

GBM. (Gradient Boosting Machine)

★ For Regression.

核心思维 ① 先创建一个叶子节点，其值为所有 y 的均值。

② 创建一系列 tree 来拟合上一个 tree 带来的 residual。

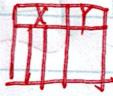
③ 直到模型达到我们的要求 / 再增加 tree 不会
显著提升模型效果。

Intuition: 每增加一棵树，模型结果都向正确的方向前进一步。
(若这一步过大，则容易过拟合。若每棵树过深，也容易
过拟合。)

★ 所以控制模型参数：树叶子数、树个数 和 learning rate.

算法理解：

Input: Data set $D \rightarrow$



A Loss function L

$$L = \sum_{i=1}^{N-1} (y^{(i)} - \hat{y}^{(i)})^2$$

A base learner Tree .

The number of iterations M 也可理解为树的个数

The learning rate η 每个树中更新的比例 (即 η 对树的影响)
增加到原先 y 上。

Step 1: Initialize $\hat{f}^{(0)}(x) = \hat{f}_{(0)}(x) = \hat{\theta}_0 = \arg \min_{\theta} \sum_{i=1}^N L(y_i, \theta);$

这一步实际在做的是取所有 y 的均值。

$$\arg \min_{\theta} \sum_{i=1}^N L(y_i, \theta) = \arg \min_{\theta} L(y_i, F(x)) \quad \xrightarrow{\text{predicted value}}$$

$$\frac{d}{d\text{predicted}} \cdot \frac{1}{2} (\text{observed} - \text{Predicted})^2 = -(\text{observed} - \text{Predicted})$$

$$\Rightarrow \text{这个 } \hat{\theta}_0 \text{ 满足 } \sum_{i=1}^N (y^{(i)} - \hat{\theta}_0) = 0$$

$$\Rightarrow N \cdot \hat{\theta}_0 = \sum_{i=1}^N y^{(i)} \Rightarrow \hat{\theta}_0 = \frac{1}{N} \cdot \sum_{i=1}^N y^{(i)}$$

Step 0): for $m=1, 2, \dots, M$ do 每一个 m 均为一棵树.

$$(A) \hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f_{\text{tree}})}{\partial f_{\text{tree}}} \right]_{f_{\text{tree}} = \hat{f}^{(m-1)}(x_i)} \quad \text{这里 } f_{\text{tree}} \text{ 指上一棵树的预测值}$$

这一步在做的实际上就是在算 residual

从之前推第一步时可知 $-\frac{\partial L(y_i, f_{\text{tree}})}{\partial f_{\text{tree}}}$ = observed - predicted.

$\hat{g}_m(x_i)$ 理解: $\hat{g}_m(x_i)$ 代表是第 i 个样本在算第 m 棵树时的偏差. x_i 代表第 i 个样本.

所以它的意思为在算第 m 棵树时. 第 i 样本与真实值之间的偏差. 即 p -residual.

每一个类别的对应 output

$$(B) \hat{\phi}_m = \arg \min_{\phi \in \Phi, \beta} \sum_{i=1}^n \left[(-\hat{g}_m(x_i)) - \beta \phi(x_i) \right]^2 \quad \text{判断是否属于}$$

这一步实际上在寻找一个最优秀的 tree model 来将样本划分到 n 个叶子节点之中.

★ 值得注意的是这里 $\phi(x_i)$ 并非一个数值. 而更类似于一个

于性函数. 判断该样本划分在哪个叶子节点. e.g.

$$\beta = (20, 21, 22, 23), \quad \phi(x_i) = (0, 0, 1, 0).$$

$$\beta \cdot \phi(x_i) = 22.$$

$$(C) \hat{\rho}_m = \arg \min_{\rho} \sum_{i=1}^n L(y_i, \hat{f}^{(m-1)}(x_i) + \rho \hat{g}_m(x_i))$$

$$= \arg \min_{\rho} \sum_{i=1}^n L(y_i - \hat{f}^{(m-1)}(x_i) - \rho \hat{g}_m(x_i))^2$$

$$= \arg \min_{\rho} \sum_{i=1}^n \left[(-\hat{g}_m(x_i)) - \rho \hat{\phi}(x_i) \right]^2$$

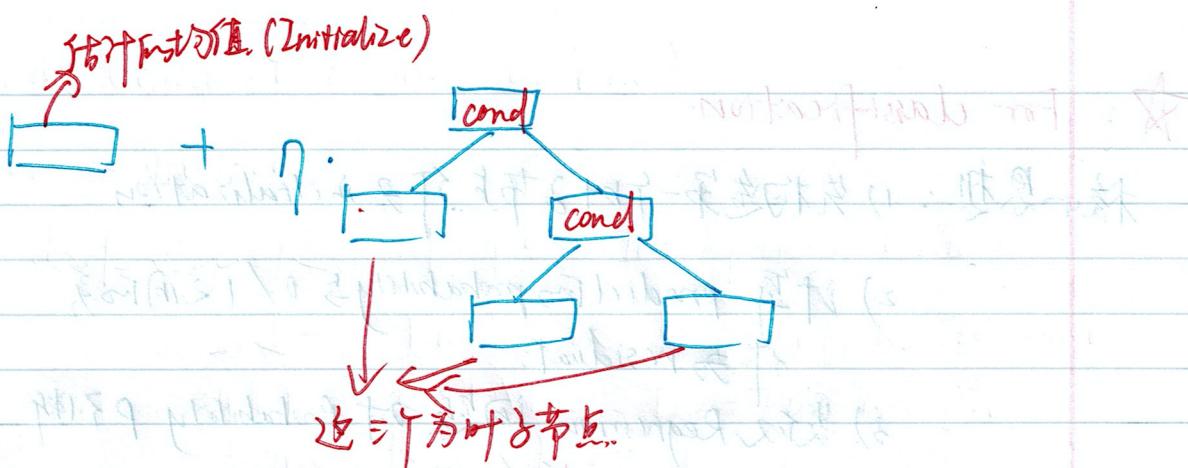
所以这一步实际上在估计上一步中的参数 β .

FIVE STAR

FIVE STAR

FIVE STAR

FIVE STAR



在每一个节点中，都对应向量中的一个值。这个 β 取值要让 Loss function 最小。那么它就是先分别这个叶子节点和所有 sample y 的平均值。

$$P_{im} = \underset{P_{im}}{\operatorname{arg\min}} \sum_{x_i \in R_{ij}} \langle (y_i, F_{m-1}(x_i) + P) \rangle$$

这是对向量中每个值都做 $\operatorname{arg\min}$ 。

$$(D) f_m(x) = \eta \hat{P}_m \hat{\phi}_m(x)$$

$$\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x)$$

这一步在做的是更新 prediction 值。

η 是 learning rate.

e.g. 若 $\hat{f}^{(m-1)}(x) = 50$. $y = 60$

$$\text{则 } -\hat{g}_m(x) = 60 - 50 = 10$$

若 此 中 仅 有 1 个 样 本 分 到 这 个 叶 子 节 点 则 有: $\hat{P}_m = 10$

若 $\eta = 0.1$

$$\text{则 } \hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \eta \hat{P}_m \hat{\phi}_m(x)$$

$$= 50 + 0.1 \times 10 = 51$$

step 13): Output: $\hat{f}(x) = \hat{f}^{(m)}(x) = \sum_{m=0}^M \hat{f}_m(x)$

将所有树的 predict 值加起来即为我们的最终结果

(每个树的 prediction 已经乘了 learning rate)

★: For classification

核心思想: 1) 先构造第一个叶子节点作为 initialization

2) 计算 prediction probability 与 0/1 之间的差
作为 residual.

3) 类似 Regression 问题对 probability P 不断更新迭代.

4). 由最终每个样本的 P 来给出分类结果

首先我们要搞清这里的 Loss function.

$\langle y_i, F(x) \rangle = \log(\text{likelihood of the observed Data given the prediction})$

$$= -\sum_{i=1}^N [y_i \log(p) + (1-y_i) \log(1-p)]$$

我们) 现在只看括号内:

$$y_i \cdot \log(p) + (1-y_i) \log(1-p)$$

$$= y_i [\log(p) - \log(1-p)] - \log(1-p)$$

$$= y_i \cdot \log \frac{p}{1-p} - \log(1-p) \quad \because \log \frac{p}{1-p} = \log(\text{odds})$$

$$\therefore \log(1-p) = \log(1 - \frac{e^{\log(\text{odds})}}{1+e^{\log(\text{odds})}}) = \log \left(\frac{1}{1+e^{\log(\text{odds})}} \right)$$

$$= \log(1) - \log(1+e^{\log(\text{odds})}) = -\log(1+e^{\log(\text{odds})})$$

$$= y_i \cdot \log(\text{odds}) + \log(1+e^{\log(\text{odds})})$$

\Rightarrow minimize. $\langle y_i, F(x) \rangle \Leftarrow \text{寻找 log odds} \text{ 使 } \langle y_i, F(x) \rangle \text{ 最小}$

FIVE STAR. ★★★★

FIVE STAR. ★★★★

FIVE STAR. ★★★★

FIVE STAR. ★★★★

因而我们对 $\log(\text{odds})$ 求导.

$$\frac{d}{d \log(\text{odds})} [-y_i \cdot \log(\text{odds}) + \log(1 + e^{\log(\text{odds})})] \\ = -y_i + \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}} = \underline{-y_i + \hat{P}} \rightarrow y_i - \hat{P} \text{ 是 residual}$$

Step 1: Initialization: $f^{(0)}(x) = f_0(x) = \hat{\theta}_0 = \arg \min_{\theta_0} \sum_{i=1}^n \mathcal{L}(y_i, \theta)$

这里其实和 regression 一样. 要想要 loss function 最小.

$$\text{就要 } \sum_{i=1}^n (y_i - P) = 0 \Rightarrow P = \frac{1}{n} \cdot \sum_{i=1}^n y_i.$$

\Rightarrow 实际上这个 P 就是样本中所有 positive 的个数占样本个数

Step 2: for $m = 1, 2, \dots, M$ do

$$(A) \hat{g}_m(x_i) = \left[\frac{\partial \mathcal{L}(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=\hat{f}^{(m)}(x)}$$

根据之前推导.

$$\hat{g}_m(x_i) = -(y_i - \hat{P}) \text{ 这里其实也是 residual.}$$

(B) Determine the structure $\{\hat{R}_{j,n}, \hat{Y}_{j-1}\}$ by selecting splits which maximize $Gain = \frac{1}{2} \left[\frac{\hat{\sigma}_L^2}{n_L} + \frac{\hat{\sigma}_R^2}{n_R} - \frac{\hat{\sigma}_{j,n}^2}{n_{j,n}} \right]$

(C) Determine the leaf weights $\{\hat{w}_{j,n}\}_{j=1}^J$ for the learnt structure by $\hat{w}_{j,n} = \arg \min_{w_j} \sum_{i \in j,n} \mathcal{L}(y_i, \hat{f}^{(m)}(x_i) + w_j)$.

这两步和之前 regression 推导的一样 只不过在这里 我们要

找 w_j 使得 $\sum (y_i, \hat{f}^{(m)}(x_i) + w_j)$ 最小很麻烦.

所以这里用暴力展开.

$$\begin{aligned} L(y_i, \hat{f}^{(m-1)}(x_i) + w_j) &\approx L(y_i, \hat{f}^{(m-1)}(x_i)) + \frac{d}{df}(y_i, \hat{f}^{(m-1)}(x_i)) \cdot w_j \\ &\quad - \frac{1}{2} \frac{d^2}{df^2}(y_i, \hat{f}^{(m-1)}(x_i)) w_j^2 \\ \frac{\partial L}{\partial w_j} &\approx \frac{d}{df}(y_i, \hat{f}^{(m-1)}(x_i)) + \frac{d^2}{df^2}(y_i, \hat{f}^{(m-1)}(x_i)) w_j = 0 \\ \Rightarrow w_j &= -\frac{\frac{d}{df}(y_i, \hat{f}^{(m-1)}(x_i))}{\frac{d^2}{df^2}(y_i, \hat{f}^{(m-1)}(x_i))} \end{aligned}$$

→ 这是在边上中仅有一个样本的情况下下面讨论多个样本

$$\begin{aligned} \frac{\partial L(y_1, \hat{f}^{(m-1)}(x_1) + w_j)}{\partial w_j} &\approx \frac{d}{df}(y_1, \hat{f}^{(m-1)}(x_1)) + \frac{d^2}{df^2}(y_1, \hat{f}^{(m-1)}(x_1)) w_j \\ \frac{\partial L(y_2, \hat{f}^{(m-1)}(x_2) + w_j)}{\partial w_j} &\approx \frac{d}{df}(y_2, \hat{f}^{(m-1)}(x_2)) + \frac{d^2}{df^2}(y_2, \hat{f}^{(m-1)}(x_2)) w_j \\ \Rightarrow w_j &= -\frac{\left[\frac{d}{df}(y_1, \hat{f}^{(m-1)}(x_1)) + \frac{d}{df}(y_2, \hat{f}^{(m-1)}(x_2)) \right]}{\frac{d^2}{df^2}(y_1, \hat{f}^{(m-1)}(x_1)) + \frac{d^2}{df^2}(y_2, \hat{f}^{(m-1)}(x_2))} \end{aligned}$$

→ 下面解决分子 分母 分母是什么

看分子.

$$\begin{aligned} -\frac{d}{df}(y_i, \hat{f}^{(m-1)}(x_i)) &\rightarrow \text{根据推导 它就是 residual} \\ &\rightarrow y_i - \hat{P}_i \end{aligned}$$

看分子.

$$\frac{d}{d \log(\text{odds})} \left[-y_i + \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}} \right]$$

FIVE STAR

FIVE STAR

FIVE STAR

FIVE STAR

$$\begin{aligned}
 &= - (1 + e^{\text{log odds}})^{-2} e^{2\text{log odds}} + (1 + e^{\text{log odds}})^{-1} \cdot e^{\text{log odds}} \\
 &= - \frac{e^{2\text{log odds}}}{(1 + e^{\text{log odds}})^2} + \frac{e^{\text{log odds}}(1 + e^{\text{log odds}})}{(1 + e^{\text{log odds}})^2} \\
 &= \frac{e^{\text{log odds}} \cdot 1}{(1 + e^{\text{log odds}})(1 + e^{\text{log odds}})} = \frac{e^{\text{log odds}}}{1 + e^{\text{log odds}}} \cdot \frac{1}{1 + e^{\text{log odds}}}
 \end{aligned}$$

$$= \hat{p} \cdot (1 - \hat{p})$$

$$(1) \hat{f}_m(x) = \sum_{j=1}^J w_{j,n}^* \cdot I(x_j \in R_{j,n}^*)$$

$$\hat{f}_m(x) = \hat{f}_{m-1}(x) + \hat{f}_{m,n}(x)$$

这一步更新为 regression 模型

$$\text{Output: } f^{(M)}(x) = \sum_{m=0}^M \hat{f}_{m,n}(x) \rightarrow \text{再去到分类器}$$