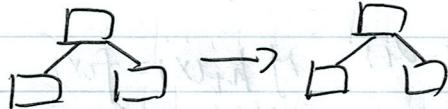


Adaboost

adaboost的核心思想：(也与 random forest 区别)

① weak learners → stump.



一个叶子两个结点的叫 stump.

由于其结构简单，称为 weak learner

② 为每个样本赋予权重。训练时，每个样本等权 在每个 tree 结束后，重新为每个样本赋予权重。

③ 每个 tree 的分类结果会乘一个比例 (weight) 然后决定最终的分类结果

• 树与树之间并不 independent，前两棵树的 prediction 会通过给样本重新赋权重的方式影响下一棵树的构建。

算法详解

训练集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$;

基学习算法 g ; 迭代轮数 T

step 1: $D_t(x) = 1/m \rightarrow$ 即为每个样本赋予相同的权重。

for $t = 1, 2, 3, \dots, T$ do

step 2: $h_t = g(D, D_t)$ 训练一个 stump. h_t 代表第 t 个基学习器

step 3: $E_t = \Pr_{x \sim D_t} (h_t(x) \neq f(x))$ E_t 指在分布 D_t 下训练误差

这里指的是若所有样本 weight 相加和为 1。
那么 E_t 就是所有分类错样本的 weight 之和

step 4: If $E_t > 0.5$ then break

这里加入了限定条件，若错误率过大，则停止

$$\text{Step 5: } \alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

这一步是在计算该 stump 的权重. 即 account of say
该 stump 在决定最终结果时的话语权

$$\text{Step 6: } D_{t+1}(x) = \frac{D_t(x)}{Z_t} \times \begin{cases} \exp(-\alpha_t) & \text{if } h_t(x) = f(x) \\ \exp(\alpha_t) & \text{if } h_t(x) \neq f(x) \end{cases}$$

这一步是在更新 sample 的权重. 判断对的权重降低.

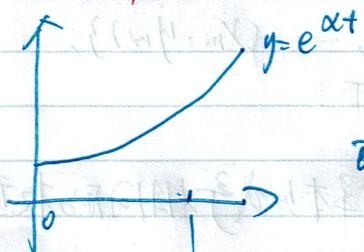
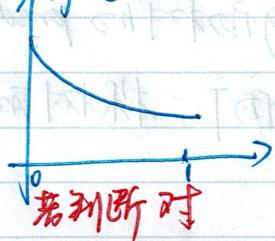
判断错的权重增加

Intuition.

$$y = e^{-\alpha t}$$

当 $\alpha_t \uparrow$, $\exp(-\alpha_t) \downarrow$.

即越有把握判断对 sample weight 越小



当 $\alpha_t \uparrow$, $\exp(\alpha_t) \uparrow$.

即越有把握判断对, 但实际却错时. 这样的 sample weight 越大

Step 7:

$$\text{Output: } H(x) = \text{sign} \left(\sum_{i=1}^T \alpha_i h_i(x) \right)$$

下面将围绕导数法推导理论背后的数学原理:

① 优化损失函数相当于优化贝叶斯最优错误率

② 为什么每个 stump 的权重为 $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$

③ 为什么每次更新 weight 用 $D_{t+1}(x) = \frac{D_t(x)}{Z_t} \times \begin{cases} \exp(-\alpha_t) & h_t(x) = f(x) \\ \exp(\alpha_t) & h_t(x) \neq f(x) \end{cases}$

问题1:

各基本学习器的线性组合

$$H(x) = \sum_{t=1}^T \alpha_t h_t(x)$$

$h_t(x)$ 是第 t 个学习器对 y 的判断
 α_t 是问题 2 的目标.

现在设我们的损失函数:

$$\text{Loss}(H|D) = \mathbb{E}_{x \sim D} [e^{-f(x)H(x)}] \rightarrow \text{模型最终判断}$$

\downarrow
数据集 x 在 D 分布下

现在要优化损失函数令其最小化; 则对其实导.

$$\frac{\partial \text{Loss}(H|D)}{\partial H(x)} = -e^{-H(x)} \cdot P(f(x)=1|x) + e^{H(x)} \cdot P(f(x)=-1|x) = 0$$

$$\text{解得: } H(x) = \frac{1}{2} \ln \frac{P(f(x)=1|x)}{P(f(x)=-1|x)}$$

$$\begin{aligned} \text{即 } \text{sign}(H(x)) &= \text{sign} \left(\frac{1}{2} \ln \frac{P(f(x)=1|x)}{P(f(x)=-1|x)} \right) \\ &= \begin{cases} 1, & P(f(x)=1|x) > P(f(x)=-1|x) \\ -1, & P(f(x)=1|x) < P(f(x)=-1|x) \end{cases} \\ &= \arg \max P(f(x)=y|x) \end{aligned}$$

因而从以上推导中. 可以发现找到的使 Loss function 最小的 $H(x)$. 这个 $H(x)$ 恰好在给定数据集下分类错误最小

问题2:

在解决上述问题后 下面我们) 将验证分类器权重 $\alpha_t, \bar{\alpha}_t$.

$$\text{Loss}(\alpha_t h_t | D_t) = \mathbb{E}_{x \sim D_t} [e^{-f(x) \cdot \alpha_t h_t(x)}]$$

\rightarrow 二项分布

$$\begin{aligned} &= \mathbb{E}_{x \sim D_t} [e^{-\alpha_t} \mathbb{I}(f(x) = h_t(x)) + e^{\alpha_t} \mathbb{I}(f(x) \neq h_t(x))] \\ &= e^{-\alpha_t} \cdot P_{x \sim D_t} (f(x) = h_t(x)) + e^{\alpha_t} P_{x \sim D_t} (f(x) \neq h_t(x)) \\ &= e^{-\alpha_t} (1 - \epsilon_f) + e^{\alpha_t} \epsilon_f \end{aligned}$$

$\downarrow \epsilon_f$

下面寻找最优的 α_t , 使得 loss function 最小。

$$\frac{\partial \text{Loss}(\alpha_t h_t | D_t)}{\partial \alpha_t} = -e^{-\alpha_t(1 - \epsilon_t)} + e^{\alpha_t \epsilon_t} = 0$$

$$\Rightarrow \alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

问题3:

在找到每个分类器的最优 α_t 后, 我们要对 sample 重新进行
试校。在这个过程中, 当前的分类器已经训练结束, 我们
希望下一个分类器可以纠正当前分类器的所有错误或能
纠正越多越好, 即最小化:

$$\text{Loss}(h_{t+1} + h_t | D_t) = E_{x \sim D} [e^{-f(x)(h_{t+1}(x) + h_t(x))}]$$

$$= E_{x \sim D} [e^{-f(x)h_{t+1}(x)} \cdot e^{-f(x)h_t(x)}]$$

$$f(x) = h_t^2(x) = 1$$

$$f(x)h_t^2(x) = 1$$

下面对 $e^{-f(x)h_t(x)}$ 展开:

$$\text{Loss}(h_{t+1} + h_t | D_t) \approx E_{x \sim D} [e^{-f(x)h_{t+1}(x)} (1 - f(x)h_t(x) + \frac{f^2(x)h_t^2(x)}{2})]$$

$$= E_{x \sim D} [e^{-f(x)h_{t+1}(x)} (1 - f(x)h_t(x) + \frac{1}{2})]$$

对于理想分类器

$$h_t(x) = \arg \min_h \text{Loss}(h_{t+1} + h_t | D_t) \rightarrow \text{这个定值因为 } h_{t+1}(x) \text{ 已确定}$$

$$= \arg \min_h E_{x \sim D} [e^{-f(x)h_{t+1}(x)} (\frac{3}{2} - f(x)h_t(x))]$$

$$= \arg \max_h E_{x \sim D} [e^{-f(x)h_{t+1}(x)} \cdot f(x)h_t(x)]$$

$$= \arg \max_h E_{x \sim D} \left[\frac{e^{-f(x)h_{t+1}(x)}}{E_{x \sim D} [e^{-f(x)h_{t+1}(x)}]} \cdot f(x)h_t(x) \right]$$

我们知道 $E_{x \sim D} [e^{-f(x)h_t(x)}]$ 是常数

令 D_t 表示新后分布.

$$D_t(x) = \frac{D(x) \cdot e^{-f(x)h_t(x)}}{E_{x \sim D} [e^{-f(x)h_t(x)}]}$$

根据概率期望的定义，相当于令

$$\begin{aligned} h_t(x) &= \arg \max_h E_{x \sim D} \left[\frac{e^{-f(x)h(x)}}{E_{x \sim D} [e^{-f(x)h(x)}]} f(x) h(x) \right] \\ &= \arg \max_h E_{x \sim D} [f(x) h(x)] \end{aligned}$$

$\forall f(x), h(x) \in \mathbb{R}^1, \exists$ 有：

$$f(x)h(x) = 1 - 2 \mathbb{I}(f(x) \neq h(x))$$

$$\therefore h_t(x) = \arg \min_h E_{x \sim D_t} [\mathbb{I}(f(x) \neq h(x))]$$

下面我们在最优化框架下，利用梯度的分布。

$$\begin{aligned} D_{t+1}(x) &= \frac{D(x) \cdot e^{-f(x)h_t(x)}}{E_{x \sim D} [e^{-f(x)h_t(x)}]} \\ &= \frac{D(x) \cdot e^{-f(x)h_t(x)} e^{-f(x) \Delta_t h_t(x)}}{E_{x \sim D} [e^{-f(x)h_t(x)}]} \quad \text{调整次} \\ &= D_t(x) \cdot e^{-f(x) \Delta_t h_t(x)} \end{aligned}$$

原分布 对分布进行更新.