosting

- Gradient boosting algorithm can be viewed as a statistical model of the generalized additive model class.



Component-wise gradient boosting

- Only the best performed base-learner is chosen into the model in every iteration.

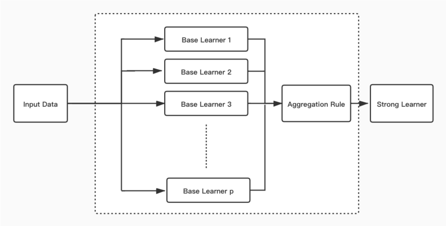
$$f(\boldsymbol{x}) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

Methodology

Functional Gradient Descent Boosting Algorithm

Statistical Boosting

- Gradient boosting algorithm can be viewed as a statistical model of the generalized additive model class.



Component-wise gradient boosting

- Only the best performed base-learner is chosen into the model in every iteration.

$$f(\boldsymbol{x}) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

Regressed iteratively

- The model complexity is controlled by the number of iteration.

Methodology

Component-Wise Gradient Boosting Algorithm

1. Set the initial iteration m=0. Given the initialized value of $\hat{f}^{[0]}(\cdots)$, common choices are

$$\hat{f}^{[0]} \equiv \arg\min_{c} \frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, c)$$

or $\hat{f}^{[0]} \equiv 0$.

2. For $m = 1$ to m_{stop}

(a). Obtain the negative gradient vector at the previous iteration $m-1$

$$\boldsymbol{g}^{[m]} = g_i^{[m]} = \left(\left[\frac{\partial\rho(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)}\right)_{(i=1,\dots,n)}$$

(b). Fit the negative gradient vector $\boldsymbol{g}^{[m]}$ to the input variables $\boldsymbol{x}$ by the base-learner procedure.

$$(\boldsymbol{x_1}, \boldsymbol{g}^{[m]}), (\boldsymbol{x_2}, \boldsymbol{g}^{[m]}), \dots, (\boldsymbol{x_p}, \boldsymbol{g}^{[m]}) \xrightarrow{procedure} \hat{h}_i^m(x_i)_{i=1,\dots,p}$$

Methodology

Component-Wise Gradient Boosting Algorithm

(c). Select the component j^* that best fits the negative gradient vector $\boldsymbol{g}_m$

$$j^* = \arg\min_{1 \le j \le p} \sum_{i=1}^{n} (g_i^{[m]} - \hat{h}_j^{[m]}(x_j))^2$$

(d). The model $\hat{f}^{[m]}(\cdot)$ is updated by

$$\hat{f}^{[m]}(\cdot) = \hat{f}^{[m-1]}(\cdot) + \theta \cdot \hat{h}_{j^*}^{[m]}(x_{j^*})$$

where θ denotes a step length.

3. After m_{stop} iterations, the model is obtained by

$$\hat{f}(\cdot) = \hat{f}^{[m]}(\cdot)$$

Methodology

Variable Selection Criterion
Selection Frequency

- Currently implemented in the algorithm

Methodology

Variable Selection Criterion

Empirical Risk Reduction

- The empirical risk reduction from each base learner in every iteration is calculated

$$VI^{[j]}_{risk}(\hat{h}_j(\cdot)) = \sum_{m:j^*_m} (\rho(y, \hat{f}^{[m]}) - \rho(y, \hat{f}^{[m-1]}))$$

l_2-norm Contribution

- The l_2-norm of every base-learner is used as a measure of the variable importance

$$\|\hat{h}_j(\cdot)\| = \sqrt{\sum_{i=1}^{n} (\hat{h}_j^{[m_{stop}]}(x_{ij}))^2}$$

$$VI^{[j]}_{norm}(\hat{h}_j(\cdot)) = \frac{\|\hat{h}_j(\cdot)\|}{\sum_{j=1}^{p} \|\hat{h}_j(\cdot)\|}$$

Simulation Data

Linear Model

- Simple Linear Model as base learners

Non-linear Model

- B-spline as base learners

Table 3: Sample size n and number of iterations m_{stop}

Sample size n	number of iterations m_{stop}
$n = 50$	$m_{stop} = 40$
	$m_{stop} = m_{stop}^{[cvrisk]}$
	$m_{stop} = 500$
$n = 200$	$m_{stop} = 40$
	$m_{stop} = m_{stop}^{[cvrisk]}$
	$m_{stop} = 500$
$n = 1000$	$m_{stop} = 40$
	$m_{stop} = m_{stop}^{[cvrisk]}$
	$m_{stop} = 500$
$n = 2000$	$m_{stop} = 40$
	$m_{stop} = m_{stop}^{[cvrisk]}$
	$m_{stop} = 500$

Simulation Data

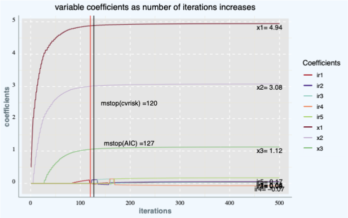
High-Dimensional Data

Table 5: Simulation design for high-dimensional scenario

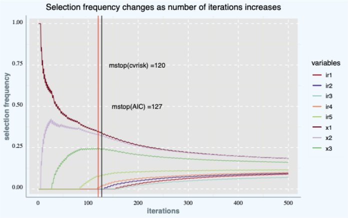
Sample size n	number of influential variables k	number of non-influential variables j	number of variables
$n = 50$	$k = 2$	$j = 100$	$p = 102$
$n = 100$	$k = 3$	$j = 500$	$p = 503$
$n = 500$	$k = 8$	$j = 1000$	$p = 1008$

Main Result

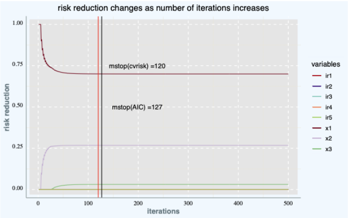
Linear Model



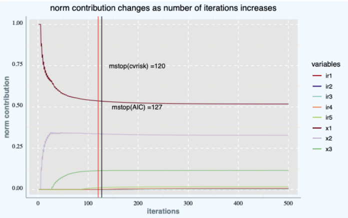
(a) change in variable coefficients



(b) change in selection frequency



(c) change in risk reduction



(d) change in norm contribution

Main Result

High-dimensional Data

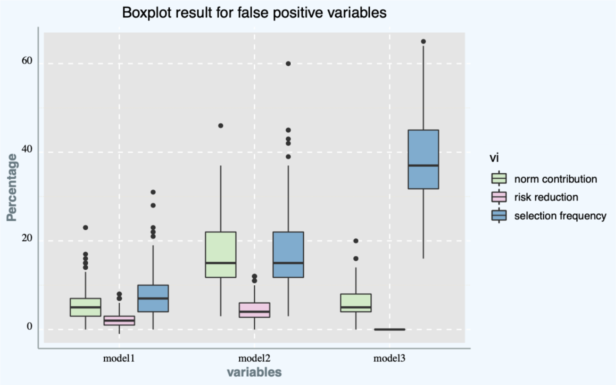


Figure 31: Number of false positive variables in high-dimensional scenario

Conclusion

Overfitting

- The variable importance measures based on empirical risk reduction and norm contribution in the FGDB algorithm are stable in resisting overfitting problem.

Conclusion

Overfitting

- The variable importance measures based on empirical risk reduction and norm contribution in the FGDB algorithm are stable in resisting overfitting problem.

High-Dimensional Data

- In high-dimensional data scenario, VI risk and VI norm also have a good ability to distinguish and rank variables by their importance.

Conclusion

Overfitting

- The variable importance measures based on empirical risk reduction and norm contribution in the FGDB algorithm are stable in resisting overfitting problem.

High-Dimensional Data

- In high-dimensional data scenario, VI risk and VI norm also have a good ability to distinguish and rank variables by their importance.

Multicollinearity

- They are also stable when existing multicollinear variables.

Outlook

More Complex Data

- In future research, more complex data scenarios need to be considered.

More Real-World Applications

- More real-world data needs to be validated, especially in the field of biometrics and bioinformatics when the dimensionality of the data is very high.

Outlook

More Complex Data

- In future research, more complex data scenarios need to be considered.

More Real-World Applications

- More real-world data needs to be validated, especially in the field of biometrics and bioinformatics when the dimensionality of the data is very high.

Thanks for your attention!

Reference

- Bühlmann, P., Gertheiss, J., Hieke, S., Kneib, T., Ma, S., Schumacher M., Ziegler, A. (2014). Discussion of the evolution of boosting algorithms and extending statistical boosting. Methods of information in medicine, 53(06), 436-445.
- Bühlmann, Peter, and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting Statistical science 22.4 (2007): 477-505.
- B. Hofner, L. Boccuto, and M. Goeker. Controlling false discoveries in high- dimensional situations: boosting with stability selection. BMC Bioinformatics, 144(16), 2015.
- Torsten Hothorn, Peter Bühlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. mboost: Model-Based Boosting, 2020. R package version 2.9-2.
- Mayr, Andreas, et al. The evolution of boosting algorithms-from machine learning to statistical modelling. arXiv preprint arXiv:1403.1452 (2014).

Reference

- Mayr, A., Hofner, B., Schmid, M. (2012). The importance of knowing when to stop. Methods of Information in Medicine, 51(02), 178-186.
- Hofner, B., Hothorn, T., Kneib, T., Schmid, M. (2011). A framework for unbiased model selection based on boosting. Journal of Computational and Graphical Statistics, 20(4), 956-971.
- Bühlmann, P., Yu, B. (2003). Boosting with the l_2 loss: regression and classification. Journal of the American Statistical Association, 98(462), 324-339.

Appendix

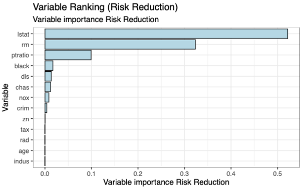
Boston House Price Data

Table 6: Boston Housing Dataset: variable explanation

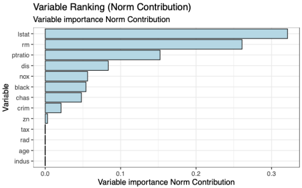
Variable abbreviation	Variable explanation
`crim`	per capita crime rate by town
`zn`	proportion of residential land zoned for lots over 25,000 sq.ft
`indus`	proportion of non-retail business acres per town
`chas`	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
`nox`	nitrogen oxides concentration (parts per 10 million)
`rm`	average number of rooms per dwelling
`age`	proportion of owner-occupied units built prior to 1940
`dis`	weighted mean of distances to five Boston employment centres
`rad`	index of accessibility to radial highways
`tax`	full-value property-tax rate per $10,000
`ptratio`	pupil-teacher ratio by town
`black`	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
`lstat`	lower status of the population (percent)
`medv`	median value of owner-occupied homes in $1000s

Appendix

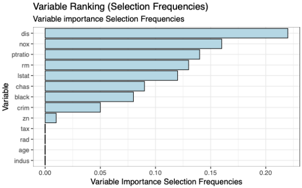
Boston House Price Data



(a) Variable importance by VI_{risk}



(b) Variable importance by VI_{norm}



(c) Variable importance by Selection frequency

Figure 36: Relative importance result of FGDB algorithm

Appendix

Boston House Price Data

Table 7: Boston Housing Dataset: Measures of Variable Importance

Variable	bagging		randomForest		gbm	VI_{risk}	VI_{norm}	SeleFreq
	IncMSE	IncNodePurity	IncMSE	IncNodePurity				
crim	0.156	0.038	0.128	0.052	0.034	0.004	0.021	0.050
zn	0.038	0.001	0.031	0.005	0.000	0.000	0.003	0.010
indus	0.118	0.006	0.091	0.051	0.000	0.000	0.000	0.000
chas	0.002	0.001	0.020	0.003	0.008	0.012	0.048	0.090
nox	0.236	0.027	0.176	0.092	0.042	0.008	0.056	0.160
rm	0.641	0.443	0.320	0.282	0.389	0.323	0.261	0.130
age	0.175	0.012	0.094	0.022	0.002	0.000	0.000	0.000
dis	0.307	0.065	0.158	0.064	0.047	0.014	0.084	0.220
rad	0.501	0.003	0.046	0.006	0.003	0.000	0.000	0.000
tax	0.155	0.014	0.089	0.018	0.010	0.000	0.000	0.000
ptratio	0.187	0.015	0.133	0.033	0.028	0.099	0.152	0.140
black	0.100	0.011	0.047	0.013	0.004	0.017	0.054	0.08
lstat	0.374	0.364	0.320	0.358	0.433	0.522	0.321	0.12