What Functions Does XGBoost Learn?

Dohyeong Ki & Adityanand Guntuboyina

Department of Statistics, University of California, Berkeley

XGBoost

Although XGBoost (eXtreme Gradient Boosting) has achieved remarkable empirical success, it has not been theoretically well-understood yet.

XGBoost fits a finite sum of regression trees to data.

XGBoost aims to (approximately) minimize least squares plus

$$\gamma \sum_k T_k + \alpha \sum_k \|w_k\|_1$$

squared L^2 norm is also common

where (1) T_k is the number of leaves in the kth regression tree,
(2) w_k is its vector of leaf weights.

XGBoost produces a discrete-valued tree fit, but it seems it also learns continuous functions quite effectively.

Q. What kinds of functions does XGBoost learn well?

Function Class Extending Finite Sums of Trees

Every regression tree can be expressed as a finite linear combination of

$$b^S_{\mathbf{p},\mathbf{q},\mathbf{t}}(x_1,\ldots,x_d) := \prod_{j\in S}\{\mathbf{1}(q_j=0)\mathbf{1}(p_j(x_j-t_j)\ge 0)+\mathbf{1}(q_j=1)\mathbf{1}(p_j(x_j-t_j)>0)\}$$

where (1) $S \subseteq \{1,\ldots,d\}$, (2) each $q_j \in \{0,1\}$, and (3) $p_j \in \{-1,1\}$.

- q_j determines whether the inequality is weak ($\ge$) or strict ($>$).
- p_j controls the direction of the inequality.
- t_j is a threshold associated with variable x_j.

Example: $d=2$ and $S=\{1,2\}$

(1) $(p_1,q_1)=(1,0)$ and $(p_2,q_2)=(-1,0)$,

$$b^S_{\mathbf{p},\mathbf{q},\mathbf{t}}(x_1,x_2)=\mathbf{1}(x_1\ge t_1)\cdot\mathbf{1}(x_2\le t_2)$$

(2) $(p_1,q_1)=(1,1)$ and $(p_2,q_2)=(-1,1)$,

$$b^S_{\mathbf{p},\mathbf{q},\mathbf{t}}(x_1,x_2)=\mathbf{1}(x_1> t_1)\cdot\mathbf{1}(x_2< t_2)$$

We consider infinite linear combinations of these basis functions with $|S|\le s$.

We define $\mathcal{F}^{d,s}_{\infty\text{-st}}$ as the collection of all functions $f:\mathbb{R}^d\to\mathbb{R}$ of the form:

$$f_{c,\{\nu^S_{\mathbf{p},\mathbf{q}}\}}(x_1,\ldots,x_d)=c+\sum_{S:0<|S|\le s}\sum_{\mathbf{p}\in\{-1,1\}^{|S|}}\sum_{\mathbf{q}\in\{0,1\}^{|S|}}\int_{\mathbb{R}^{|S|}} b^S_{\mathbf{p},\mathbf{q},\mathbf{t}}(x_1,\ldots,x_d)\,d\nu^S_{\mathbf{p},\mathbf{q}}(t_j, j\in S)$$

$\mathcal{F}^{d,s}_{\infty\text{-st}}$ is an infinite dimensional extension of the class $\mathcal{F}^{d,s}_{\text{st}}$ of finite sums of regression trees with maximum depth s.

→ consistent with XGBoost whose **max_depth** = 6 by default.

Complexity Extending XGBoost Penalty

Infinite linear combination representation for $f\in\mathcal{F}^{d,s}_{\infty\text{-st}}$ is not unique.

Define the complexity of $f\in\mathcal{F}^{d,s}_{\infty\text{-st}}$ as

$$V^1_{\infty\text{-xgb}}(f):=\inf\left\{\sum_{S:0<|S|\le s}\sum_{\mathbf{p}\in\{-1,1\}^{|S|}}\sum_{\mathbf{q}\in\{0,1\}^{|S|}}\|\nu^S_{\mathbf{p},\mathbf{q}}\|_{\text{TV}}: f_{c,\{\nu^S_{\mathbf{p},\mathbf{q}}\}}\equiv f\right\}$$

where $\|\nu\|_{\text{TV}}$ denotes the total variation of a signed measure ν.

Main Result 1:

If $f\in\mathcal{F}^{d,s}_{\text{st}}$, i.e., f is a finite sum of regression trees,

$$V^1_{\infty\text{-xgb}}(f)=\inf\left\{\sum_k\|w_k\|_1\right\}$$

where the infimum is over all representations of f in a finite sum of trees.

→ $V^1_{\infty\text{-xgb}}(\cdot)$ is an extension of the XGBoost penalty with $\gamma=0$

$\gamma=0$ means no penalty on numbers of leaves; the default choice by XGBoost

Relation to Hardy–Krause Variation

As the domain $\mathbb{R}^d$ is unbounded, we need to place an anchor for Hardy–Krause variation at infinity (either $-\infty$ or $+\infty$ for each coordinate).

Let $\mathbf{a}=(a_1,\ldots,a_d)\in\{-\infty,\infty\}^d$ denote the anchoring point.

For a function $f:\mathbb{R}^d\to\mathbb{R}$ and $S\subseteq\{1,\ldots,d\}$, define

$$f^S_{(a_j,j\in S^c)}(x_j,j\in S)=\lim_{(x_j,j\in S^c)\to(a_j,j\in S^c)} f(x_1,\ldots,x_d)\quad\text{for}\quad (x_j,j\in S)\in\mathbb{R}^{|S|}$$

Hardy–Krause variation of f anchored at $\mathbf{a}$ is defined by

$$\text{HK}_{\mathbf{a}}(f)=\sum_{\emptyset\ne S\subseteq\{1,\ldots,d\}}\text{Vit}(f^S_{(a_j,j\in S^c)}).$$

where $\text{Vit}(\cdot)$ denotes Vitali variation.

Hardy–Krause variation is asymmetric, whereas $V^1_{\infty\text{-xgb}}(\cdot)$ is symmetric.

Example: $d=s=2$ and $\mathbf{a}=(-\infty,-\infty)$

$$\text{HK}_{\mathbf{a}}\big(\mathbf{1}(\cdot_1\ge t_1,\cdot_2\ge t_2)\big)=1\ \text{ but }\ \text{HK}_{\mathbf{a}}\big(\mathbf{1}(\cdot_1< t_1,\cdot_2< t_2)\big)=3$$

$$V^1_{\infty\text{-xgb}}\big(\mathbf{1}(\cdot_1\ge t_1,\cdot_2\ge t_2)\big)=V^1_{\infty\text{-xgb}}\big(\mathbf{1}(\cdot_1< t_1,\cdot_2< t_2)\big)=1$$

In fact, $V^1_{\infty\text{-xgb}}(\cdot)$ is a symmetrized version of Hardy–Krause variation;

$V^1_{\infty\text{-xgb}}(\cdot)$ is the infimal convolution of $\text{HK}_{\mathbf{a}}(\cdot)$ over all anchors $\mathbf{a}\in\{-\infty,\infty\}^d$, when restricted to the subclass $\mathcal{F}^{d,s}_{\infty\text{-rst}}$ consisting of right-continuous functions

$$V^1_{\infty\text{-xgb}}(f)=\inf\left\{\sum_{\mathbf{a}\in\{-\infty,\infty\}^d}\text{HK}_{\mathbf{a}}(f_{\mathbf{a}}):\sum_{\mathbf{a}\in\{-\infty,\infty\}^d} f_{\mathbf{a}}\equiv f\ \text{ and }\ f_{\mathbf{a}}\in\mathcal{F}^{d,s}_{\infty\text{-rst}}\right\}$$

Least Squares Estimator (LSE)

A central object of interest is the following least squares estimator:

$$\underset{f}{\text{argmin}}\left\{\sum_{i=1}^n\big(y_i-f(\mathbf{x}^{(i)})\big)^2: f\in\mathcal{F}^{d,s}_{\infty\text{-st}}\ \text{ and }\ V^1_{\infty\text{-xgb}}(f)\le V\right\}.$$

Let $\mathcal{F}^{d,s}_{\text{rstm}}$ denote the collection of all finite linear combinations of $b^S_{\mathbf{p},\mathbf{q},\mathbf{t}}$ where

- $|S|\le s$ → depth is no larger than s
- $\mathbf{p}=\mathbf{1}-2\mathbf{q}$ → $b^S_{\mathbf{p},\mathbf{q},\mathbf{t}}$ are products only of $\mathbf{1}(x_j\ge t_j)$ and $\mathbf{1}(x_j<t_j)$

→ aligns with XGBoost's tree-splitting scheme where one branch corresponds to $\mathbf{1}(x_j\ge t_j)$ and the other to $\mathbf{1}(x_j<t_j)$

- t_j are midpoints between observed values of the jth covariate

→ aligns with XGBoost's split points for numerical variables

→ By default (**tree_method** = **auto**), XGBoost uses midpoints when datasets are small but switches to quantiles for larger datasets.

Main Result 2:

The least squares estimator $\hat{f}^{d,s}_{n,V}$ over all $f\in\mathcal{F}^{d,s}_{\text{rstm}}$ with $V^1_{\infty\text{-xgb}}(f)\le V$ is a least squares estimator over all $f\in\mathcal{F}^{d,s}_{\infty\text{-st}}$ with $V^1_{\infty\text{-xgb}}(f)\le V$.

XGBoost can be viewed as a greedy solver for the penalized version of this least squares problem over $\mathcal{F}^{d,s}_{\text{rstm}}$.

Theoretical Accuracy of LSE

Main Result 3:

Assume the following random design setting:

(1) $y_i=f^*(\mathbf{x}^{(i)})+\epsilon_i$ where $f^*\in\mathcal{F}^{d,s}_{\infty\text{-st}}$ and $\epsilon_i\overset{\text{i.i.d.}}{\sim}N(0,\sigma^2)$ — can be replaced by a more general assumption

(2) $\mathbf{x}^{(i)}\overset{\text{i.i.d.}}{\sim}p_0$ for some density p_0

(3) p_0 has compact support: there exist $M_1,\ldots,M_d>0$ such that

$$p_0(\mathbf{x})=0\ \text{ unless }\ \mathbf{x}\in\prod_{j=1}^d\left[-\frac{M_j}{2},\frac{M_j}{2}\right]$$

(4) p_0 is bounded above; $B:=M_1\cdots M_d\cdot\sup_{\mathbf{x}}p_0(\mathbf{x})<+\infty$.

If $V>V^1_{\infty\text{-xgb}}(f^*)$, then we have

$$\mathbb{E}\left[\int\big(\hat{f}^{d,s}_{n,V}(\mathbf{x})-f^*(\mathbf{x})\big)^2\cdot p_0(\mathbf{x})\,d\mathbf{x}\right]=O\big(n^{-2/3}(\log n)^{4(s-1)/3}\big).$$

constant factor depends on $B,d,V,$ and σ

It can also be proved that this rate is nearly minimax optimal.

Whether XGBoost itself achieves a similar nearly dimension-free rate of convergence is an open problem.

dohyeong_ki@berkeley.edu