What Functions Does XGBoost Learn?

Dohyeong Ki & Adityanand Guntuboyina

Department of Statistics, University of California, Berkeley

XGBoost

XGBoost (eXtreme Gradient Boosting) has achieved huge empirical success, but it is not well-understood theoretically.

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 tree,
(2) w_k is its vector of leaf node values.

Although computationally infeasible, the minimizer of this problem can be viewed as an idealized target of XGBoost.

Studying this idealized target is helpful in understanding XGBoost better and in answering questions like:

XGBoost produces a discrete-valued tree fit, yet it seems to learn continuous functions quite effectively. How so?

Q. What kinds of functions can XGBoost, in principle, learn efficiently?

Function Class Extending Finite Sums of Regression Trees

We restrict to regression trees whose splits are based on whether $\mathbf{1}(x_j \geq t_j)$ or $\mathbf{1}(x_j < t_j)$ (not $\mathbf{1}(x_j > t_j)$ or $\mathbf{1}(x_j \leq t_j)$).

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

$$b_{\mathbf{l},\mathbf{u}}^{L,U}(x_1, \dots, x_d) := \prod_{j \in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j \in U} \mathbf{1}(x_j < u_j)$$

where (1) $L, U \subseteq \{1, \dots, d\}$ (not necessarily disjoint) and
(2) each $l_j, u_j \in \mathbb{R}$.

Why? Every regression tree can be decomposed into paths from the root node to each leaf node.

For each path, L (resp., U) is the set of indices j for which the condition $\mathbf{1}(x_j \geq l_j)$ (resp., $\mathbf{1}(x_j < u_j)$) appearing on the path.

Example) $d = 2$, $L = \{1\}$, and $U = \{1,2\}$

$$b_{\mathbf{l},\mathbf{u}}^{L,U}(x_1, x_2) = \mathbf{1}(l_1 \leq x_1 < u_1) \cdot \mathbf{1}(x_2 < u_2)$$

We consider infinite linear combinations of these basis functions $b_{\mathbf{l},\mathbf{u}}^{L,U}$ with $|L| + |U| \leq s$ for some fixed s.

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

$$f_{c,\{\nu_{L,U}\}}(x_1, \dots, x_d) := c + \sum_{0<|L|+|U|\leq s} \int_{\mathbb{R}^{|L|+|U|}} b_{\mathbf{l},\mathbf{u}}^{L,U}(x_1, \dots, x_d)\, d\nu_{L,U}(\mathbf{l}, \mathbf{u})$$

where $\nu_{L,U}$ are finite signed (Borel) measures on $\mathbb{R}^{|L|+|U|}$.

$\mathcal{F}_{\infty-\mathrm{ST}}^{d,s}$ is an infinite dimensional extension of the class $\mathcal{F}_{\mathrm{ST}}^{d,s}$ 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 $f_{c,\{\nu_{L,U}\}}$ is not unique for each $f \in \mathcal{F}_{\infty-\mathrm{ST}}^{d,s}$.

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

$$V_{\infty-\mathrm{XGB}}^{d,s}(f) := \inf \left\{ \sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{\mathrm{TV}} : f_{c,\{\nu_{L,U}\}} \equiv f \right\}$$

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

Main Result 1:

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

$$V_{\infty-\mathrm{XGB}}^{d,s}(f) = \inf \left\{ \sum_k \|w_k\|_1 \right\} =: V_{\mathrm{XGB}}^{d,s}(f)$$

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

→ $V_{\infty-\mathrm{XGB}}^{d,s}(\cdot)$ is an extension of XGBoost penalty with $\gamma = 0$

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

Idealized Target of XGBoost

Let $\mathcal{F}_{\mathrm{STM}}^{d,s}$ be the sub-collection of $\mathcal{F}_{\mathrm{ST}}^{d,s}$ where $\nu_{L,U}$ are discrete and supported on the midpoints of observations.

By default, XGBoost uses such midpoints for tree splits when datasets are small but switches to quantiles for larger datasets.

Least squares estimator $\hat{f}_{n,V}^{d,s}$ over all $f \in \mathcal{F}_{\mathrm{STM}}^{d,s}$ with $V_{\mathrm{XGB}}^{d,s}(f) \leq V$ can be seen as an idealized target of XGBoost.

Main Result 2:

$\hat{f}_{n,V}^{d,s}$ is also a least squares estimator over all $f \in \mathcal{F}_{\infty-\mathrm{ST}}^{d,s}$ with $V_{\infty-\mathrm{XGB}}^{d,s}(f) \leq V$, i.e.,

$$\hat{f}_{n,V}^{d,s} \in \operatorname*{argmin}_f \left\{ \sum_{i=1}^n (y_i - f(\mathbf{x}^{(i)}))^2 : f \in \mathcal{F}_{\infty-\mathrm{ST}}^{d,s}, V_{\infty-\mathrm{XGB}}^{d,s}(f) \leq V \right\}$$

Accuracy of the Idealized Target

Main Result 3:

Assume the following random design setting: (can be more general)

(1) $y_i = f^*(\mathbf{x}^{(i)}) + \epsilon_i$ where $f^* \in \mathcal{F}_{\infty-\mathrm{ST}}^{d,s}$ and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$,

(2) $\mathbf{x}^{(i)} \overset{\text{i.i.d.}}{\sim} p_0$ for some density p_0 that has compact support and is bounded above, i.e., $\|p_0\|_\infty \leq B$.

If $V > V_{\infty-\mathrm{XGB}}^{d,s}(f^*)$, then

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

(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