The Additive Tree

Gilmer Valdes PhD, DABR

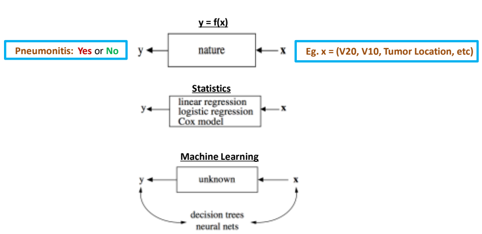
Outline

1. Machine Learning algorithms.
2. Few thoughts on Interpretability of Machine Learning algorithms.
3. The Additive Tree Framework.

Statistical Modeling: The Two Cultures

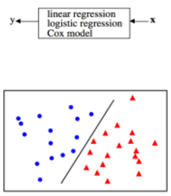
Leo Breiman. Statistical Science. 2001, Vol. 16, No. 3, 199-231.

https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726



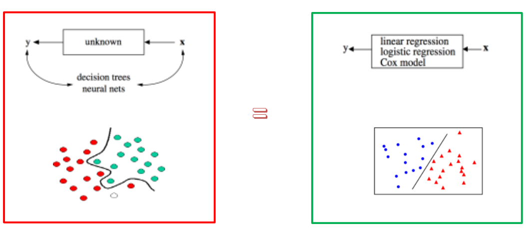
Before Machine Learning

Y = f(x)



Machine Learning vs Statistics (Not really)

Y = f(x)



The reasons for the hype

Press release: " Artificial Intelligence classifies Husky vs Wolfs with super human performance……"

https://www.slideshare.net/0xdata/explaining-blackbox-machine-learning-predictions

Reality:Accuracy is not Intelligence

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

https://www.slideshare.net/0xdata/explaining-blackbox-machine-learning-predictions

https://arxiv.org/pdf/1602.04938.pdf

The reasons for the hype

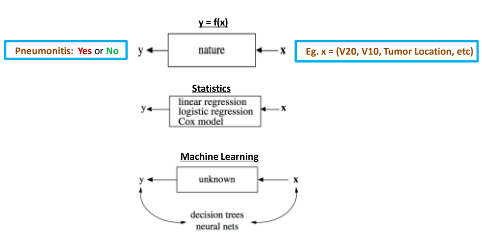
Press release: " Artificial Intelligence classifies Husky vs Wolfs with super human performance……."

https://www.slideshare.net/0xdata/explaining-blackbox-machine-learning-predictions

Statistical Modeling: The Two Cultures

Leo Breiman. Statistical Science. 2001, Vol. 16, No. 3, 199-231.

https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726



Reasons for interpretability

The need for interpretable models in medicine rises from practical and theoretical reasons:

1. *Acceptance.*

2. *Known limitations of observational training data (cofounders, noise, bias, etc)*

3. *Mismatch between the objective function maximized by machine learning algorithms and the ethical needs in medicine.*

1. Acceptance

TABLE V

MEANS RATINGS AND (STANDARD DEVIATIONS) FOR DEMAND STATEMENTS*

	Physicians $n = 146$	Nonphysicians $n = 29$
D.1 *Should* be able to explain their diagnostic and treatment decisions to physician users	1.42 (0.80)	1.78 (0.42)
D2. *Should* be portable and flexible so that MD can access them at any time and place	1.4 (0.81)	1.52 (0.51)
D3. *Should* display an understanding of their own medical knowledge	0.99 (0.94)	1.48 (0.80)
D4. *Should* improve the cost efficiency of tests and therapies	0.85 (0.99)	1.11 (1.58)
D5. *Should* automatically learn new information when interacting with medical experts	0.84 (1.02)	1.41 (0.75)
D6. *Should* display common sense	0.75 (1.20)	1.11 (0.97)
D7. *Should* simulate physicians' thought processes	0.64 (1.16)	0.93 (1.07)
D8. *Should* not reduce the need for specialists	0.46 (1.18)	0.70 (1.07)
D9. *Should* demand little effort from physician to learn or use	0.35 (1.20)	1.19 (0.92)
D10. *Should* respond to voice command and not require typing	0.26 (1.23)	0.56 (1.05)
D11. *Should* not reduce the need for paraprofessionals	0.26 (1.06)	0.85 (1.03)
D12. *Should* significantly reduce amount of technical knowledge physician must learn and remember	−0.08 (1.34)	0.00 (1.49)
D13. *Should* never make an error in treatment planning	−0.25 (1.33)	−0.22 (1.34)
D14. *Should* never make an incorrect diagnosis	−0.45 (1.31)	−0.26 (1.46)
D15. *Should* become the standard for acceptable medical practice	−0.80 (1.13)	0.00 (1.07)
Total scale	0.44	0.81

* Scale: −2.0 = unimportant to 2.0 = important.

Teach, R.L. and E.H. Shortliffe, *An analysis of physician attitudes regarding computer-based clinical consultation systems.* Computers and Biomedical Research, 1981. **14**(6): p. 542-558.

2. Limitations of observational data

Example: *Predicting Risk of dying of Pneumonia for In-hospital patients*

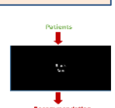
Most accurate model trained: Multi-purpose neural net….

Rule Based Model

Asthmatic → **Lower Risk**

- **Harmful to patients**
- **High Risk of Liability**

=



https://www.ncbi.nlm.nih.gov/pubmed/9040894

2. Limitations of observational data

Example: *Predicting Risk of stroke for Emergency Department patients*

Table 1—Predicting and Mispredicting

	Stroke	30-day mortality
Prior stroke	0.302	0.041
	(0.012)	(0.014)
Prior accidental injury	0.285	0.007
	(0.095)	(0.101)
Abnormal breast finding	0.224	0.162
	(0.092)	(0.110)
Cardiovascular disease history	0.218	−0.017
	(0.029)	(0.034)
Colon cancer screening	0.242	−0.475
	(0.178)	(0.222)
Acute sinusitis	0.220	0.056
	(0.155)	(0.166)

Notes: Logistic regression on demographics and prior diagnoses in EHR data. Sample: 177,825 ED visits in 2010–2012 to a large academic hospital.

"Does Machine Learning Automate Moral Hazard and Error?" .American Economic Review: Papers & Proceedings 2017, 107(5): 476–480

2. Limitations of observational data

Bias in Medicine

1. Psychologically salient diseases are over-diagnosed.

"Cognitive Biases and Heuristics in Medical Decision Making: A Critical Review Using a Systematic Search Strategy." Medical Decision Making **35 (4): 539–57.**

2. Physicians are 40 percent less likely to refer female or black patients for catherization.

"Effect of Race and Sex on Physicians' Recommendations for Cardiac Catheterization." New England Journal of Medicine 340: 618–26.

3. Minorities receive less aggressive cancer treatment.

"Racial Differences in the Treatment of Early-Stage Lung Cancer." New England Journal of Medicine **341: 1198–1205.**

3. Mismatch between the objective function maximized by machine learning algorithms and the ethical needs in medicine.

Machine Learning

$$\mathrm{E}_{y,x}(1(Y = Y_{ml})|\boldsymbol{x}) \geq \mathrm{E}_{y,x}\big(1\big(Y = Y_{ph}\big)\big|\boldsymbol{x}\big)$$

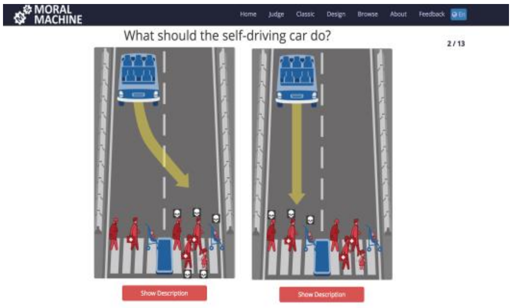
Wish to have:

$$\mathrm{E}_{y}(1(Y = Y_{ml})|\boldsymbol{x}) \geq \mathrm{E}_{y}\big(1\big(Y = Y_{ph}\big)|\boldsymbol{x}\big)\ \forall\ \boldsymbol{x}\ \in \boldsymbol{P}$$

More realistic and also impossible.

$$\mathrm{E}_{y,x}(1(Y = Y_{ml})|\boldsymbol{x}) \geq \mathrm{E}_{y,x}(1(Y = Y_{h})|\boldsymbol{x})\ \forall\ \boldsymbol{x}\ \in \boldsymbol{P_s} \subseteq\ \boldsymbol{P}$$

3. Moral Problem



http://moralmachine.mit.edu/

Reasons for interpretability

The need for interpretable models in medicine rises from practical and theoretical reasons:

1. Acceptance.

2. Known limitations of observational training data (cofounders, noise, bias, etc)

3. Mismatch between the objective function maximized by machine learning algorithms and the ethical needs in medicine.

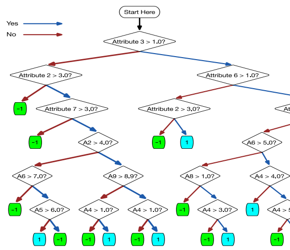
Outline

1. Machine Learning algorithms.

2. Few thoughts on Interpretability of Machine Learning algorithms.

3. The Additive Tree Framework.

Decision Trees are globally interpretable



	Attribute	Value
1	Clump Thickness	1 – 10
2	Uniformity of Cell Size	1 – 10
3	Uniformity of Cell Shape	1 – 10
4	Marginal Adhesion	1 – 10
5	Single Epithelial Cell Size	1 – 10
6	Bare Nuclei	1 – 10
7	Bland Chromatin	1 – 10
8	Normal Nucleoli	1 – 10
9	Mitoses	1 – 10
	Diagnosis (Output)	–1 benign 1 malign

Example Interpretable Rules Induced by MediBoost:

A3 Uniformity of Cell Shape $\leq$ 1.0 $\wedge$ A2 Uniformity of Cell Size > 3.0 $\wedge$ A7 Bland Chromatin $\leq$ 3.0 $\Rightarrow$ predict benign

A3 Uniformity of Cell Shape > 1.0 $\wedge$ A6 Bare Nuclei $\leq$ 1.0 $\wedge$ A2 Uniformity of Cell Size > 3.0 $\Rightarrow$ predict malignant

Decision Tree Accuracy

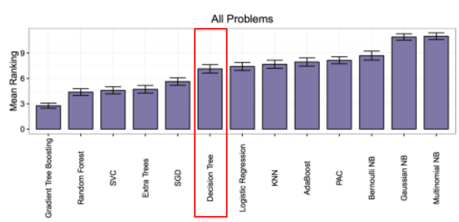
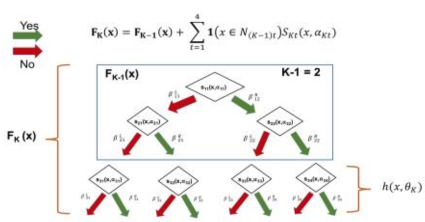


Fig. 1. Average ranking of the ML algorithms over all datasets. Error bars indicate the 95% confidence interval.

" Data-driven Advice for Applying Machine Learning to Bioinformatics Problems." Randal S. Olson et al.
https://arxiv.org/pdf/1708.05070.pdf

The Additive Tree

Single Trees that offer a Continuum between
CART and Gradient Boosting



Example of a tree built using MediBoost and its full boosting interpretation. K-1 assumed to be 2. $S_{Kt}(x_i, a_{Kt})$ assumed to be stumps to obtain same architecture as CART. a_{Kt} in represents the index of the feature where the split will be made, the cutoff value and the constant β that the stumps predict. The Function $\mathbf{F_K(x)}$ at each terminal node is given by the sum of the βs.

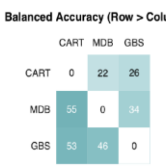
Experiments

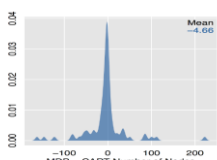
Statistical characteristics of the 95 Penn ML Benchmarks used in the experimental validation of TSB

	Mean	Std deviation
# Instances	1713	2900
# Attributes	36	111

The Additive Tree vs CART

Accuracy comparable to Gradient Boosting while building smaller trees than CART





Number of wining vs losing of all algorithms in 95 classification datasets. (Ties have been omitted). For each row, the number of times that each algorithm wins is represented. MediBoost (MDB) wins 55 vs 22 compare to CART ($p= 3.0 \times 10^{-7}$) and 34 vs 46 compare to Gradient Boosting with Stumps (GBS), not statistically significantly different

Number of nodes of trees built with MediBoost minus those of CART for the same problem (95 total different problems). The distribution is slightly tilted towards the left (smaller number of nodes for MediBoost).

The Additive Tree with Linear Models in the Nodes: (LIMB)

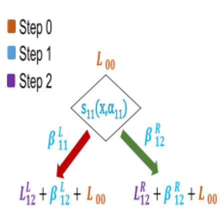


Illustration of a simple tree built with Linear MediBoost (LiMB). The Linear Model $\boldsymbol{L}_{00}$ is the first to be built. Then, the constant terms $\beta^L_{12}, \beta^R_{12}$ are added to each node defined by $s_{11}(x,\alpha_{11})$. Finally, new linear models $\boldsymbol{L}^L_{12}, \boldsymbol{L}^R_{12}$ are added at each node. Further partitions/linear models can be added to get a deeper tree.

LiMB	LR with Lasso	CART	Logit Boost	Random Forest
0.003	0.031	0.000	0.000	0.000
0.124	0.439	0.112	0.104	0.114
0.118	0.227	0.124	0.136	0.107
0.168	0.229	0.170	0.144	0.139
0.211	0.238	0.240	0.189	0.188
0.031	0.034	0.047	0.043	0.025
0.176	0.172	0.240	0.171	0.186
0.245	0.250	0.356	0.299	0.313
0.455	0.576	0.415	0.352	0.435
0.226	0.236	0.273	0.255	0.213
0.167	0.183	0.184	0.174	0.183
0.152	0.210	0.199	0.120	0.118
0.244	0.249	0.268	0.269	0.257
0.241	0.244	0.292	0.272	0.236
0.358	0.443	0.464	0.474	0.470

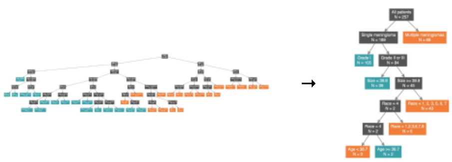
Comparison of LiMB vs different algorithms in 15 classification problems. 5 fold cross-validated balanced classification error shown. LiMB consistently wins over LR and CART, it wins 8 times compare to Gradient Boosting (with depth of trees optimized), 5 times compare to Random Forest.

The Additive Tree in Radiation Oncology

rtemis::s.MDB

Prune: `prune.mdb()`
- Remove empty leaves
- Remove siblings with same prediction
- Remove inner nodes with no siblings

Plot: `mplot3.mdb()`
- Interface to graphviz
- Option to prune internally



Conclusions

- The Additive Trees creates Decision Trees that would be drop-in replacements from currently used trees.
- The Additive Tree is the most accurate decision tree algorithm up to date that keeps the same structure as CART.

Acknowledgment

- Jerome H. Friedman (Stanford)
- Lyle Ungar (U Penn)
- Eric Eaton (U Penn)
- Marcio Lunas (U Penn)
- Efstathios Gennatas (UCSF)
- Timothy D Solberg (UCSF)

An R package will be released soon: gilmer.valdes@ucsf.edu

Available in Matlab and R

- Valdes et al "MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. " Nat Sci Rept. 2016

- Luna et al "Tree-Structured Boosting: Connection between Gradient Boosted Stumps and Full Decision Trees." NIPS 2017.

http://www.mediboostml.com/

https://egenn.github.io/rtemis/rtemis_mediboost_vignette.html