Gradient boosting

adaboost use expontential loss $\Rightarrow$ sensitive to outliers.

GB: additive model

$$f_m(x) = f_{m-1}(x) + \rho_m h_m(x) \quad ①$$

mth iteration: $Loss = \sum_{i=1}^{N} L(y_i, f_m(x_i))$. If we think $f(x)$ as parameters (like gradient descend: $\theta = \theta - \alpha \frac{\partial}{\partial \theta} L(\theta)$)

$$f_m(x) = f_{m-1}(x) - \rho_m \cdot \frac{\partial}{\partial f_{m-1}(x)} L(y, f_{m-1}(x)) \quad ②$$

compare ① and ②, if we set

$$h_m(x) = -\frac{\partial}{\partial f_{m-1}(x)} L(y, f_{m-1}(x))$$

it means we use $h_m(x)$, a base learner to estimate the negative gradient of loss in $(m-1)$th iteration. so that we can minimize $L(f)$ by gradient descend.

Negative gradient, called "response" or "pseudo residual".

If residual $r = y - f(x)$ larger, the difference is larger, then in next iteration, we can correct this.

Pseudocode:

① Initialize: $f_{(0)} = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$

② for $m=1$ to M:

(a) negative gradient: $\tilde{y}_i = -\frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)}$ $(i=1,2,\cdots,N)$

(b) use base learner $h_m(x)$ to estimate $\tilde{y}_i$ by MSE:

$$w_m = \arg\min_{w} \sum_{i=1}^{N} [\tilde{y}_i - h_m(x_i; w)]^2$$

(c) decide ρ_m by line search, to min L:

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \rho h_m(x_i; w_m))$$

(d) $f_m(x) = f_{m-1}(x) + \rho_m h_m(x; w_m)$

③ output $f_m(x)$

Gradient Boosting Decision Tree (GBDT)

Base learner is DT.

Single DT: $h(x; \{R_j, b_j\}_1^J) = \sum_{j=1}^{J} b_j I(x \in R_j)$. $\{R_j\}_1^J$ is the leaf node and $\{b_j\}_1^J$ is the output of each $\{R_j\}_1^J$.

change 2.b to $\{R_{jm}\}_1^J = \arg\min_{\{R_{jm}\}_1^J} \sum_{i=1}^{N} [\tilde{y}_i - h_m(x_i; \{R_{jm}, b_{jm}\}_1^J]^2$

where $b_{jm} = \underset{x \in R_{jm}}{\text{mean}} \tilde{y}_{im}$.

Compute optimal value $\gamma_{jm} = \rho_m b_{jm}$ for each R_j. so

2.C : $\gamma_{jm} = \underset{\gamma}{\arg\min} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$

Pseudocode:

① Initialize: $f_0(x) = \underset{\gamma}{\arg\min} \sum_{i=1}^{N} L(y_i, \gamma)$

② for $m=1$ to M:

(a) negative gradient: $\tilde{y}_i = -\dfrac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)}$ $(i=1,2,\cdots,N)$

(b) $\{R_{jm}\}_1^J = \underset{\{R_{jm}\}_1^J}{\arg\min} \sum_{i=1}^{N} [\tilde{y}_i - h_m(x_i; \{R_{jm}, b_{jm}\}_1^J)]^2$

(c) $\gamma_{jm} = \underset{\gamma}{\arg\min} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$

(d) $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$

③ Output $f_M(x)$