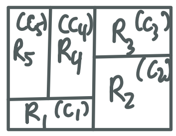
Extreme gradient boosting

$\vdots$		
i	x_i^T	y_i
$\vdots$		
n		

$$f(x) = \sum_{k=1}^{K} h_k(x)$$

$\downarrow$



$$s_i = f(x_i) = \sum_{k=1}^{K} h_k(x_i)$$

$$p_i = \frac{e^{s_i}}{1+e^{s_i}} = Pr(y_i = 1 \mid s_i)$$

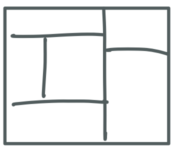
$$1 - p_i = \frac{1}{1+e^{s_i}}$$

Boosting

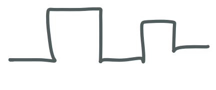
At iteration t

$$s_i = \boxed{\sum_{k=1}^{t-1} h_k(x_i)} + \underbrace{h_t(x_i)}_{\text{new tree}}$$

current committee $\hat{s}_i$ (fixed)



Δs_i



Recall

$$\text{Likelihood} = \prod_{i=1}^{n} \boxed{p_i^{y_i}(1-p_i)^{1-y_i}} \;\; = \begin{cases} p_i & y_i = 1 \\ 1-p_i & y_i = 0 \end{cases}$$

$$\text{log-lik} = \sum_{i=1}^{n} \left[y_i \log p_i + (1-y_i)\log(1-p_i) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \left(s_i - \log(1+e^{s_i})\right) + (1-y_i)\left(-\log(1+e^{s_i})\right) \right]$$

$$= \sum_{i=1}^{n} \underbrace{\left[y_i s_i - \log(1+e^{s_i}) \right]}_{}$$

$$L(s_i) = L(\hat{s}_i + h_t(x_i))$$

$$\doteq L(\hat{s}_i) + L'(\hat{s}_i)\, h_t(x_i) - \frac{1}{2} L''(\hat{s}_i)\, h_t(x_i)^2$$

$$= \text{const} +$$

$$L'(s_i) = y_i - \frac{e^{s_i}}{1+e^{s_i}} = y_i - p_i = e_i$$

$$L''(s_i) = \frac{d}{ds_i}\left(-\frac{e^{s_i}}{1+e^{s_i}} + 1 \right) = \frac{d}{ds_i}\left(\frac{1}{1+e^{s_i}} \right) = \frac{-e^{s_i}}{(1+e^{s_i})^2} = -\frac{e^{s_i}}{1+e^{s_i}} \, \frac{1}{1+e^{s_i}} = -p_i(1-p_i) = -w_i$$

$$= -\frac{1}{2}\hat{w}_i\left[h_t(x_i)^2 - 2\frac{\hat{e}_i}{\hat{w}_i}h_t(x_i)\right] + \text{constant}$$

$$= -\frac{1}{2}\hat{w}_i\Big[\underbrace{\boxed{\frac{\hat{e}_i}{\hat{w}_i}}}_{\hat{y}_i} - h_t(x_i)\Big]^2 + \text{constant}$$

iterated reweighted least squares

each step grows a new tree. to expand imperfections

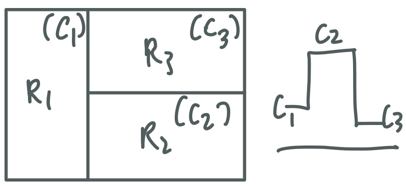
process

current committee $\rightarrow \hat{s}_i \rightarrow \hat{p}_i \rightarrow \begin{cases} \hat{e}_i = y_i - \hat{p}_i \\ \hat{w}_i = \hat{p}_i(1-\hat{p}_i) \end{cases} \rightarrow \hat{y}_i = \frac{\hat{e}_i}{\hat{w}_i}$

$\xrightarrow{\hat{y}_i, \hat{w}_i}$ grow a new tree $h_t(x_i)$ $\rightarrow$ add to committee $\rightarrow$ current committee

$$\min \sum_{i=1}^{n} \hat{w}_i\big(\hat{y}_i - \overset{(x_i^T\Delta\beta)}{h_t(x_i)}\big)^2 + \lambda M + \gamma\sum_{m=1}^{M} c_m^2$$

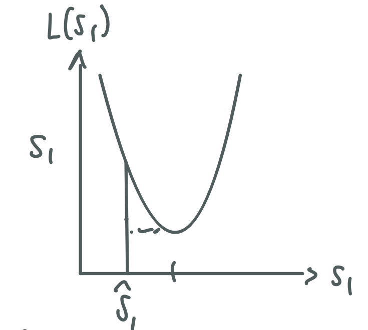
we minimize this loss function b/c we want to grow a newtree to fit the data



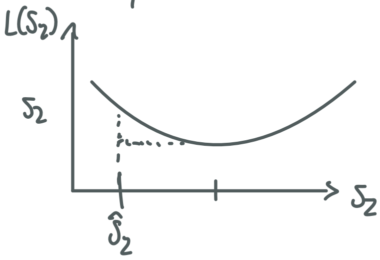
$\hat{y}_i$ represent the imperfections we want to expand

$\hat{e}_i$ gives us the gradient

gradient is scaled by $\hat{w}_i$



big $\hat{w}_i$



small $\hat{w}_i$

$$h_t(x_i) = \sum_{m=1}^{n} c_m \mathbb{1}(x_i \in R_m)$$

$$\hat{c}_m = \frac{\sum_{i=1}^{n} y_i \mathbb{1}(x_i \in R_m)}{\sum_{i=1}^{n} \mathbb{1}(x_i \in R_m) + \gamma}$$

shrinkage parameter

$$\boxed{\sum_{i=1}^{n} \hat{w}_i(\hat{y}_i - \hat{c}_m)\mathbb{1}(x_i \in R_m) + \lambda M}$$ weighted loss function

growtree by recursive partitioning, only depends on regions/partitioning

Adaboost (origin)

$$f(x) = \sum_{k=1}^{K} \beta_k \boxed{h_k(x)} \rightarrow \text{classifier} \rightarrow \{+,-\}$$

in XGboost, $h_k(x) \in \mathbb{R}$, regression tree as opposed to binary classifier

$$\hat{y} = \text{sign}(f(x)) = \begin{cases} + & f(x) > 0 \\ - & f(x) < 0 \end{cases}$$

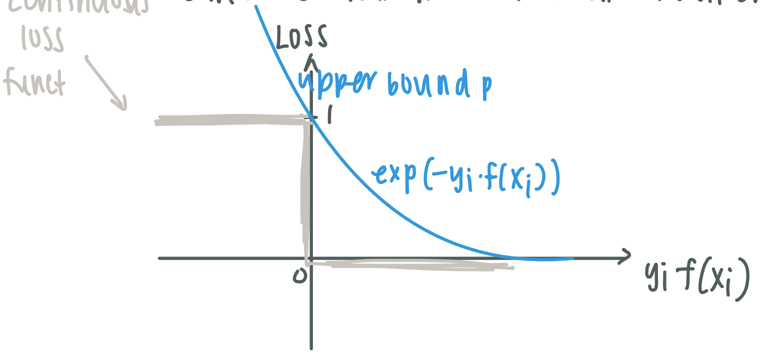
final result is a classifier; each tree is a classifier

withholds information as it does not include info on magnitude

purpose is to answer a theoretical question

weak learn = strong learn??

can the combination of weak classifiers yield a strong classifier?



$$\text{Loss} = \sum_{i=1}^{n} \exp(-y_i s_i) = \sum_{i=1}^{n} \exp\left(-y_i(\hat{s}_i + \beta_t h_t(x_i))\right)$$

$$= \boxed{\sum_{i=1}^{n} \exp(-y_i \hat{s}_i)} \exp(-\beta_t y_i h_t(x_i))$$

D_i (distribution)

$$s_i = \underbrace{\boxed{\sum_{k=1}^{t-1} \beta_k h_k(x_i)}}_{\text{current committee} \rightarrow \hat{s}_i \text{ fixed}} + \underbrace{\beta_t h_t(x_i)}_{\text{to be learned}}$$

$+/-$ $\quad$ $+/-$

$$D_i \xleftarrow{\text{normalize}} \frac{D_i}{\sum_{i=1}^{n} D_i}$$

$$D_i \rightarrow \sum_{i=1}^{n} D_i = 1$$

Difficult examples

$D = (D_i, i = 1, \ldots, n)$ pays attention to examples that current committee did not do well

$$\text{Loss} = \sum_{i=1}^{n} D_i \exp(-\beta_t \overset{+/-}{y_i} \overset{+/-}{h_t(x_i)})$$

$+1$: $y_i = h_t(x_i)$

-1 : $y_i \neq h_t(x_i)$ did not classify correctly

$$= \underbrace{\sum_{y_i = h_t(x_i)} D_i e^{-\beta_t}}_{a} + \underbrace{\sum_{y_i \neq h_t(x_i)} D_i e^{\beta_t}}_{b}$$

where $\sum_{y_i = h_t(x_i)} D_i = 1 - \varepsilon$, $\sum_{y_i \neq h_t(x_i)} D_i = \varepsilon$

$$a+b \geq 2\sqrt{ab}$$

$$(\sqrt{a}-\sqrt{b})^2 \geq 0$$

"=" $\quad a=b$

$\min \beta_t \rightarrow \geq (1-\varepsilon)(\varepsilon)$ lower bound

$$(1-\varepsilon)e^{-\beta_t} = \varepsilon e^{\beta_t}$$

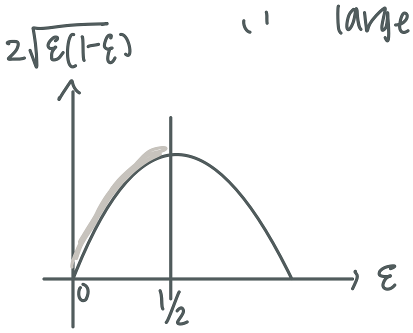
$$\frac{1-\varepsilon}{\varepsilon} = e^{2\beta_t}$$

$$\hat{\beta}_t = \frac{1}{2}\log\frac{1-\varepsilon}{\varepsilon}$$

voting right of $h_t(\,)$

If ε is small, then ratio of $\frac{1-\varepsilon}{\varepsilon}$ is very large, so $\hat{\beta}_t$ is large

" large, " " small " small



from range $(0,\frac{1}{2})$ function is increasing

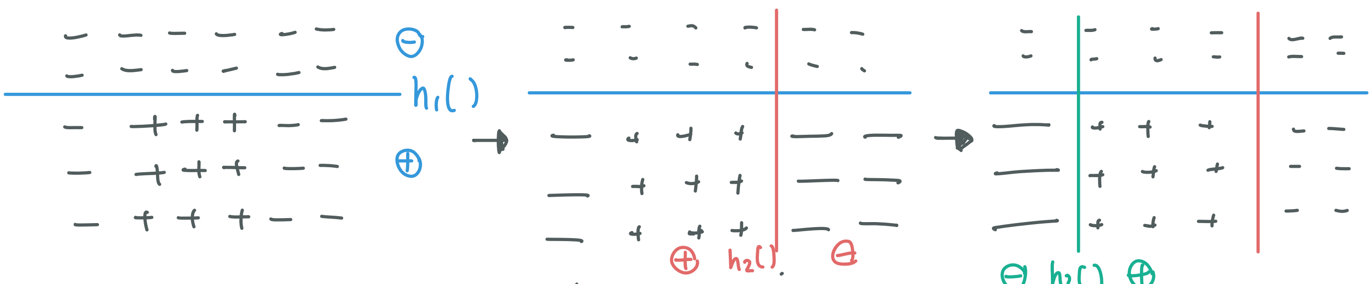
recut $h_t(\,)$ with minimal ε

assign $\hat{\beta}_t = \frac{1}{2}\log\frac{1-\varepsilon}{\varepsilon}$

process:

current committee $\longrightarrow D \longrightarrow$ grow $h_t(\,)$

$\hat{\beta}_t$



data is reweighted to heavily weight incorrectly classified data