Gradient Boosting Decision Tree

Zengchang Qin (PhD)

Gradient Boosting

① Given a training data $D = \{(x^{(1)}, y^{(1)}) \cdots (x^{(N)}, y^{(N)})\}$
In boosting, we aim to fit residual of previous weak classifiers.

$$F_{m+1}(x) = F_m(x) + h(x)$$

Our aim is to learn $h(x^{(i)})$ that:

$$F_{m+1}(x^{(i)}) = F_m(x^{(i)}) + h(x^{(i)}) = y^{(i)}$$

② In supervised learning, the goal is to find an approximation $\hat{F}(x)$ to the function $F(x)$ that minimize the expected value of the loss function.

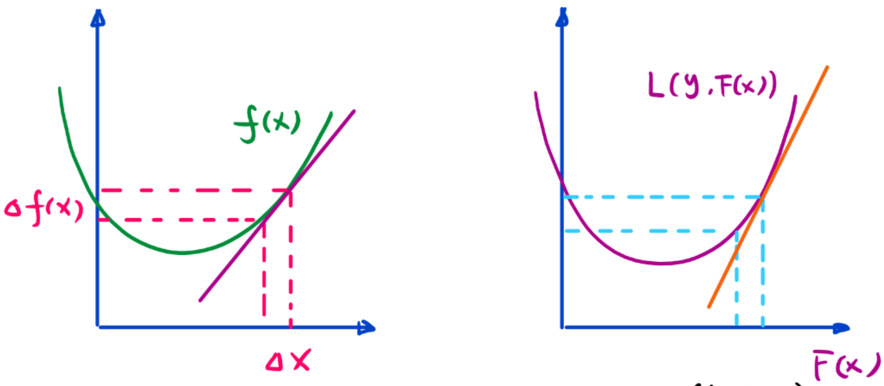
$$\hat{F} = \underset{F}{\arg\min}\, E_{x,y}[L(y, F(x))]$$

$$F(x) = \sum_{i=1}^{M} \alpha_i h_i(x)$$

$$F_0(x) = \underset{c}{\arg\min} \sum_{i=1}^{N} L(y^{(i)}, c)$$

$$F_m(x) = F_{m-1}(x) + \underset{h_m \in \mathcal{H}}{\arg\min} \sum_{i=1}^{N} L(y^{(i)}, F_{m-1}(x^{(i)}) + h_m(x^{(i)}))$$

where loss function $L(y, f(x)) = \sum_i (y^{(i)} - f(x^{(i)}))^2$



$$x \leftarrow x - \alpha \frac{\partial f(x)}{\partial x}$$

$$-\frac{\partial L(y, F(x))}{\partial F(x)}$$

③ If we use tree model, e.g. CART,

For $m = 1 \cdots M$, Calculate pseudo-residuals training on $\{x^{(i)}, r_m^{(i)}\}_{i=1,2\cdots N}$

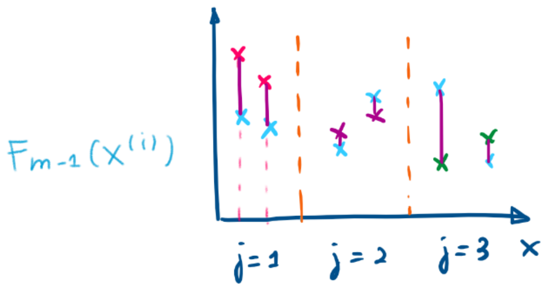
$$r_m^{(i)} = -\left[\frac{\partial L(y^{(i)}, F(x^{(i)}))}{\partial F(x^{(i)})}\right]_{F(x)=F_{m-1}(x)}$$

$r_m^{(i)}$ is the residual of data $x^{(i)}$ given $F_{m-1}(x^{(i)})$

Learn a CART (with best j, s) to fit $r_m^{(i)}$

(It is like we are learning a CART given a new training data: $D = \{(x^{(1)}, r_m^{(1)}), (x^{(2)}, r_m^{(2)}) \cdots (x^{(N)}, r_m^{(N)})\}$ for $m = 1, 2 \cdots M$)

$$C_{mj} = \underset{c}{\arg\min} \sum_{x \in R(m,j)} L(y^{(i)}, F_{m-1}(x^{(i)}) + c)$$



minimize residual by choosing proper C_{mj} for each leaf node of the tree.

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} C_{mj} I(x \in R(m,j))$$

The final tree is

$$F_m(x) = \sum_{m=1}^{M} \sum_{j=1}^{J_m} C_{mj} I(x \in R(m,j))$$

That is what we call GBDT (Gradient Boosting Decison Tree)

④ Shrinkage

In updating of $F_m(x)$: $F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} C_{mj} I$

if we add a learning rate θ ($\theta > 0$)

$$F_m(x) = F_{m-1}(x) + \theta \sum_{j=1}^{J} C_{mj} I(x \in R(m,j))$$

usually, $\theta \in [0.001, 0.01]$ to reduce the fitting speed.