§ 讲 义

2019 年 4 月 17 日

1 AdaBoost理论研究

集成学习是一类著名机器学习方法, 通过使用一系列学习器进行学习, 并使用某种规则将各个学习结果进行整合从而获得比单个学习器显著优越的泛化性能, 常用方法包括Boosting, Bagging, random forest等. Boosting 是一族可将弱学习器提升为强学习器的学习方法, 这族算法的基本原理类似: 先从初始训练集训练出一个基学习器, 再根据基学习器的预测能力对训练样本分布进行调整, 使得先前基学习器做错的训练样本在后续得到更大的权重, 从而受到更多的关注, 然后基于调整后的样本分布来训练下一轮基学习器; 如此重复进行, 直到基学习器数目达到事先指定的值T, 最终将这T 个基学习器进行加权结合. Boosting族算法最著名的代表是AdaBoost.

1.1 Boosting起源

Boosting起源于1989年计算机理论科学STOC会议中强学习与弱学习之间是否存在等价性问题. 给定输入空间$\mathcal{X}$和输出空间$\mathcal{Y}=\{+1,-1\}$, 假设$\mathcal{D}$ 为空间$\mathcal{X}$上潜在未知分布函数, 函数$f\colon \mathcal{X}\to\mathcal{Y}$为所学习的目标函数. 一个训练集由$m$个样本组成

$$S=\{(x_1,f(x_1)),(x_2,f(x_2)),\cdots,(x_m,f(x_m))\},$$

其中每个样本是根据分布$\mathcal{D}$独立同分布采样所得. 函数空间$\mathcal{H}=\{h\colon \mathcal{X}\to\mathcal{Y}\}$, 学习过程则是根据训练集$S$产生一个假设$h\in\mathcal{H}$，使其最小化错误率

$$\Pr_{x\sim\mathcal{D}}\left[I[f(x)\neq h(x)]\right].$$

一个学习问题是弱可学习的是指存在一种算法, 对任意给定的分布$\mathcal{D}$, 都能在关于m和d的多项式时间内学习得到一个分类器h满足

$$\Pr_{(x,y)\sim\mathcal{D}}[I[h(x)\neq f(x)]]<\frac{1}{2}-\frac{1}{\mathrm{poly}(m,d)},$$

这里$\mathrm{poly}(m,d)$是关于训练集大小m和维度d的多项式.

一个学习问题是强可学习的是指对任意给定的分布$\mathcal{D}$, 任意给定$\epsilon > 0$和$\delta > 0$, 存在一种关于m和d的多项式时间学习算法, 学习得到一个分类器h满足

$$\Pr_{S\sim\mathcal{D}^m}[\Pr_{x\sim\mathcal{D}}[I[h(x) \neq f(x)]] > \epsilon] \leq \delta.$$

从直观上看, 弱学习器指在多项式时间内存在一个学习器, 其错误率略低于50%, 即存在比随机猜略好的学习方法; 强可学习器指在多项式时间内存在高精度学习器. Kearns & Valiant 在1989年STOC会议论文中提出一个重要的理论问题: 弱学习算法与强学习算法之间是否存在等价性? 如果两者等价, 那么只要找到一些比随机猜略好的弱学习算法就可以直接将其提升为强学习算法, 而不必直接去找通常情况下很难获得的强学习算法.

Schapire于1990年对这个问题通过构造性证明给出了肯定性回答, 而这个证明就是最初的boosting算法, Freund于1995年提出一种更高效的算法. 这两种算法在解决实际问题时存在一个问题, 即需要事先知道弱学习算法在数据分布$\mathcal{D}$中学习正确率的下限. 然而在实际应用中, 仅仅知道训练数据集, 而不知道数据的整体分布, 前面所提两种算法不能用于实际问题, 仅存在理论意义.

1.2 AdaBoost算法

第一个可用于实际问题的Boosting算法是Freund & Schapire于1995年提出的AdaBoost (Adaptive Boosting)算法, 如Algorithm 1所示, 该算法不需要任何关于弱学习器在真实分布中正确率等先验信息, 可直接应用于各种实际学习问题.

AdaBoost算法可描述为: 在每一轮迭代中, 产生一个弱学习器h_t, 该学习器比随机猜算法的性能略好. 然后根据产生的弱学习器h_t, 每一个样本被赋予一个权重, 如果某样本点已经被准确地分类, 那么它的权重越小, 表明在下一论训练被选中的概率越低; 相反, 如果某样本点没有被准确地分类, 那么它的权重就会得到提高, 即根据分类器来提高被它错分的样本权重, 并降低被正确分类的样本权重. 然后更新过样本集并用于下一轮的学习训练, 整个训练过程如此迭代地进行下去, 最后将所有得到的弱分类器通过权重结合起来.

AdaBoost算法有很多种推导, 比较容易理解的是基于"加性模型"(additive model), 即基学习器的线性组合

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

优化指数损失函数(exponential loss)

$$\ell_{\exp}(H|\mathcal{D}) = E_{x\sim\mathcal{D}}[e^{-f(x)H(x)}].$$

在AdaBoost算法中, 第一个基分类器h_1是将基学习器算法用于初始数据训练数据得到的; 此后迭代生

Algorithm 1 AdaBoost算法

输入: 训练数据集$S=\{(x_1,f(x_1)),(x_2,f(x_2)),\cdots,(x_m,f(x_m))\}$和迭代轮数$T$

初始化: $\mathcal{D}_1(x_i)=1/m$

1: **for** $t=1$ to T **do**
2: 根据分布$\mathcal{D}_t$构造基学习器$h_t\colon \mathcal{X}\to\mathcal{Y}$
3: 计算基学习器h_t的错误率$\epsilon_t=\Pr_{x\sim\mathcal{D}_t}[f(x)\neq h_t(x)]$
4: **if** $\epsilon_t>1/2$ **then**
5: 算法停止 %%%基学习器比随机猜性能还差
6: **end if**
7: 计算学习器h_t权重$\alpha_t=\frac{1}{2}\ln[(1-\epsilon_t)/\epsilon_t]$
8: 根据基学习器预测能力更新分布

$$\mathcal{D}_{t+1}(x_i)=\mathcal{D}_t(x_i)\exp(-\alpha_t f(x_i)h_t(x_i))/Z_t$$

其中Z_t是规范化因子, 保证D_{t+1}是一个分布.

9: **end for**

输出: 最后的分类器$\mathrm{sgn}[H(x)]$, 其中$\mathrm{sgn}[\cdot]$ 表示符号函数, 以及

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

成h_t和α_t, 当基分类器h_t基于分布$\mathcal{D}_t$产生后, 该基分类器的权重α_t应使得$\alpha_t h_t$最小化损失函数

$$\begin{aligned}\ell_{\exp}(\alpha_t h_f|\mathcal{D}_t)&=E_{x\sim\mathcal{D}_t}[e^{-f(x)\alpha_t h_t(x)}]\\&=E_{x\sim\mathcal{D}_t}[e^{-\alpha_t}I[f(x)=h_t(x)]]+E_{x\sim\mathcal{D}_t}[e^{\alpha_t}I[f(x)\neq h_t(x)]]\\&=e^{-\alpha_t}E_{x\sim\mathcal{D}_t}[I[f(x)=h_t(x)]]+e^{\alpha_t}E_{x\sim\mathcal{D}_t}[I[f(x)\neq h_t(x)]]\\&=(1-\epsilon_t)e^{-\alpha_t}+\epsilon_t e^{\alpha_t}\end{aligned}$$

其中$\epsilon_t=E_{x\sim\mathcal{D}_t}[I[f(x)\neq h_t(x)]]=\Pr_{x\sim\mathcal{D}_t}[f(x)\neq h_t(x)]$. 对上面指数函数求导, 并令其导数为零, 即

$$\frac{d\ell_{\exp}(\alpha_t h_f|\mathcal{D}_t)}{d\alpha_t}=-(1-\epsilon_t)e^{-\alpha_t}+\epsilon_t e^{\alpha_t}=0,$$

求解可得

$$\alpha_t=\frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t}$$

这正是Algorithm 1中第7行的分类器权重更新公式.

下面的定理表明AdaBoost算法能在训练集上将一系列弱学习器提升为一强学习器:

定理1.1. *假设AdaBoost算法每轮迭代过程中基学习器产生的错误率分别为$\epsilon_1,\epsilon_2,\ldots,\epsilon_T$, 则最终产生的学习器$H(x)=\sum_{t=1}^{T}\alpha_t h_t(x)$ 满足*

$$\frac{1}{m}\sum_{i=1}^{m}I[sgn(H(x_i))\neq f(x_i)]\leq 2^T\prod_{t=1}^{T}\sqrt{\epsilon_t(1-\epsilon_t)}.$$

如果存在$\gamma \in (0, 1/2)$, 使得错误率$\epsilon_i < 1/2 - \gamma$ $(i \in [T])$成立, 则有

$$\frac{1}{m}\sum_{i=1}^{m} I[sgn(H(x_i)) \neq f(x_i)] \leq (1 - 4\gamma^2)^{T/2} \leq e^{-2\gamma^2 T}.$$

因此, 只要AdaBoost在每次迭代过程中比随机猜略好, 即$\epsilon_i < 1/2 - \gamma$, 则训练错误率随迭代轮数的增加而呈指数性下降, 从而构造出一个强学习器.

Proof. 在训练过程中, 初始分布$\mathcal{D}_1 = 1/m$, 以及在第t轮迭代中

$$\mathcal{D}_{t+1}(x_i) = \mathcal{D}_t(x_i)\exp(-\alpha_t f(x_i)h_t(x_i))/Z_t. \tag{1}$$

这里Z_t是使$\mathcal{D}_{t+1}(x_i)$成为分布的正规化项, 即

$$Z_t = \sum_{i=1}^{m} \mathcal{D}_t(x_i)\exp(-\alpha_t f(x_i)h_t(x_i)).$$

由$\alpha_t = \frac{1}{2}\ln[(1-\epsilon_t)/\epsilon_t]$可知

$$\begin{aligned} Z_t &= e^{-\alpha_t}\sum_{i=1}^{m}\mathcal{D}_t(x_i)I[f(x_i) = h_t(x_i)] + e^{\alpha_t}\sum_{i=1}^{m}\mathcal{D}_t(x_i)I[f(x_i) \neq h_t(x_i)] \\ &= (1-\epsilon_t)e^{-\alpha_t} + \epsilon_t e^{\alpha_t} = 2\sqrt{\epsilon_t(1-\epsilon_t)}. \end{aligned}$$

由式(1)可知

$$\begin{aligned} Z_t\mathcal{D}_{t+1}(x_i) &= \mathcal{D}_t(x_i)\exp(-\alpha_t f(x_i)h_t(x_i)) \\ Z_{t-1}\mathcal{D}_t(x_i) &= \mathcal{D}_{t-1}(x_i)\exp(-\alpha_{t-1} f(x_i)h_{t-1}(x_i)) \\ &\vdots \\ Z_1\mathcal{D}_2(x_i) &= \mathcal{D}_1(x_i)\exp(-\alpha_1 f(x_i)h_1(x_i)) \end{aligned}$$

对上式两边同时相乘, 以及$\mathcal{D}_1(x_i) = 1/m$可得

$$\mathcal{D}_{t+1}(x_i)\prod_{t=1}^{T} Z_t = \frac{1}{m}\exp(-f(x_i)H_t(x_i)).$$

再对上式两边中x_i求和, 以及利用

$$\frac{1}{m}\sum_{i=1}^{m} I[\mathrm{sgn}(H(x_i)) \neq f(x_i)] \leq \frac{1}{m}\sum_{i=1}^{m}\exp(-f(x_i)H(x_i)) = \sum_{i=1}^{m}\mathcal{D}_{t+1}(x_i)\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} Z_t.$$

定理得证. □

1.3 AdaBoost VC维泛化性

首先给出基于VC维的定理:

定理1.2. 假设函数空间$\mathcal{F}$是VC维为d, S表示一个大小为m的训练集. 对任意$\delta > 0$, 以及任意的学习器$f \in \mathcal{F}$, 下式至少以$1-\delta$ 的概率有

$$\Pr_{\mathcal{D}}[y \neq f(x)] \leq \Pr_{S}[y \neq f(x)] + O(\sqrt{(d\ln(m/d) + \ln 1/\delta)/m}).$$

假设AdaBoost基学习器函数空间为$\mathcal{H}$, 以及AdaBoost算法运行了T 轮, 根据Baum & Haussler的理论可知$\mathcal{C}(\mathcal{H})$ 的VC维为$O(Td)$, 根据上面的定理有:

$$\Pr_{\mathcal{D}}[y \neq f(x)] \leq \Pr_{S}[y \neq f(x)] + O(\sqrt{(Td\ln(m/Td) + \ln 1/\delta)/m}).$$

从这个结论可知: 随着迭代轮数T的增加, AdaBoost所产生的学习器函数空间复杂度将增加, 从而导致AdaBoost学习方法过拟合的风险.

1.4 AdaBoost间隔理论

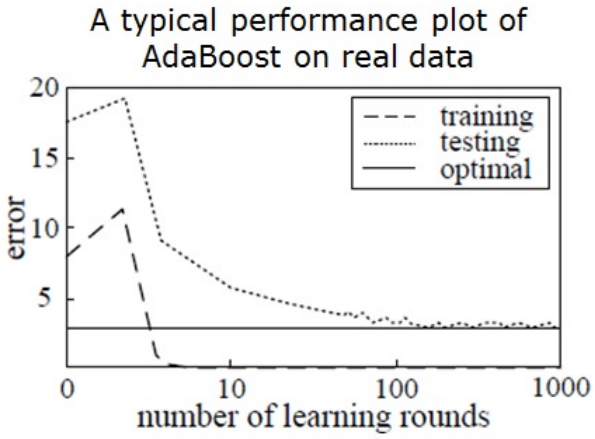

图 1: *AdaBoost*算法不易陷入过拟合现象

图1表明: AdaBoost算法即使训练轮数增加, 仍不易陷入过拟合, 甚至当训练错误率为零时还可以降低泛化错误率. 这一现象显然违背了奥卡姆剃刀原理: 在与经验数据一致的模型中, 模型越简单, 泛化性越好. 为解释AdaBoost的这种现象, Schapire等人于1998 年提出了著名的间隔理论.

给定基学习器的函数空间$\mathcal{H}$, 定义函数空间$\mathcal{H}$的凸包$\mathcal{C}(\mathcal{H})$为

$$\mathcal{C}(\mathcal{H}) = \left\{ H = \sum_i \alpha_i h_i \colon h_i \in \mathcal{H}, \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1 \right\}.$$

下面给出AdaBoost的间隔的定义:

定义1.1. 给定一个样本(x,y)和*AdaBoost*算法返回的学习器$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \in \mathcal{C}(\mathcal{H})$, 间隔定义为$yH(x)$, 即

$$yH(x) = \sum_{i\colon y=h_i(x)} \alpha_i - \sum_{i\colon y\neq h_i(x)} \alpha_i.$$

间隔计算正确/错误分类样本(x,y)的基分类器权重之差, 在一定程度上反映了分类器的可信度.

根据间隔定义, 下面给出著名的间隔理论:

定理1.3. 假设$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$是一个大小为$m$的训练集. 对任意$\delta$和$\theta > 0$, 每个分类器$H \in \mathcal{C}(\mathcal{H})$至少以$1 - \delta$的概率满足:

$$\Pr_{\mathcal{D}}[yH(x) < 0] \leq \Pr_{S}[yH(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\left(\frac{\ln m \ln|\mathcal{H}|}{\theta^2} + \ln\frac{1}{\delta}\right)^{1/2}\right),$$

其中$|\mathcal{H}|$表示空间$\mathcal{H}$所包含函数的个数(这里讨论有限情况), $\Pr_D[yH(x) < 0]$表示学习器H在真实分布D上的泛化错误率, $\Pr_S[yH(x) \leq \theta] = \sum_{(x_i,y_i)} I[y_i f(x_i) \leq \theta]/m$表示在训练集$S$中间隔小于$\theta$的样本所占比例.

该定理表明AdaBoost的泛化性与训练集的大小, 基学习器以及间隔θ相关, 但与算法的迭代轮数无关. AdaBoost算法当训练错误率为零时不停止训练, 可以进一步增大间隔θ, 从而提高学习方法的泛化性而没有过拟合, 间隔理论成功解释了AdaBoost不易过拟合现象.

Proof. 关键证明技术: Chernoff不等式和全概率公式$\Pr[A] = \Pr[A \cap B] + \Pr[A \cap \bar{B}]$.

对每个函数$H = \sum_{i=1}^{T} \alpha_i h_i \in \mathcal{C}(\mathcal{H})$, 其系数$\alpha_i$ 可以看作选择函数h_i 的概率. 从而每个函数H与函数空间$\mathcal{H}$的分布建立联系, 记为$\mathcal{Q}(H)$, 即$\Pr_{h\sim\mathcal{Q}(H)}[h = h_i] = \alpha_i$. 给定整数$N > 0$, 假设函数$g = \sum_{j=1}^{N} h_j/N$其中$h_j \sim \mathcal{Q}(H)$, 即函数$g$是根据分布$\mathcal{Q}(f)$从$\mathcal{Q}(H)$中随机选择$N$个函数的平均. 于是有

$$E[g] = E_{h_j\sim\mathcal{Q}(f)}\left[\sum_{j=1}^{N}\frac{h_j}{N}\right] = E_{h_1\sim\mathcal{Q}(f)}[h_1] = \sum_{i=1}^{T}\alpha_i h_i \in \mathcal{C}(\mathcal{H}) = H.$$

主要证明分为下面三步:

I: $\Pr_{\mathcal{D}}[yH(x) \leq 0] \Longleftrightarrow \Pr_{\mathcal{D}}[yg(x) \leq \theta/2]$

II: $\Pr_{\mathcal{D}}[yg(x) \leq \theta/2] \Longleftrightarrow \Pr_{S}[yg(x) \leq \theta/2]$

III: $\Pr_{S}[yg(x) \leq \theta/2] \Longleftrightarrow \Pr_{S}[yH(x) \leq \theta]$

对于第I步, 由全概率公式可得

$$\begin{aligned}
&\Pr_{\mathcal{D}}[yH(x) \leq 0] \\
&= \Pr_{\mathcal{D}}[yH(x) \leq 0, yg(x) \leq \theta/2] + \Pr_{\mathcal{D}}[yH(x) \leq 0, yg(x) > \theta/2] \\
&\leq \Pr_{\mathcal{D}}[yg(x) \leq \theta/2] + \Pr_{\mathcal{D}}[yH(x) \leq 0, yg(x) > \theta/2].
\end{aligned}$$

由$\Pr[A \cap B] = \Pr[A]\Pr[A|B] \leq \Pr[A|B]$, 以及Chernoff不等式可知

$$\Pr_{\mathcal{D}}[yH(x) \leq 0, yg(x) > \theta/2] \leq \Pr_{\mathcal{D}}[yg(x) > \theta/2 | yH(x) \leq 0] \leq \exp(-N\theta^2/8).$$

于是得到

$$\Pr_{\mathcal{D}}[yH(x) \leq 0] \leq \Pr_{S}[yg(x) \leq \theta/2] + \exp(-N\theta^2/8). \tag{2}$$

对于第II步, 首先发现

$$\begin{aligned}\Pr_S[yg(x) \le \theta/2] &= \frac{1}{m}\sum_{i=1}^{m} I[y_i g(x_i) \le \theta/2] \\ \Pr_{\mathcal{D}}[yg(x) \le \theta/2] &= E_S[\Pr_S[yg(x) \le \theta/2]].\end{aligned}$$

利用Union不等式和Chernoff不等式可得到

$$\Pr_{\mathcal{S}\sim\mathcal{D}^m}\left[\Pr_{\mathcal{D}}[yg(x) \le \theta/2] \ge \Pr_{\mathcal{S}}[yg(x) \le \theta/2] + \epsilon\right] \le (N+1)|\mathcal{H}|^N \exp(-2m\epsilon^2).$$

令$\delta = (N+1)|\mathcal{H}|^N \exp(-2m\epsilon^2)$, 则下面表达式至少以$1-\delta$的概率成立:

$$\Pr_{\mathcal{D}}[yg(x) \le \theta/2] < \Pr_{\mathcal{S}}[yg(x) \le \theta/2] + \sqrt{(1/2m)\ln((N+1)|\mathcal{H}|^N/\delta)}. \tag{3}$$

对于第III步, 再次利用全概率公式有

$$\begin{aligned}&\Pr_S[yg(x) \le \theta/2] \\ &= \Pr_S[yg(x) \le \theta/2, yH(x) \le \theta] + \Pr_S[yg(x) \le \theta/2, yH(x) > \theta] \\ &\le \Pr_S[yH(x) \le \theta] + \Pr_S[yg(x) \le \theta/2, yH(x) > \theta].\end{aligned}$$

由Chernoff不等式可知

$$\Pr_S[yg(x) \le \theta/2, yH(x) > \theta] \le \Pr_S[yg(x) \le \theta/2 | yH(x) > \theta] \le \exp(-N\theta^2/8).$$

于是有

$$\Pr_S[yg(x) \le \theta/2] \le \Pr_S[yH(x) \le \theta] + \exp(-N\theta^2/8). \tag{4}$$

根据式(2)-(4), 下式至少以$1-\delta$ 的概率成立

$$\Pr_{\mathcal{D}}[yf(x) \le 0] \le \Pr_{\mathcal{S}}[yf(x) \le \theta] + 2\exp(-N\theta^2/8) + \sqrt{(1/2m)\ln((N+1)|\mathcal{H}|^N/\delta)}$$

通过令$N = 8\ln m/\theta^2$以及整理计算, 定理得证. □

习题1.1. 假设*AdaBoost*算法每轮迭代过程中基学习器产生的错误率分别为$\epsilon_1, \epsilon_2, \ldots, \epsilon_T$. 对任何$\theta > 0$, 最终产生的学习器$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$ 满足

$$\frac{1}{m}\sum_{i=1}^{m} I[y_i H(x_i) \le \theta] \le 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t^{1-\theta}(1-\epsilon_t)^{\theta}}.$$

1.5 AdaBoost的Rademacher泛化性分析

本节将利用Rademacher复杂度来研究AdaBoost算法的间隔泛化性, 经验Rademacher复杂度和分布Rademacher复杂度相关定义如下:

定义1.2. 给定样本数据$Z=\{z_1,\ldots,z_m\}$, 函数空间$\mathcal{H}$关于Z的经验*Rademacher*复杂度为

$$\hat{R}_Z(\mathcal{H})=\mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\Big].$$

函数空间$\mathcal{F}$关于$\mathcal{Z}$在分布$\mathcal{D}$上的*Rademacher*复杂度为

$$R_m(\mathcal{F})=\mathbb{E}_{Z\subset\mathcal{Z}:|Z|=m}\Big[\hat{R}_Z(\mathcal{F})\Big].$$

基于上述定义有

定理1.4. 对实值函数空间$\mathcal{H}:\mathcal{Z}\to[0,1]$, 根据分布$\mathcal{D}$从$\mathcal{Z}$独立同分布采样得到示例集$Z=\{z_1,z_2,\ldots,z_m\}$,$z_i\in\mathcal{Z}$, $0<\delta<1$, 对任意$f\in\mathcal{F}$, 下列不等式以至少$1-\delta$的概率成立:

$$\begin{aligned}E[f(z)] &\le \frac{1}{m}\sum_{i=1}^{m}f(z_i)+2R_m(\mathcal{F})+\sqrt{\frac{\ln(1/\delta)}{2m}},\\ E[f(z)] &\le \frac{1}{m}\sum_{i=1}^{m}f(z_i)+2\hat{R}_Z(\mathcal{F})+3\sqrt{\frac{\ln(2/\delta)}{2m}}.\end{aligned}$$

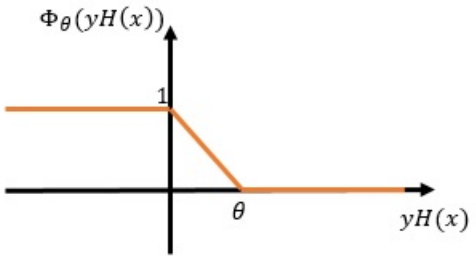

图 2: θ-间隔损失函数Φ_θ

下面研究基于Rademacher复杂度的Boosting间隔分布理论, 首先引入一个新的函数.

定义1.3. 给定任何$\theta>0$, 定义θ-间隔损失函数Φ_θ为

$$\Phi_\theta(t)=\begin{cases}1 & t<0\\ 1-t/\theta & t\in[0,\theta]\\ 0 & t>\theta\end{cases}$$

图 2 给出了θ-间隔损失函数Φ_θ的线段, 进一步得到如下性质:

引理1.1. 对任意给定样本(x,y)和函数H, 有如下关系成立

$$I[yH(x)<0]\le\Phi_\theta(yH(x))\le I[yH(x)<\theta].$$

引理1.2. 给定实值函数空间$\mathcal{H}$, 假设函数ϕ: $\mathbb{R}\to\mathbb{R}$满足李普西兹条件且其系数为L, 有

$$R_S(\phi\circ\mathcal{H})\le LR_S(\mathcal{H})\quad 以及\quad R(\phi\circ\mathcal{H})\le LR(\mathcal{H}).$$

引理1.3. 对任意给定的函数空间$\mathcal{H}$, 以及样本数据$Z=\{z_1,z_2,\ldots,z_m\}$, 有

$$R_S(\mathcal{C}(\mathcal{H}))=R_S(\mathcal{H})$$

成立, 其中$\mathcal{C}(\mathcal{H})=\{H=\sum_{t=1}^{T}\alpha_t h_t\colon \alpha_t\geq 0,\sum_t\alpha_t=1,h_t\in\mathcal{H}\}$.

Proof. 容易得到

$$\begin{aligned}
R_S(\mathcal{C}(\mathcal{H})) &= \frac{1}{m}E_\sigma\left[\sup_{H\in\mathcal{C}(\mathcal{H})}\sum_{i=1}^{m}\sigma_i H(z_i)\right]\\
&= \frac{1}{m}E_\sigma\left[\sup_{h_t\in\mathcal{H},\sum\alpha_t=1}\sum_{t=1}^{T}\alpha_t\sum_{i=1}^{m}\sigma_i h_t(z_i)\right]\\
&= \frac{1}{m}E_\sigma\left[\sup_{h_t\in\mathcal{H}}\max_{t\in[T]}\sum_{i=1}^{m}\sigma_i h_t(z_i)\right]\\
&= \frac{1}{m}E_\sigma\left[\sup_{h_t\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h_t(z_i)\right]=R_S(\mathcal{H}).
\end{aligned}$$

引理得证. □

下面给出基于Rademacher复杂度的间隔理论:

定理1.5. 假设$S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$是一个大小为$m$的训练集. 对任意$\delta$和$\theta>0$, 每个分类器$H\in\mathcal{C}(\mathcal{H})$至少以$1-\delta$的概率满足:

$$\Pr_{\mathcal{D}}[yH(x)<0] \leq \Pr_{S}[yH(x)\leq\theta]+\frac{2}{\theta}R_m(\mathcal{H})+\sqrt{\frac{\ln(1/\delta)}{2m}} \tag{5}$$

$$\Pr_{\mathcal{D}}[yH(x)<0] \leq \Pr_{S}[yH(x)\leq\theta]+\frac{2}{\theta}R_S(\mathcal{H})+3\sqrt{\frac{\ln(1/\delta)}{2m}} \tag{6}$$

习题1.2. 证明上述定理中式(6).

Proof. 假设函数空间$\hat{\mathcal{H}}=\{z=(x,y)\to yH(x)\colon H\in\mathcal{C}(\mathcal{H})\}$, 考虑一个新的函数集合

$$\bar{\mathcal{H}}=\{\Phi_\theta\circ f\colon f\in\hat{\mathcal{H}}\}.$$

首先根据引理1.1可知

$$\Pr_{\mathcal{D}}[yH(x)<0]\leq\Pr_{\mathcal{D}}[\Phi_\theta(yH(x))].$$

再根据定理1.4得到

$$\Pr_{\mathcal{D}}[\Phi_\theta(yH(x))]\leq\Pr_{S}[\Phi_\theta(yH(x))]+2R_m(\bar{\mathcal{H}})+\sqrt{\frac{\ln(1/\delta)}{2m}}.$$

由于函数Φ_θ满足李普西兹条件, 且其李普西兹系数为$1/\theta$, 根据引理1.2可得

$$\begin{aligned} R_m(\bar{\mathcal{H}}) = R_m(\Phi_\theta \circ \hat{\mathcal{H}}) &= \frac{1}{\theta} R_m(\hat{\mathcal{H}}) \\ &= \frac{1}{\theta m} \sum_{i=1}^{m} E_\sigma \left[\sup_{H \in \mathcal{C}(\mathcal{H})} \sum_{i=1}^{m} \sigma_i y_i H(x_i) \right] \\ &= \frac{1}{\theta m} \sum_{i=1}^{m} E_\sigma \left[\sup_{H \in \mathcal{C}(\mathcal{H})} \sum_{i=1}^{m} \sigma_i H(x_i) \right] \\ &= \frac{1}{\theta} R_m(\mathcal{C}(\mathcal{H})) = \frac{1}{\theta} R_m(\mathcal{H}), \end{aligned}$$

其中最后一个等式成立是因为引理1.3. 式(5)得证. □