Boosting Intro

Mathematics for Machine Learning — Lecture

Q: Guess the chest?

"Diagnosis"

Agenda:

1. Bagging Cont.
2. Full Search
3. Boosting Intuition
4. Boosting Math
5. Gradient Boosting

Problem Importance:

C.C.T. Algorithm

"Low complexity learning"

stump tree $\Rightarrow$ min $\left[\frac{\text{Loss}}{\text{complexity}}\right] \Rightarrow \lambda \cdot \text{complexity}$

stump tree $\Rightarrow$ min $\left[\frac{\text{Loss}}{\text{complexity}}\right] + \alpha\left(\begin{array}{c}\text{Minimum weight} \\ [\ldots]\end{array}\right)$

$\alpha \downarrow \uparrow \rightarrow$ underfit

$\alpha = 0 \downarrow \rightarrow$ overfit

Cons:

Decision

$\boxed{\text{RF}}$: $\boxed{\text{DT}} + \boxed{\text{Bagging}} + \boxed{\text{Rgs}}$

Bagging

Correlation $\rightarrow$ $\text{Variance}_{\text{Ensemble}} > \text{Variance}_{\text{Tree}}$?

The Lemma Inside me.

1. Problem Importance.
2. Why Decision Tree? $\Rightarrow \boxed{\text{RF}}$, $\boxed{\text{CCT}}$, $\boxed{\text{GBT}}$, ...

Decision Tree

Full Search

"Brute force"

is automatic hyper-parameter tuning

$\boxed{\text{RF}}$

$\text{Depth} \Rightarrow [1, 2, 3, \ldots, 7]$

$\text{Trees} \Rightarrow [1, 5, 7, 9, 12]$

"Brute force/Approach"

$\Rightarrow$ Exponential.

$\Rightarrow$ Problematic (optimal trees)

$\Rightarrow$ Combinations

Trees $\Rightarrow \boxed{12}$ RF models

$\boxed{\text{RF}_1} \rightarrow \boxed{\text{P}_1}$

$\boxed{\text{RF}_2} \rightarrow \boxed{\text{P}_2}$

$\boxed{\text{RF}_{12}} \rightarrow \boxed{\text{P}_{12}}$

$\boxed{\text{P}_{\text{Best}}}$ $\boxed{(2, 7)}$

$\rightarrow$ Parameter

$\rightarrow$ Problematic (optimal trees)

$\boxed{\ldots}$ $\boxed{\text{Best}}$

Randomsearch:

"Random Combination"

Non-right Distribution

$\boxed{\text{D.F.S.}}$

$\Rightarrow$ unbound $\Rightarrow$ $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

Fine tuning:

$\text{Max-depth} \rightarrow [\text{Low}, \text{Low}] \boxed{\text{Mid}} [\text{Mid}, \text{Mid}]$

$\boxed{\text{Low}}$ $\boxed{\text{Mid}}$ $\boxed{\text{Low}}$ $\boxed{\text{Low}}$ $\boxed{\text{Mid}}$

$\boxed{\text{Mid}}$ $\boxed{\text{Mid}}$ $\boxed{\text{Mid}}$ $\boxed{\text{Mid}}$ $\boxed{\text{Mid}}$

$\boxed{\text{Mid}}$ $\boxed{\text{Mid}}$ $\boxed{\text{Mid}}$ $\boxed{\text{Mid}}$ $\boxed{\text{Mid}}$

Boosting Intuition.

Bagging: Base learners [high var + low bias]

Boosting: Base learners [high bias + low var]

$\rightarrow$ Bias-variance

$\rightarrow$ Sequential Bagging

Base Learner: $\boxed{\text{D.T (low variance)}} \rightarrow \boxed{\text{GBDT}}$

Key Idea: "Address the bias"

Intuition: Train on Algorithm

Dataset: $\left\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\right\}_{m}$

Step-1

a) $\text{Train}_{\text{data}} \rightarrow$ $\boxed{\text{M}_1}$ $\left\{\begin{array}{l}\text{Mistakes} \\ \text{low var.}\end{array}\right\}$ $\left\{\begin{array}{l}\text{Fixed} \\ \text{Bias}\end{array}\right\}$

$\boxed{y_1 = h(x_1)}$

"Big mistake"

"Very inaccurate"

b) $\text{errors}^{(1)}$ / "rounds"

$\text{error}^{(1)} \rightarrow y^{(1)}$

$\text{error}^{(2)} \rightarrow y^{(2)}$, $h_2(x^{(1)})$

error $\rightarrow$ $y^{(1)}$, $h_2(x^{(1)})$

for each data point $\{x^{(1)}, y^{(1)}, \text{error}^{(1)}\}_{m}$

$\text{M}_1 \rightarrow h_1(x)$

Step-2.a

$\text{Error}_m(x) \Rightarrow$ mistake on each of stage 2

$\text{Error}_m(x) = \underbrace{h_1(x)}_{\text{mistake}} + \underbrace{\alpha_1^{*} h_2(x)}_{\text{mistake}}$

Step-2.b

$\text{Error}_m(x) \Rightarrow \alpha_1^{*} h_1(x) + \alpha_2^{*} h_2(x)$

Step-n

$\text{Error}_m(x) = h^{(1)}(x) + \alpha_1^{*} h_2(x) + \alpha_2^{*} h_3(x) \ldots + \alpha_n^{*} h_n(x)$

$\boxed{\text{Error}(x): \quad h_m(x) = \sum_{i=1}^{n} \alpha_i^{*} h_i(x)}$

Example:

$\text{Mistake}_{\text{Tolerance}} \Rightarrow \boxed{\text{Error}_m(x)}$

$\boxed{y_{\text{Pred}}}$

$\text{error}^{(1)}, y^{(1)}, h_1(x)$

$= 1 + 0 = 0$

$\boxed{0.5}$

$h_1(x) \rightarrow$ 3.5 (Mistake)

$\Rightarrow$ 3 $\rightarrow$ 0.5 0.5

$\Rightarrow$ 3.5 $\rightarrow$ 1.5 1.5

$\Rightarrow$ 4.5

$\boxed{\text{Error}^{(1)}}$ $h_2(x) + \alpha_1^{*} h_2^{(1)}(x)$

$\Rightarrow$ 3 $\rightarrow$ 0.5 0.5

$\Rightarrow$ 3.5 $\rightarrow$ 1.5 1.5

$\Rightarrow$ 4.5

$\text{error}^{(2)} \Rightarrow y^{(2)} - \hat{F}_1(x)$

$\Rightarrow$ 1.5 $\rightarrow$ 1.5 - 3.5

$\Rightarrow$ 4.5

Model

$x \rightarrow \boxed{\text{Error}^{(1)}} \rightarrow (y^{(1)} - \hat{F}_1(x))$

$h_2^{(1)}(x)$ (residual)

$\Rightarrow$ 4.5

$y^{(1)} \rightarrow 2$

$\hat{F}_2(x) = h_1(x) + \alpha_1^{*} h_2(x) \Rightarrow \hat{F}_1(x) + h_2(x)$

$\Rightarrow 3.0 + 1.5 + 0.5 = 5$

$\Rightarrow 3.0 + 1.5 - 1.5$

$\Rightarrow 4.5$

$\boxed{\hat{F}_2(x)}$

Model

$\hat{F}_M(x) = h_1 + \sum_{i=1}^{M} \alpha_i \cdot h_i(x)$

$\hat{F}_M(x) = h_1(x) + \alpha_1 h_1(x) + \alpha_2 h_2(x) + \ldots + \alpha_M h_M(x)$

Decision Boundary.

Linear Reg. vs GBDT

$\hat{y} = w_0 + \sum_{j=1}^{D} w_j x_j$

$\hat{y}$ $\rightarrow$ $\boxed{\text{Model}}$ $\rightarrow$ $y^{(1)}$

$\hat{y}$ $\rightarrow$ $\boxed{\text{Model}}$ $\rightarrow$ $\text{residual}^{(1)}$

$\hat{y} \mid \hat{F}_1(x) \Rightarrow w_0^{(1)} + \sum_{j=1}^{D} w_j^{(1)} \cdot h_2^{(1)}(x)$

$\boxed{\text{residual}}$

$\Rightarrow \underbrace{\boxed{w_0^{(1)}} + \boxed{w_1^{(1)}} + \boxed{w_2^{(1)}}}_{\text{Learning of Sequential residual}}$

Weight are not learning?

$\#\text{Trees} = 3 = M$

$\hat{F}_3(x) \Rightarrow \underbrace{h_1(x)}_{\text{weight}} + \underbrace{\alpha_1^{*} \cdot h_2(x)}_{\text{weight}} + \underbrace{\alpha_2^{*} \cdot h_3(x)}_{\text{weight}}$

$\Rightarrow$ 1.5 $\rightarrow$ 0.5 $\rightarrow$ 0.25

Why learn slower.

$\hat{y} = \begin{cases} 1 & \\ -1 & \end{cases}$

$1.5 + 0.5 \cdot 1 + 0.25 \cdot 1 = 2.25$

$\Rightarrow$ ____

Trees.

$y^{(1)} \Rightarrow 22.2$

h_1

h_2

$\Rightarrow$ 15 $\Rightarrow$ 22.2

$\#\text{Trees} = 3 = M$

$\hat{F}_3(x) \Rightarrow \underbrace{h_1(x)}_{\text{Mean}} + \underbrace{\alpha_1^{*} \cdot h_2(x)}_{\text{Error}} + \underbrace{\alpha_2^{*} \cdot h_3(x)}_{\text{Error}}$

$(1, 2, 3)$

$\Rightarrow \boxed{\text{M.V.E}}$