Gradient boost for classification :

step 1 :→ Initial guess is log(odds) ⇒ $\text{odds} = \frac{\#\text{true}}{\#\text{false}}$

step 2 :→ We calculate the probability of ith datapoint ⇒ $p^{(i)}$ → using (m-1)th model

$$r_{i,m} = \left(y^{(i)} - p^{(i)}\right)$$

~~If we have multiple residuals in a leaf~~

$$\gamma_{j,m} = \text{output} = \frac{\sum_{i \in R_{j,m}} \text{residuals}}{\sum_{i \in R_{j,m}} \left[\text{previous prob}_i \times (1 - \text{previous probability}_i) \right]}$$

→ sum of all residuals in that leaf.

↓ for any leaf.

↓ previously predicted probability of that residual.

suppose ith data point ends up in $R_{j,m}$

$$\left[\log(\text{odds})\right]^{(i)} := \left[\log(\text{odds})\right]^{(i)} + \nu\, \gamma_{j,m}$$

step 3 :→ we find new residuals by converting $\left[\log(\text{odds})\right]^{(i)}$ to $p^{(i)}$ using ;

$$p^{(i)} = \frac{e^{[\log(\text{odds})]^{(i)}}}{1 + e^{[\log(\text{odds})]^{(i)}}}$$

$$r_{i,m} = y^{(i)} - p^{(i)}$$ and so onwards repeat.

~~some proofs:~~

<u>more mathematically:</u> $\rightarrow$

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$ and differentiable loss function $L(y_i, \overset{[\log(\text{odds})]^{(i)}}{F(x)})$

$$\sum_{i=1}^{n} -y^{(i)} \times [\log(\text{odds})]^{(i)} + \log\left(1+e^{[\log(\text{odds})]^{(i)}}\right) \leftarrow \text{here we use log likelihood}$$

$$-y^{(i)} + p^{(i)} = \frac{\partial L(y_i, F(x))}{\partial [\log(\text{odds})]^{(i)}}$$

step 1: Initialise model with constant value.

$$F_0(X) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma) = \log(\text{odds})$$

$$\sum_{i=1}^{n} L(y_i, \gamma) = \sum_{i=1}^{n} -y^{(i)} \times \gamma + \log(1+e^{\gamma})$$

$$\frac{\partial\left(\sum_{i=1}^{n} L(y_i, \gamma)\right)}{\partial \gamma} = \sum_{i=1}^{n} -y^{(i)} + \frac{e^{\gamma}}{1+e^{\gamma}} = 0.$$

$$n\left(\frac{e^{\gamma}}{1+e^{\gamma}}\right) = \sum_{i=1}^{n} y^{(i)}$$

$$\frac{1}{1+e^{-\gamma}} = \frac{\sum_{i=1}^{n} y^{(i)}}{n} \Rightarrow e^{-\gamma} = \frac{n - \sum_{i=1}^{n} y^{(i)}}{\sum_{i=1}^{n} y^{(i)}}$$

$$\Rightarrow \gamma = \log(\text{odds})$$

step 2: for $m = 1$ to M

A) compute $r_{im} = -\left[\dfrac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x) = F_{m-1}(x)}$ for $i = 1, \ldots, n$

here $L(y_i, F_{m-1}(x_i)) = -y^{(i)} F_{m-1}(x^{(i)}) + \log\left(1 + e^{F_{m-1}(x^{(i)})}\right)$

so, $r_{i,m} = y^{(i)} - \dfrac{e^{F_{m-1}(x^{(i)})}}{1 + e^{F_{m-1}(x^{(i)})}} = y^{(i)} - p^{(i)}$

B) Fit a regression tree to $r_{i,m}$ values and create terminals $R_{j,m}$ for $j = 1, 2, \ldots, J_m$

$\hookrightarrow$ number of ~~trees~~ leaf nodes

C) For $j = 1, 2, \ldots, J_m$

compute $\gamma_{j,m} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

output for each leaf node

$$\gamma_{j,m} = \underset{\gamma}{\operatorname{argmin}} \underbrace{\sum_{x_i \in R_{ij}} -y_i \times \left(F_{m-1}(x_i) + \gamma\right) + \log\left(1 + e^{F_{m-1}(x_i) + \gamma}\right)}$$

derivative wrt γ

$\downarrow$

$$\sum_{x_i \in R_{ij}} -y_i + \frac{e^{F_{m-1}(x_i) + \gamma}}{1 + e^{F_{m-1}(x_i) + \gamma}} = 0$$

$$\sum_{x_i \in R_{ij}} \frac{1}{1+e^{-(F_{m-1}(x^{(i)})+\gamma)}} - \frac{1}{1+e^{-F_{m-1}(x^{(i)})}} = \sum_{x_i \in R_{ij}} y^{(i)} - p^{(i)}$$

"<u>THIS GETS US NOWHERE</u>"

we approximate loss function as second order Taylor polynomial.

$$L(y^{(i)}, F_{m-1}(x^{(i)}) + \gamma) \approx L(y^{(i)}, F_{m-1}(x^{(i)})) + \left[\frac{d}{dF()}(y^{(i)}, F_{m-1}(x^{(i)}))\right]_{x=x^{(i)}} \gamma$$
$$+ \frac{1}{2}\left[\frac{d^2}{dF()^2}(y^{(i)}, F_{m-1}(x^{(i)}))\right]_{x=x^{(i)}} \gamma^2.$$

Hence

$$\left[\frac{\partial}{\partial \gamma} L(y^{(i)}, F_{m-1}(x^{(i)}) + \gamma)\right]_{x=x^{(i)}} = \left[\frac{d}{dF()}(y^{(i)}, F_{m-1}(x^{(i)}))\right]_{x=x^{(i)}}$$
$$+ \left[\frac{d^2}{dF()^2}(y^{(i)}, F_{m-1}(x^{(i)}))\right]_{x=x^{(i)}} \gamma = 0.$$

$$\Rightarrow \gamma = -\frac{\left[\frac{d}{dF()}(y^{(i)}, F_{m-1}(x^{(i)}))\right]_{x=x^{(i)}}}{\left[\frac{d^2}{\partial F()^2}(y^{(i)}, F_{m-1}(x^{(i)}))\right]_{x=x^{(i)}}}.$$

hence,

$$\gamma_{j,m} = \underset{\gamma}{\operatorname{argmin}} \sum_{x^{(i)} \in R_{ij}} L(y^{(i)}, F_{m-1}(x^{(i)}) + \gamma)$$

$$\gamma_{j,m} = \underset{\gamma}{\operatorname{argmin}} \sum_{x^{(i)} \in R_{ij}} L(y^{(i)}, F_{m-1}(x^{(i)})) + \left[\frac{d}{dF()} L(y^{(i)}, F_{m-1}(x))\right]_{x=x^{(i)}} \gamma$$

$$\underbrace{+ \frac{1}{2}\left[\frac{d^2}{dF()^2} L(y^{(i)}, F_{m-1}(x))\right]_{x=x^{(i)}} \gamma^2}_{\text{derivative wrt } \gamma = 0.}$$

$$\gamma = \frac{- \sum_{x^{(i)} \in R_{ij}} \left[\frac{d}{dF()} L(y^{(i)}, F_{m-1}(x))\right]_{x=x^{(i)}}}{\sum_{x^{(i)} \in R_{ij}} \left[\frac{d^2}{dF()^2} L(y^{(i)}, F_{m-1}(x))\right]_{x=x^{(i)}}}$$

$$\gamma_{j,m} = \frac{\text{sum of residuals in node}}{\sum_{x^{(i)} \in R_{ij}} p^{(i)} \cdot (1-p^{(i)})}$$

D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} \mathbb{1}(x \in R_{jm})$