MCW Biostatistics Technical Report 71:
Novel pediatric height outlier detection methodology for electronic health records via machine learning with monotonic Bayesian additive regression trees

RA Sparapani

October 13, 2021

1 Introduction

Our novel outlier detection methodology relies mainly on nonparametric machine learning via Bayesian Additive Regression Trees (BART) Chipman et al. (2010); Sparapani et al. (2021); speci-cally, an extension known as Monotonic BART or MBART Chipman et al. (2016). BART/MBART is an ensemble model of binary regression trees. Ensembles are the best-known predictive models in out-of-sample performance as assessed by an independent validation data set Baldi and Brunak (2001); Kuhn and Johnson (2013), i.e., ensembles will not over- t to the training data at the expense of predictive performance on the unseen validation data (thus providing robustness to outliers whether a particular patient has an outlier is NOT needed to be known for the training cohort

Figure 1: Friedman's partial dependence function when the strength of the relationship between age and weight is mistakenly ignored. In this gure, males are blue dots/lines and females are red dots/lines with 95% credible intervals around the marginal e ects.

2.3 Marginal e ects and dependent variables

Friedman's partial dependence function works well when there are only weak relationships between the covariates. However, when there are strong relationships, such as between age and weight here; then, we need to extend this approach which we illustrate via our example.

We adopt the following notation for our variables: a for age, g for gender, r for race, w for weight and y for height. If we are interested in the marginal e ects due to age and gender, then we could consider the FPDF $f_y(a; g) = E[y|do(a; g)]$, i.e., the expected height on a grid of speci ed values for age and gender. However, this will y.cgl tuns/F1gleci ede1(mult6 0330)-3e3(ho)28iedman's e adopt thende41femad

Figure 2: Friedman's partial dependence function when the strength of the relationship between age and weight is properly accounted for. In this gure, males are blue dots/lines and females are red dots/lines with 95% credible intervals around the marginal e ects.

Figure 3: Friedman's partial dependence function when the strength of the relationship between age and weight is mistakenly ignored. In this gure, males are blue dots/lines and females are red dots/lines with 95% credible intervals around the marginal e ects.

Figure 4: Friedman's partial dependence function when the strength of the relationship between age and weight is properly accounted for. In this gure, males are blue dots/lines and females are red dots/lines with 95% credible intervals around the marginal e ects.

References

P Baldi and S Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, MA, 2nd edition, 2001.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1):266{298, 2010. doi: 10.1214/09-aoas285.

Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. High-dimensional nonparametric monotone function estimation using BART. *arXiv preprint arXiv:1612.01619*, 2016.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189{1232, 2001. URL http://www.jstor.org/stable/2699986 .

M Kuhn and K Johnson. *Applied Predictive Modeling*. Springer-Verlag, New York, NY, 2013. doi: 10.1007/978-1-4614-6849-3.

Hang TT Phan, Florina Borca, David Cable, James Batchelor, Justin H Davies, and Sarah Ennis. Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. *Scienti c reports*, 10(1):1{9, 2020.

R Sparapani, C Spanbauerni,Yu(2763Abauerni,)-28(nca(.-28(ncav)5auerni,Mosting)-30uerni,Le75(L)52(8(Y(bauerni