MSE:

$r_i = (y_i - f_{M-1}(x_i))$ ← pseudo resid.

Why is this equal to $\partial L / \partial f = \partial L / \partial \hat{y}_i$

MSE:

$$L(y, f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\frac{\partial L}{\partial f} = \frac{-1}{n} \sum_{i=1}^{n} 2(y_i - f(x_i))$$

For any obs x_i, what's its contribution to the loss?

$$\left.\frac{\partial L}{\partial f}\right|_{f=f(x_i)} = \frac{-2}{n}(y_i - f(x_i))$$

$$\left.-\frac{\partial L}{\partial f}\right|_{f=f(x_i)} = \frac{2}{n}(y_i - f(x_i))$$

Why can I take just

$r_i = (y_i - f(x_i))$?

The reason is:

Minimizing MSE $\Leftrightarrow$ Minimizing $\frac{n}{2}$ MSE

The loss function for a single point $(x_i, f(x_i))$

$$L(y_i, f(x_i)) = (y_i - f(x_i))^2$$

$$= (y_i - f_i)^2$$

$f(x_i) = f_i$

$$-\frac{\partial L}{\partial f_i} = +2(y_i - f_i)$$

Gradient Boosting Theory:

In general, the boosting setting uses an additive model:

$$f(x) = \sum_{m=1}^{M} T_m(x)$$

We have a loss function

$$L(y, f(x))$$

and we want to minimize

$$\sum_{i=1}^{n} L(y_i, f(x_i)) =$$

$$\sum_{i=1}^{n} L\left(y_i, \sum_{m=1}^{M} T_m(x_i)\right)$$

Fitting $(T_1, T_2, \ldots, T_M)$ at the same time is intractable, but I can make it easier by trying to fit one tree at a time. This is called <u>forward stagewise fitting</u>.

At any given stage m, we want to solve:

"Find T such that

$$\sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + T(x_i))$$

is minimized."

The first step (Step 0) is just to fit a constant.

<u>Step 0:</u>

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \overset{0+\alpha}{\underline{\underline{\alpha}}})$$

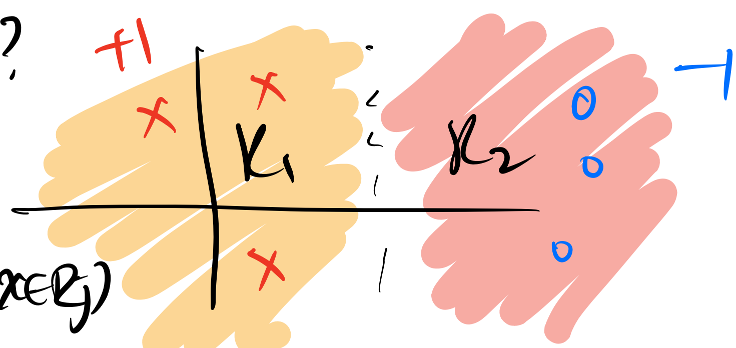
After that we'll move onto fitting trees, $T_m(x)$, which divide the predictor space into regions $R_1, R_2, \ldots, R_J$. In each of these regions, our tree is just a constant value.

The best constant for each region is the solution to:

$$\beta_j^* = \arg\min_{\beta_j} \sum_{x_i \in R_j} L(y_i, f_{m-1}(x_i) + \beta_j)$$

Why is it a constant for the tree $T_m(x)$?

$$T_m(x) = \sum_{j=1}^{J} \beta_j \mathbb{1}(x \in R_j)$$



The only issue is how to optimize

$$\sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + \underline{T(x_i)})$$

Can we do grad. descent in the space of functions?

Grad. Descent:

$$W \leftarrow W - \eta \frac{\partial F}{\partial W}$$

extend this idea to our current set up.

We are trying to optimize

$$\underline{\underline{L(f)}} = \sum_{i=1}^{n} L(y_i, f(x_i))$$

(f is a sum of trees.)

Consider this representation of f:

$$\tilde{f} = [\underline{\underline{f(x_1)}}, \underline{\underline{f(x_2)}}, \ldots, f(x_N)]$$

$$L(\tilde{f}) = \sum_{i=1}^{n} L(y_i, \hat{f}_i)$$

Now gradient descent looks like:

$$f_M(x_i) \leftarrow f_{M-1}(x_i) - \eta \frac{\partial L}{\partial f}\Big|_{f=\tilde{f}_i=f(x_i)}$$

compare w/

$$f_M(x) = f_{M-1}(x) + \nu T_M(x)$$

turn this into a tree

⇒ All I have to do is fit a tree

to $-\frac{\partial L}{\partial f}\Big|_{f=\tilde{f}_i=f(x_i)}$ ← pseudoresiduals.

Example: $f_M(x) = f_0 + \nu \sum_{m=1}^{M} T_m(x)$

Binary Classification for logloss

$$L(y, f) = \log(1 + e^{-yf})$$

Step 0:

Find the constant f_0 which minimizes

the total loss:

$$\circledast \quad \sum_{i=1}^{n} L(y_i, f(x_i))$$

$$= \sum_{i=1}^{n} \log\left(1+e^{-y_i f_0}\right) = \hat{L}$$

const

$$\frac{\partial \hat{L}}{\partial f_0} = \sum_{i=1}^{n} \frac{-y_i e^{-y_i f_0}}{1+e^{-y_i f_0}} \left(\frac{e^{y_i f_0}}{e^{y_i f_0}}\right)$$

$$= \sum_{i=1}^{n} \left(\frac{-y_i}{e^{y_i f_0}+1}\right) \stackrel{!}{=} 0$$

Notice $y_i \in \{+1, -1\}$.

Define $N_{\oplus} = \#\text{ of } +1\text{s} = \sum_{i=1}^{n} \mathbb{1}(y_i = +1)$

$N_{\ominus} = \text{"} \quad -1\text{s} = \sum_{i=1}^{n} \mathbb{1}(y_i = -1)$

$\downarrow$

$$= -\left(\sum_{(i:y_i=+1)} \left(\frac{y_i}{e^{y_i f_0}+1} \right) + \sum_{(i:y_i=-1)} \left(\frac{y_i}{e^{y_i f_0}+1} \right) \right)$$

$$= -\left(\sum_{\underline{\underline{(i:y_i=+1)}}} \frac{1}{e^{f_0}+1} + \sum_{(i:y_i=-1)} \frac{-1}{e^{-f_0}+1} \right)$$

$$= -\left(\frac{N_\oplus}{e^{f_0}+1} - \frac{N_\ominus}{e^{-f_0}+1} \right) = 0$$

$$\Rightarrow \quad \frac{N_\oplus}{e^{f_0}+1} = \frac{N_\ominus}{e^{-f_0}+1} \left(\frac{e^{f_0}}{e^{f_0}} \right)$$

$$= \frac{N_\ominus e^{f_0}}{1+e^{f_0}}$$

$$N_\oplus = N_\ominus e^{f_0} \Rightarrow f_0 = \log\left(\frac{N_\oplus}{N_\ominus} \right)$$

Exercise:

$$f_0 = \log\left(\frac{N_\oplus}{N_\ominus}\right) = \log\left(\frac{1+\bar{y}}{1-\bar{y}}\right).$$

<u>Stage m</u>: $(m=1,2,\dots,M)$

We need the pseudo residuals:

$$r_i = -\frac{\partial L}{\partial f_i}$$

Still in logloss:

$$L(y,f) = \log\left(1+e^{-yf}\right)$$

$$\frac{\partial L}{\partial f} = \frac{-ye^{-yf}}{1+e^{-yf}} = \frac{-y}{e^{yf}+1}$$

$$r_i = -\frac{\partial L}{\partial f}\bigg|_{f=f_{m-1}(x_i),\, y=y_i}$$

$$= y_i \Big/ \left(e^{y_i f_{m-1}(x_i)}+1\right).$$

Let's focus on just one region R_j &

try to find its best constant:

$$\beta_j^* = \underset{\beta}{\operatorname{argmin}} \sum_{x_i \in R_j} L(y_i, f_{m-1}(x_i) + \beta)$$

Define:

$$L_{R_j}(\beta) = \sum_{x_i \in R_j} L(y_i, f_{m-1}(x_i) + \beta)$$

Want to solve:

$$\frac{\partial L_{R_j}}{\partial \beta} = G(\beta) = 0$$

$$L_{R_j}(\beta) = \sum_{x_i \in R_j} L(y_i, f_{m-1}(x_i) + \beta)$$

$$= \sum_{x_i \in R_j} \log\left(1 + e^{-[y_i (f_{m-1}(x_i) + \beta)]}\right)$$

$$G(\beta) = \frac{\partial L_{R_j}}{\partial \beta} = \sum_{x_i \in R_j} \frac{-y_i e^{-[y_i(f_{m-1}(x_i)+\beta)]}}{\left(1+e^{-[y_i(f_{m-1}(x_i)+\beta)]}\right)} \overset{!}{=} 0$$

This doesn't have a closed form sol.

but I can approximate the solution by:

Claim:

$$\beta \approx \frac{-G(0)}{G'(0)}.$$

Here's why:

Let's do a Taylor Exp @ $\beta = 0$.

$$G(\beta) = G(0) + (\beta - 0)\, G'(0) + O(\beta^2)$$

$$\approx G(0) + \beta\, G'(0) \overset{!}{=} 0$$

$$\beta = \frac{-G(0)}{G'(0)}.$$

Num:

$$-G(0)=\sum_{x_i\in R_j}\frac{+y_i\,e^{-[y_i(f_{m-1}(x_i))]}}{1+e^{-[y_i(f_{m-1}(x_i))]}}$$

$$=\sum_{x_i\in R_j}\frac{y_i}{e^{[y_i f_{m-1}(x_i)]}+1}$$

$$=\sum_{x_i\in R_j} r_j$$

Den:

$$G'(\beta)=\frac{\partial}{\partial\beta}\sum_{x_i\in R_j}\frac{-y_i\,e^{-[y_i(f_{m-1}(x_i)+\beta)]}}{(1+e^{-[y_i(f_{m-1}(x_i)+\beta)]})}$$

$$=\frac{\partial}{\partial\beta}\sum_{x_i\in R_j}\frac{-y_i}{(e^{y_i(f_{m-1}(x_i)+\beta)}+1)}$$

$$=\sum_{x_i\in R_j}\frac{+y_i}{(e^{y_i(f_{m-1}(x_i)+\beta)}+1)^2}\left[y_i e^{[y_i(f_{m-1}(x_i)+\beta)]}\right]$$

$$= \sum_{x_i \in R_j} \frac{1}{\left(e^{y_i(f_{m-1}(x_i)+\beta)}+1\right)} \frac{e^{[y_i(f_{m-1}(x_i)+\beta)]}}{\left(e^{y_i(f_{m-1}(x_i)+\beta)}+1\right)}$$

Claim:

$$= |r_i|$$

$$= \left| \frac{y_i}{e^{y_i[f_{m-1}(x_i)+\beta]}+1} \right|$$

$$= \frac{1}{e^{y_i(f_{m-1}(x_i)+\beta)}+1}$$

$$= 1 - |r_i|$$

$$G'(\beta) = \sum_{x_i \in R_j} |r_i|(1-|r_i|)$$

$$\beta_j^* \approx \frac{-G(0)}{G'(0)} = \frac{\sum_{x_i \in R_j} r_i}{\sum_{x_i \in R_j} |r_i|(1-|r_i|)}$$