§ 讲 义

2019 年 4 月 24 日

1　一致性

1.1　一致性定义

给定示例空间$\mathcal{X} \subseteq \mathbb{R}^d$以及标记空间$\mathcal{Y} = \{0, 1\}$, 假设$\mathcal{D}$是空间$\mathcal{X} \times \mathcal{Y}$的一个联合分布. 分布$\mathcal{D}$可分解为在示例空间$\mathcal{X}$的边缘分布$\mathcal{D}_{\mathcal{X}}$ 和条件概率$\eta(x)$, 其中

$$\eta(x) = \Pr[Y = 1|X = x].$$

函数$g\colon \mathcal{X} \to \{0, 1\}$给出了一个分类器或决策函数, 其错误率(error/classification probability)定义为

$$R(g) = \Pr_{(X,Y)\sim\mathcal{D}}[g(X) \neq Y] = E_{(X,Y)\sim\mathcal{D}}[I[g(X) \neq Y]].$$

下面考虑最优分类器及其性质. 首先定义贝叶斯风险(Bayes risk)和贝叶斯分类器(Bayes classifier)分别为

$$R^* = \min_g R(g) = R(g^*) \quad \text{和} \quad g^* = \arg\min_g R(g)$$

这里考虑所有可测函数g. 贝叶斯分类器和贝叶斯风险与数据的条件分布有如下关系:

定理1.1. 贝叶斯分类器和贝叶斯风险可分别表示为

$$g^*(x) = I[\eta(x) > 1/2] \quad \text{和} \quad R^* = E_{x\sim\mathcal{D}_{\mathcal{X}}}[\min\{\eta(x), 1 - \eta(x)\}].$$

对任意分类器$g\colon \mathcal{X} \to \mathcal{Y}$, 有$R(g) \geq R(g^*)$成立.

Proof. 对任意的分类器$g\colon \mathcal{X} \to \mathcal{Y}$, 其错误率为

$$\begin{aligned} R(g) &= \Pr_{(X,Y)\sim\mathcal{D}}[g(X) \neq Y] = E_{(X,Y)\sim\mathcal{D}}[I[g(X) \neq Y]] \\ &= E_{X\sim\mathcal{D}_{\mathcal{X}}}[\eta(X)I[g(X) \neq 1] + (1 - \eta(X))I[g(X) \neq 0]] \\ &= E_{X\sim\mathcal{D}_{\mathcal{X}}}[\eta(X)I[g(X) = 0] + (1 - \eta(X))I[g(X) = 1]]. \end{aligned}$$

考虑贝叶斯分类器$g^* = \arg\min_g R(g)$, 对任意$X \in \mathcal{X}$, 如果$\eta(X) \leq 1 - \eta(X)$, 即$\eta(X) \leq 1/2$, 那么有$g^*(X) = 0$ 成立; 如果$\eta(X) \geq 1 - \eta(X)$, 即$\eta(X) \geq 1/2$, 那么有$g^*(X) = 1$成立. 因此得到贝叶斯分类器$g^*(x) = I[\eta(x) \geq 1/2]$和贝叶斯错误率为$R^* = E_{x\sim\mathcal{D}_{\mathcal{X}}}[\min\{\eta(x), 1 - \eta(x)\}]$. 定理得证. □

定理1.2. *对任意给定的学习器g, 其与贝叶斯分类器之间的错误率之差为:*

$$R(g) - R^* = E_{x\sim\mathcal{D}_\mathcal{X}}[|1-2\eta(x)|I[g(x)\neq g^*(x)]].$$

Proof. 由定义可知

$$\begin{aligned} R(g) - R^* &= E_{(X,Y)\sim\mathcal{D}}I[g(X)\neq Y] - E_{(X,Y)\sim\mathcal{D}}I[g^*(X)\neq Y] \\ &= E_{X\sim\mathcal{D}_\mathcal{X}}[\eta(X)I[g(X)=0] + (1-\eta(X))I[g(X)=1]] \\ &\quad -E_{X\sim\mathcal{D}_\mathcal{X}}[\eta(X)I[g^*(X)=0] + (1-\eta(X))I[g^*(X)=1]] \end{aligned}$$

给定$X\in\mathcal{X}$, 如果$\eta(X)\leq 1/2$, 则有$g^*(X)=0$, 以及

$$\begin{aligned} &\eta(X)I[g(X)=0] + (1-\eta(X))I[g(X)=1] - (\eta(X)I[g^*(X)=0] + (1-\eta(X))I[g^*(X)=1]) \\ &= \eta(X) + (1-2\eta(X))I[g(X)=1] - \eta(X) = |1-2\eta(X)|I[g(X)\neq g^*(X)]; \end{aligned}$$

如果$\eta(X)\geq 1/2$, 则有$g^*(X)=1$, 以及

$$\begin{aligned} &\eta(X)I[g(X)=0] + (1-\eta(X))I[g(X)=1] - (\eta(X)I[g^*(X)=0] + (1-\eta(X))I[g^*(X)=1]) \\ &= 1-\eta(X) + (2\eta(X)-1)I[g(X)=0] - (1-\eta(X)) = |1-2\eta(X)|I[g(X)\neq g^*(X)]. \end{aligned}$$

定理得证. □

下面考虑一种经典的条件概率估计方法: Plug-in学习方法. 由于数据分布$\mathcal{D}$和条件概率$\eta(x)$未知. 因此一种学习方法是通过训练数据集估计$\eta(x)$, 例如最近邻方法, 随机森林等投票(Voting)方法均可看作Plug-in方法的变体. 假设通过训练数据估计的条件概率$\hat\eta(x)$, 定义Plug-in学习器为$g(x) = I[\hat\eta(x) > 1/2]$.

定理1.3. *Plug-in学习器$g(x) = I[\hat\eta(x) > 1/2]$与Bayes最优分类器的错误率之差为*

$$R(g) - R^* \leq 2E_{x\in\mathcal{X}}[|\hat\eta(x) - \eta(x)|] \leq 2\sqrt{E_{x\in\mathcal{X}}[(\hat\eta(x)-\eta(x))^2]}.$$

Proof. 由定理 1.2可知

$$R(g) - g^* = E_X[|1-2\eta(X)|I[g(X)\neq g^*(X)]].$$

如果$I[g(X)\neq g^*(X)]$, 即$I[\hat\eta(X)>1/2]\neq I[\eta(X)>1/2]$, 那么有$\hat\eta(X)>1/2$和$\eta(X)\leq 1/2$, 进一步得到

$$|1-2\eta(X)| = 2|1/2-\eta(X)| \leq 2|\hat\eta(X)-\eta(X)|;$$

或者有$\hat\eta(X)\leq 1/2$和$\eta(X)>1/2$, 进一步得到

$$|1-2\eta(X)| = 2|1/2-\eta(X)| \leq 2|\hat\eta(X)-\eta(X)|.$$

由于$f(t)=t^2$是凸函数, 由Jensen不等式可直接证明定理中第二个不等式. 定理得证. □

给定训练集$S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, 通过学习方法$F$学习得到一个分类器$F_{S_n}$. 随着样本规模$n$的增加，得到一些列学习器$F_{S_1}, F_{S_2}, \ldots, F_{S_n}, \ldots$. 一致性研究在足够多样本的情形下, 学习得到的分类器F_{S_n}是否趋于贝叶斯分类器, 下面给出形式化定义:

定义1.1. *称学习方法F满足一致性(贝叶斯一致性, 弱一致性), 如果有*

$$E_{S_n}[R(F_{S_n})] \to R^* \quad \text{当} \quad n \to \infty.$$

一致性反映了学习函数的错误率随样本的增加而趋于贝叶斯错误率.

1.2 拆分算法

给定训练数据集合$S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, 很多的经典学习算法将示例空间$\mathcal{X}$划分成多个互不相交的单元格$A_1, A_2, \ldots$, 并在每个单元格中利用投票(Voting)的方式给单元格中的每个示例赋予标记, 即

$$g_n(X) = \begin{cases} 0 & \text{if } \sum_{X_i \in A(X)} I[Y_i = 1] \leq \sum_{X_i \in A(X)} I[Y_i = 0] \\ 1 & \text{otherwise} \end{cases}$$

这里$A(X)$表示包含示例X的单元格. 这种拆分的方法依赖于存于n和训练示例$X_1, X_2, \ldots, X_n$, 但标记信息$Y_1, Y_2, \ldots, Y_n$与拆分无关. 下面考虑这种拆分学习方法的一致性, 需要满足两个条件: i) 拆分单元格应足够小, 从而可以检测分布的任意局部变化; ii) 在任意单元格内应包含足够多的训练样本, 从而保证投票方法的有效性. 给定集合A, 用$\mathrm{Diam}(A)$表示集合A 的直径, 即

$$\mathrm{Diam}(A) = \sup_{x,y \in A} \|x - y\|.$$

设

$$N(X) = \sum_{i=1}^{n} I[X_i \in A(X)],$$

即训练集中与示例X落入同一单元格的示例数. 下面给出拆分算法满足一致性的充分条件:

定理1.4. *拆分算法满足一致性, 如果以下两条件以概率成立*

$$A(X) \to 0 \quad \text{和} \quad N(X) \to \infty \quad \text{当} \quad n \to \infty.$$

Proof. 定义条件概率$\eta(X) = \Pr[Y = 1|X]$, 由定理 1.3可知: 要证明$E_{S_n}[R(F_{S_n})] \to R^*$, 只需要证明$E[|\hat{\eta}(X) - \eta(X)|] \to 0$, 其中

$$\hat{\eta}(X) = \frac{1}{N(X)} \sum_{X_i \in A(X)} Y_i.$$

进一步引入$\bar{\eta}(X) = E[\eta(X')|X' \in A(X)]$. 利用三角不等式有

$$E[|\hat{\eta}(X) - \eta(X)|] \leq E[|\hat{\eta}(X) - \bar{\eta}(X)|] + E[|\bar{\eta}(X) - \eta(X)|].$$

固定$X, X_1, X_2, \ldots, X_n$, 容易发现$N(X)\hat{\eta}(X)$本质上服从二项分布$B(N(X), \bar{\eta}(X))$, 因此有

$$
\begin{aligned}
& E[|\hat{\eta}(X)-\bar{\eta}(X)||X, X_1, \ldots, X_n] \\
\leq\ & E\left[\left|\frac{1}{N(X)} \sum_{X_i \in A(X)} Y_i-\bar{\eta}(X)\right| \middle| N(X)>0, X, X_1, \ldots, X_n\right]+\operatorname{Pr}[N(X)=0] \\
\leq\ & E\left[\sqrt{\frac{\bar{\eta}(X)(1-\bar{\eta}(X))}{N(X)}} I[N(X)>0] \middle| X, X_1, \ldots, X_n\right]+\operatorname{Pr}[N(X)=0],
\end{aligned}
$$

上式中最后一个不等式成立是根据引理 1.1. 对上式两边分别对$X, X_1, X_2, \ldots, X_n$求期望, 则对任意$k>3$, 有

$$
E[|\hat{\eta}(X)-\bar{\eta}(X)|] \leq \frac{1}{2} \operatorname{Pr}[N(X) \leq k]+\frac{1}{2 \sqrt{k}}+\operatorname{Pr}[N(X)=0].
$$

取$k=\sqrt{N(X)}$, 由$N(X) \rightarrow \infty$以概率成立, 得到$E[|\hat{\eta}(X)-\bar{\eta}(X)|] \rightarrow 0$.

有$\eta(X)$的连续性, 以及$A(X) \rightarrow 0$以概率形式成立, 有

$$
E[|\bar{\eta}(X)-\eta(X)|] \rightarrow 0.
$$

定理得证. □

引理1.1. *假设$Z_1, Z_2, \ldots, Z_n$是n个独立同分布的随机变量, 并服从Bernoulli分布, 即$Z_i \sim B(p)$. 则有*

$$
E\left[\frac{1}{n} \sum_{i=1}^n Z_i-E[Z_i]\right] \leq \sqrt{p(1-p) / n}.
$$

Proof. 由Jensen不等式有

$$
\begin{aligned}
E\left[\frac{1}{n} \sum_{i=1}^n Z_i-E[Z_i]\right] & \leq \sqrt{E\left[\frac{1}{n} \sum_{i=1}^n Z_i-E[Z_i]\right]^2} \\
& =\sqrt{\frac{1}{n^2} \sum_{i=1}^n E\left[Z_i-E[Z_i]\right]^2}=\sqrt{Var(Z_1) / n}=\sqrt{p(1-p) / n}.
\end{aligned}
$$

引理得证. □

1.3 Box划分算法一致性分析

本节将考虑Box划分算法, 并利用定理 1.4证明Box划分算法一致性. Box算法基本思想是将示例空间$\mathcal{X}$划分为多个相同的立方体Box, 然后在每个Box 中使用投票方法. 假设$B_{n1}, B_{n2}, \ldots, B_{nk}, \ldots$是一些列将示例空间$\mathcal{X}$划分成立方体Box, 每个立方体边长为$h_n$. 对每个$X \in \mathcal{X}$, 如果$X \in B_{ni}$, 那么记$A_n(X)=B_{ni}$. 最后Box划分算法分类器定义为

$$
g_n(X)= \begin{cases}0 & \text { if } \sum_{X_i \in A(X)} I[Y_i=1] \leq \sum_{X_i \in A(X)} I[Y_i=0], \\ 1 & \text { otherwise. }\end{cases}
$$

如上述算法, 有如下定理:

定理1.5. 当$n \to \infty$时如果有$h_n \to 0$和$nh_n^d \to \infty$成立, 那么Box划分算法满足一致性.

Proof. 利用定理 1.4证明Box划分算法一致性, 只需要验证定理 1.4两个条件. 对每个划分单元B_{ni}, 由于其边长为h_n可知

$$A(B_{ni}) = \sqrt{d}h_n \to 0 \quad \text{当} \quad h_n \to 0.$$

对定理 1.4的另一个条件, 需要证明对任意的$M > 0$, 有$\Pr[N(X) < M] \to 0$即可. 假设$\mathcal{C}$是一个中心在坐标原点的任意圆, 那么与$\mathcal{C}$ 相交的立方体个数不会超过$c_1 + c_2/h_n^d$, 其中c_1, c_2是常数. 由全概率公式可得

$$\begin{aligned}
\Pr[N(X) < M] &\leq \sum_{i:B_{ni}\cap\mathcal{C}\neq\emptyset} \Pr[X \in B_{ni}, N(X) < M] + \Pr[X \notin \mathcal{C}] \\
&\leq \sum_{\substack{i:B_{ni}\cap\mathcal{C}\neq\emptyset \\ \Pr[B_{ni}]\leq 2M/n}} \Pr[B_{ni}] + \sum_{\substack{i:B_{ni}\cap\mathcal{C}\neq\emptyset \\ \Pr[B_{ni}]>2M/n}} \Pr[B_{ni}] \Pr\left[\sum_{i=1}^n I[X_i \in B_{ni}] < M\right] + \Pr[X \notin \mathcal{C}] \\
&\leq \frac{2M}{n}\left(c_1 + \frac{c_2}{h_n^d}\right) + \sum_{\substack{i:B_{ni}\cap\mathcal{C}\neq\emptyset \\ \Pr[B_{ni}]>2M/n}} \Pr[B_{ni}] \Pr\left[\sum_{i=1}^n I[X_i \in B_{ni}] < M\right] + \Pr[X \notin \mathcal{C}].
\end{aligned}$$

设$\mu(B_{ni}) = E[I[X_i \in B_{ni}]] = \Pr[B_{ni}]$, 由Chebyshev不等式可得

$$\begin{aligned}
&\sum_{\substack{i:B_{ni}\cap\mathcal{C}\neq\emptyset \\ \Pr[B_{ni}]>2M/n}} \Pr[B_{ni}] \Pr\left[\sum_{i=1}^n I[X_i \in B_{ni}] < M\right] \\
&= \sum_{\substack{i:B_{ni}\cap\mathcal{C}\neq\emptyset \\ \mu(B_{ni})>2M/n}} \mu(B_{ni}) \Pr\left[\frac{1}{n}\sum_{i=1}^n I[X_i \in B_{ni}] - \mu(B_{ni}) < -\frac{\mu(B_{ni})}{2}\right] \\
&\leq 4 \sum_{\substack{i:B_{ni}\cap\mathcal{C}\neq\emptyset \\ \mu(B_{ni})>2M/n}} \mu(B_{ni}) \frac{Var(\frac{1}{n}\sum_{i=1}^n I[X_i \in B_{ni}])}{\mu(B_{ni})^2} \leq \frac{4}{n}\left(c_1 + \frac{c_2}{h_n^d}\right).
\end{aligned}$$

当$n \to \infty$时, 有

$$\Pr[N(X) < M] \leq \frac{4+2M}{n}\left(c_1 + \frac{c_2}{h_n^d}\right) + \Pr[X \notin \mathcal{C}] \to \Pr[X \notin \mathcal{C}].$$

由$\mathcal{C}$的任意性, 定理得证. □

1.4 随机森林一致性

本节将利用定理 1.4来研究随机森林算法一致性. 假设示例空间$\mathcal{X} = [0,1]^d$, 以及边缘分布$\mathcal{D}_\mathcal{X}$是空间$\mathcal{X}$的连续非零函数. 假设$S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$是给定的训练数据集. 对于随机树等随机分类器, 用随机变量$Z \in \mathcal{Z}$表示刻画随机分类器的随机过程, 则一个随机分类器可表示为$g_n(X, Z, S_n)$, 通常用$g_n(X, Z)$简单表示.

随机森林则是将m颗随机树分类器$g_n(X, Z_1), g_n(X, Z_2), \ldots, g_n(X, Z_m)$再次进行投票, 由此产生的随机森林分类器可表示为$\bar{g}_n$, 即

$$\bar{g}_n(X, Z) = \begin{cases} 1 & \text{if } \sum_{i=1}^m g_n(X, Z_i) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

首先给出如下性质:

命题1.1. *假设随机树$g_n(X,Z)$满足一致性, 则随机树$g_n(X,Z_1),g_n(X,Z_2),\dots,g_n(X,Z_m)$通过投票所构成的随机森林$\bar{g}_n(X,Z)$也满足一致性.*

上述命题表明, 随机森林一致性研究可以通过随机森林中随机树的一致性来体现, 后面的研究主要集中于随机树一致性分析.

Proof. 因为随机树$g_n(X,Z)$满足一致性, 则有$E[R_{S_n}(g_n(X,Z))]\to R^*$. 对任意给定$X\in\mathcal{X}$, 因为

$$\Pr[g_n(X,Z)\neq Y|X]\geq \Pr[g^*(X,Z)\neq Y|X]=\min(\eta(X),1-\eta(X)),$$

由$E[R_{S_n}(g_n(X,Z))]\to R^*$可得, 对任意的$X\in\mathcal{X}$, 有

$$\Pr[g_n(X,Z)\neq Y|X]\to \Pr[g^*(X,Z)\neq Y|X] \quad \text{几乎处处成立.}$$

对任意给定$X\in\mathcal{X}$, 且满足$\eta(X)>1/2$, 那么有

$$\begin{aligned}\Pr[\bar{g}_n(X,Z)=0] &= \Pr\left[\frac{1}{m}\sum_{i=1}^{m}I[g_n(X,Z_i)=0]>1/2\right]\\ &\leq 2E\left[\frac{1}{m}\sum_{i=1}^{m}I[g_n(X,Z_i)=0]\right]=2\Pr[g_n(X,Z)=0]\to 0.\end{aligned}$$

命题得证. □

这里考虑一种简化版随机树, 其分类器$g_n(X,Z)$构造方式为: 树中所有节点都对应于一个长方体单元格, 所有叶节点的长方体单元格构成空间$\mathcal{X}$的一个划分. 随机树的根节点是空间$\mathcal{X}$本身, 在随机树的每一步构造中, 随机选择一个叶节点, 然后在叶节点随机选择一种划分属性特征(feature), 在所选择的属性特征随机选择一个划分点进行划分, 将上述过程重复k次. 完成划分后, 在每一个长方体单元格内进行投票进行标记预测, 即

$$g_n(X,Z)=\begin{cases}0 & \text{if } \sum_{X_i\in A(X)}I[Y_i=1]\leq\sum_{X_i\in A(X)}I[Y_i=0]\\ 1 & \text{otherwise}\end{cases}$$

这里$A(X)$表示示例X所在的长方体单元格. 于是有如下定理:

定理1.6. *当$n\to\infty$时, 如果有$k=k(n)\to\infty$以及$k/n\to 0$, 那么基于上述随机树的森林算法满足一致性.*

Proof. 根据命题 1.1, 仅需证明随机树$g_n(X,Z)$的一致性即可. 由定理 1.4可知需证明$\mathrm{Diam}(A_n(X,Z))\to 0$和$N_n(X,Z)\to\infty$, 这里$A_n(X,Z)$表示包含$X$的长方体单元格, 而$N_n(X,Z)$表示落入$A_n(X,Z)$中的训练样本数, 即

$$N_n(X,Z)=\sum_{i=1}^{n}I[X_i\in A_n(X,Z)].$$

首先证明$N_n(X,Z) \to \infty$, 随机数经过k论迭代得到$k+1$长方体单元格, 记为$A_1, A_2, \ldots, A_{k+1}$. 假设$N_1, N_2, \ldots, N_{k+1}$ 表示训练集$X_1, X_2, \ldots, X_n$ 落入这些单元格的样本数. 当给定训练集S_n和随机变量Z, 则X落入第i个单元格的条件概率为$N_i/n+1$. 因此对每个固定的$t>0$, 有

$$\Pr[N_n(X,Z)<t] = E[\Pr[N_n(X,Z)<t|S_n,Z]] = E\left[\sum_{i:\ N_i<t} \frac{N_i}{n+1}\right] = (t-1)\frac{k+1}{n+1} \to 0.$$

下面证明$\mathrm{Diam}(A_n(X,Z)) \to 0$. 假设$V_n = V_n(X,Z)$表示包含$X$的单元格中第一维长度, 假设$T_n = T_n(X,Z)$表示包含$X$的单元格被划分的次数. K_n是一个二项分布$B(T_n, 1/d)$表示第一维特征被选中划分的次数. 因此只需要证明$E[V_n(X,Z)] \to 0$即可.

假设$U_1, U_2, \ldots$表示$[0,1]$上的均匀分布, 那么有

$$\begin{aligned} E[V_n(X,Z)] &= E\left[E\left[\left.\prod_{i=1}^{K_n} \max(U_i, 1-U_i)\right| K_n\right]\right] \\ &= E[(E[\max(U_i,1-U_i)])^{K_n}] = E[(3/4)^{K_n}] \\ &= E\left[\left(1-\frac{1}{d}+\frac{3}{4d}\right)^{T_n}\right] = E\left[\left(1-\frac{1}{4d}\right)^{T_n}\right]. \end{aligned}$$

这里使用了

$$E[\max(U_i,1-U_i)] = 2\int_{1/2}^{1} U_i dU_i = 3/4$$

以及

$$\begin{aligned} E[(3/4)^{K_n}] &= \sum_{K_n=1}^{T_n} \left(\frac{3}{4}\right)^{K_n} \binom{T_n}{K_n} \left(\frac{1}{d}\right)^{K_n} \left(1-\frac{1}{d}\right)^{T_n-K_n} \\ &= \sum_{K_n=1}^{T_n} \binom{T_n}{K_n} \left(\frac{3}{4d}\right)^{K_n} \left(1-\frac{1}{d}\right)^{T_n-K_n} \\ &= \left(1-\frac{1}{d}+\frac{3}{4d}\right)^{T_n}. \end{aligned}$$

下面只需要证明$T_n \to \infty$, 首先可以发现$T_n = \sum_{i=1}^{n} \xi_i$, 其中$\xi_i \sim B(1/i)$. 进一步有

$$\begin{aligned} E[T_n] &= \sum_{i=1}^{k} \frac{1}{i} \geq \ln k. \\ V[T_n] &= \sum_{i=2}^{k} \frac{1}{i}\left(1-\frac{1}{i}\right) \leq \ln k + 1. \end{aligned}$$

由Chebyshev不等式可知

$$\Pr[|T_n - E[T_n]| \geq E[T_n]/2] \leq 4V[T_n]/E[T_n]^2 \leq 4(\ln k+1)/\ln^2 k \to 0 \quad \text{当} \quad k\to\infty,$$

从而得到$\Pr[T_n \geq E[T_n]/2] \to 0$. 定理得证. □

2 替代损失函数一致性

2.1 替代损失函数

给定示例空间$\mathcal{X} \subseteq \mathbb{R}^d$以及标记空间$\mathcal{Y} = \{+1, -1\}$, 假设$\mathcal{D}$是空间$\mathcal{X} \times \mathcal{Y}$的一个联合分布. 分布$\mathcal{D}$可分解为在示例空间$\mathcal{X}$的边缘分布$\mathcal{D}_{\mathcal{X}}$ 和条件概率$\eta(x)$, 其中

$$\eta(x) = \Pr[Y = 1|X = x].$$

目标学习一个分类器$g\colon \mathcal{X} \to \mathbb{R}$, 使得其损失的期望尽可能小, 即最小化

$$R(g) = E_{(X,Y)\sim\mathcal{D}}[\ell(g(X), Y)],$$

其中$\ell(\cdot,\cdot)$是损失函数. 例如, 对于常见的二分类问题, 最常见的损失函数是0-1损失函数, 即

$$\ell(g(X), Y) = I[Yg(X) \leq 0] = \begin{cases} 1 & \text{for } Yg(X) \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

这里$g(X) = 0$不需要过多关注.

给定训练数据集$S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, 其中每个样本根据分布$\mathcal{D}$独立采样所得. 机器学习考虑在一个函数空间$\mathcal{H}$中学习一个函数$g$, 使得其最小化在训练数据集$S_n$的分类错误率

$$\frac{1}{n}\sum_{i=1}^{n} I[Y_i g(X_i) \leq 0]. \tag{1}$$

这种方法本质上可以看作随机近似最小化期望分类错误率

$$R(g) = E_{(X,Y)\sim\mathcal{D}}[I[Yg(X) \leq 0]].$$

由于分类错误率$I[\cdot]$本身是非凸与不连续的, 直接优化式 (1)是一个典型的NP-Hard问题, 从而导致在计算上的不可行. 在实际的算法设计过程中, 一般会对分类错误率损失函数$I[\cdot]$进行凸放松, 即对分类错误率损失函数$I[\cdot]$的上界进行凸放松. 例如, AdaBoost算法优化指数损失函数$\ell(g(X), Y) = \exp(-Yg(X))$, 支持向量机(SVMs)方法优化hinge损失函数$\ell(g(X), Y) = \max(0, 1 - Yg(X))$, 等. 为此引入一个新的函数$\phi\colon \mathcal{R} \to \mathcal{R}$, 使得

$$\ell(g(X), Y) = \phi(Yg(X)),$$

这里我们称函数ϕ为替代损失函数, 一般是连续的凸函数, 如AdaBoost和SVMs算法分别对应于$\phi(t) = e^{-t}$和$\phi(t) = \max(0, 1 - t)$.

给定训练数据集$S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ 以及替代损失函数ϕ, 可以优化替代损失函数ϕ在训练集上的平均损失

$$\frac{1}{n}\sum_{i=1}^{n} \phi(Y_i g(X_i)).$$

由于函数ϕ一般情形下是连续的凸函数, 因此各种优化技术可用于上述替代损失函数. 从本质上看, 这种方法可被看作随机近似最小化替代损失函数期望风险

$$R_\phi(g) = E_{(X,Y)\sim\mathcal{D}}[\phi(Yg(X))].$$

本节主要研究下面几种常见的替代损失函数, 相关图形表示如图 1所示.

- 最小二乘损失函数$\phi(t) = (1-t)^2$ (最小二乘支持向量机SVMs方法)
- Hinge损失函数$\phi(t) = \max(0, 1-t)$ (支持向量机SVMs方法)
- 指数损失函数$\phi(t) = e^{-t}$ (Boosting方法)
- 对数损失函数$\phi(t) = \log(1+e^{-t})$ (Logistic回归方法)
- 平方Hinge损失函数$\phi(t) = (\max(0, 1-t))^2$ (平滑支持向量机SVMs方法)

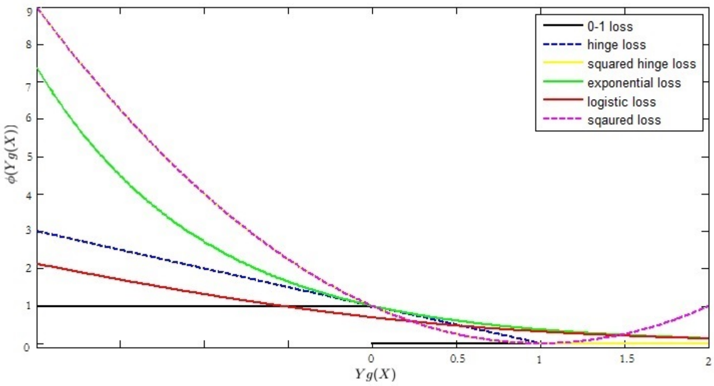

图 1: *0-1*损失函数与各种替代损失函数

本节研究的主要问题是替代损失函数与0-1损失函数之间的一致性问题, 即两类分类问题本质是优化0-1损失函数$I[\cdot]$, 由于其非凸不连续性, 在实际算法设计中往往转而优化替代损失函数ϕ. 替代函数函数一致性研究通过优化替代损失函数ϕ所学习得到的学习器, 是否真正优化优化0-1损失函数$I[\cdot]$?

为了形式化给出替代损失函数一致性定义, 首先需要用到一些重要的记号. 对0-1 损失函数, 学习器g的期望错误率和最优错误率分别表示为

$$R(g) = E_{(X,Y)\sim\mathcal{D}}[I[Yg(X) \le 0]] \quad \text{和} \quad R^* = \min_g [R(g)].$$

对替代损失函数ϕ, 学习器g的期望错误率和最优错误率分别表示为

$$R_\phi(g) = E_{(X,Y)\sim\mathcal{D}}[\phi(Yg(X))] \quad \text{和} \quad R_\phi^* = \min_g [R_\phi(g)].$$

给定训练集$S_n = \{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$, 通过优化替代损失函数

$$\frac{1}{n}\sum_{i=1}^n \phi(Y_i g(X_i))$$

而得到学习器$\hat{g}_n$. 随着训练样本的增加, 可以得到一系列学习器$\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_n, \ldots$. 替代损失函数一致性的形式化定义如下:

定义2.1 (替代损失函数一致性问题)**.** *通过优化替代损失函数所得的学习器$\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_n, \ldots$, 如果$R_\phi(\hat{g}_n) \to R_\phi^*$ $(n \to \infty)$, 则有$R(\hat{g}_n) \to R^*$. 满足这样性质的替代损失函数ϕ称为与0-1分类损失函数具有一致性.*

2.2 一致性理论

对于分类错误率损失函数$I[\cdot]$, 有

$$R(g) = E_{(X,Y)\sim\mathcal{D}}[I[Yg(X) \leq 0] = E_X\left[\eta(X)I[g(X) \leq 0] + (1-\eta(X))I[g(X) \geq 0]\right]$$

因此得到最小分类错误率$R^* = E_X[\min(\eta(X), 1-\eta(X))]$, 以及最优分类器

$$g^* \in \mathcal{G} = \{g^* \colon g^*(X)(\eta(X) - 1/2) > 0\}.$$

这里忽略了$g^*(X) = 0$的详细讨论. 对于替代损失函数ϕ, 有

$$R_\phi(g) = E_{(X,Y)\sim\mathcal{D}}[\phi(Yg(X))] = E_{X\sim\mathcal{D}_\mathcal{X}}[\eta(X)\phi(g(X)) + (1-\eta(X))\phi(-g(X))]$$

根据$R_\phi^* = \inf_g[R_\phi(g)]$, 这里函数$g$为任意可测函数. 因此, 假设$g_\phi^*$满足$R_\phi(g_\phi^*) = R_\phi^*$, 则有

$$\begin{aligned} g_\phi^*(X) &= \arg\min_{\alpha\in\mathcal{R}}[\eta(X)\phi(\alpha) + (1-\eta(X))\phi(-\alpha)] \\ R_\phi^* &= E_{X\sim\mathcal{D}_\mathcal{X}}\left[\min_{\alpha\in\mathcal{R}}[\eta(X)\phi(\alpha) + (1-\eta(X))\phi(-\alpha)]\right]. \end{aligned}$$

基于上述推导可得

引理2.1. *对最小二乘替代函数$\phi(t) = (1-t)^2$, 有$g_\phi^*(X) = 2\eta(X) - 1$, $R_\phi^* = E_X[4\eta(X)(1-\eta(X))]$;*
Hinge替代函数$\phi(t) = \max(0, 1-t)$, 有$g_\phi^*(X) = sgn(2\eta(X) - 1)$, $R_\phi^* = 2E_X[\min(\eta(X), 1-\eta(X))]$;
平方Hinge替代函数$\phi(t) = (\max(0, 1-t))^2$, 有$g_\phi^*(X) = 2\eta(X) - 1$, $R_\phi^* = E_X[4\eta(X)(1-\eta(X))]$;
指数替代函数$\phi(t) = e^{-t}$, 有$g_\phi^*(X) = \frac{1}{2}\ln\frac{\eta(X)}{1-\eta(X)}$, $R_\phi^* = 2E_X[\sqrt{\eta(X)(1-\eta(X))}]$;
对数替代函数$\phi(t) = \log(1+e^{-t})$, 有

$$g_\phi^*(X) = \ln\frac{\eta(X)}{1-\eta(X)}, \quad R_\phi^* = E_X[-\eta(X)\ln\eta(X) - (1-\eta(X))\ln(1-\eta(X))].$$

由引理可知, 对任意给定的数据分布(分布未知), 对任意可测的函数空间, 通过优化最小二乘替代函数, Hinge替代函数, 平方Hinge替代函数, 指数替代函数, 对数替代函数所得到学习器$g_\phi^* \in \mathcal{G}$, 从而优化替代损失函数所获得的最优学习器对0-1损失函数而言也是最优的.

Proof. 这里将给出最小二乘损失函数和对数函数的详细证明, 其他替代损失函数的证明类似可得. 对最小二乘替代损失函数有

$$\begin{aligned} g_\phi^*(X) &= \arg\min_{\alpha\in\mathcal{R}}[\eta(X)\phi(\alpha) + (1-\eta(X))\phi(-\alpha)] \\ &= \arg\min_{\alpha\in\mathcal{R}}[\eta(X)(1-\alpha)^2 + (1-\eta(X))(1+\alpha)^2]. \end{aligned}$$

令$f(\alpha)=\eta(X)(1-\alpha)^2+(1-\eta(X))(1+\alpha)^2$, 求导$f'(\alpha)$并令$f'(\alpha)=0$可得到$\alpha=2\eta(X)-1$. 进一步得到

$$R_\phi^*=E_X[\eta(X)(1-2\eta(X)+1)^2+(1-\eta(X))(1+2\eta(X)-1)^2]=4E_X[\eta(X)(1-\eta(X))].$$

对指数替代损失函数有

$$\begin{aligned}g_\phi^*(X) &= \arg\min_{\alpha\in\mathcal{R}}[\eta(X)\phi(\alpha)+(1-\eta(X))\phi(-\alpha)]\\ &= \arg\min_{\alpha\in\mathcal{R}}[\eta(X)e^{-\alpha}+(1-\eta(X))e^{\alpha}].\end{aligned}$$

令$f(\alpha)=\eta(X)e^{-\alpha}+(1-\eta(X))e^{\alpha}$, 求导$f'(\alpha)$并令$f'(\alpha)=0$可得到$\alpha=\frac{1}{2}\ln\frac{\eta(X)}{1-\eta(X)}$. 进一步得到

$$\begin{aligned}R_\phi^* &= E_X\left[\eta(X)\exp\left(-\frac{1}{2}\ln\frac{\eta(X)}{1-\eta(X)}\right)+(1-\eta(X))\exp\left(\frac{1}{2}\ln\frac{\eta(X)}{1-\eta(X)}\right)\right]\\ &= 2E_X\left[\sqrt{\eta(X)(1-\eta(X))}\right].\end{aligned}$$

引理得证. □

前面的引理表明: 当优化替代损失函数到最优解时所得到的学习器对0-1分类损失函数也是最优的. 但这样的结论局限于最优解, 不能用于函数的趋近过程, 也不能用于替代函数一致性研究, 但下面的定理给出了一致性的充分条件:

定理2.1. *对替代损失函数ϕ, 如果最优替代损失函数满足$g_\phi^*\in\mathcal{G}$, 并存在$c>0$和$s\geq 1$满足*

$$|0.5-\eta(X)|^s\leq c^s(\phi(0)-\eta(X)\phi(g_\phi^*(X))-(1-\eta(X))\phi(-g_\phi^*(X)))\quad \textit{for every}\quad X\in\mathcal{X},$$

那么对任何可测函数g, 有如下关系成立:

$$R(g)-R^*\leq 2c\sqrt[s]{R_\phi(g)-R_\phi^*}.$$

当替代损失函数上述定理条件时, 如果有$R_\phi(g)\to R_\phi^*$, 则有$R(g)\to R^*$, 即当优化替代损失趋于最优时, 通过优化替代函数学习得到的学习器其0-1损失也趋于最优, 从而证明了代替损失函数对0-1损失满足一致性.

Proof. 由前面的定义可知$R(g)-R^*=E_X[\Delta(X)]$, 其中

$$\Delta(X)=\eta(X)I[g(X)\leq 0]+(1-\eta(X))I[g(X)\geq 0]-\min\{\eta(X),1-\eta(X)\}$$

如果$\eta(X)>0.5$且函数$g(X)>0$, 则$\Delta(X)=0$; 如果$\eta(X)>0.5$且函数$g(X)\leq 0$, 则$\Delta(X)=2\eta(X)-1$; 如果$\eta(X)<0.5$且函数$g(X)\geq 0$, 则$\Delta(X)=1-2\eta(X)$; 如果$\eta(X)<0.5$且函数$g(X)>0$, 则$\Delta(X)=0$. 由此可得

$$\begin{aligned}R(g)-R^* &= E_{\eta(X)>0.5,g(X)\leq 0}[2\eta(X)-1]+E_{\eta(X)<0.5,g(X)\geq 0}[1-2\eta(X)]\\ &= 2E_{(\eta(X)-0.5)g(X)\leq 0}[|\eta(X)-0.5|]\\ &\leq 2\sqrt[s]{E_{(\eta(X)-0.5)g(X)\leq 0}[|\eta(X)-0.5|^s]}\end{aligned}$$

最后一个不等式成立是因为Jensen不等式$(E[x])^s \leq E[x^s]$ $(s \geq 1)$. 根据假设可得

$$R(g) - R^* \leq 2c\sqrt{E_{(\eta(X)-0.5)g(X)\leq 0}[\phi(0) - \eta(X)\phi(g_\phi^*(X)) - (1-\eta(X))\phi(-g_\phi^*(X))]}$$

如果能够证明当$(\eta(X)-0.5)g(X) \leq 0$时有$\phi(0) \leq \eta(X)\phi(g(X)) + (1-\eta(X))\phi(-g(X))$成立, 那么有

$$\begin{aligned}
&E_{(\eta(X)-0.5)g(X)\leq 0}[\phi(0) - \eta(X)\phi(g_\phi^*(X)) - (1-\eta(X))\phi(-g_\phi^*(X))] \\
&\leq E_{(\eta(X)-0.5)g(X)\leq 0}[\eta(X)\phi(g(X)) + (1-\eta(X))\phi(-g(X)) - \eta(X)\phi(g_\phi^*(X)) - (1-\eta(X))\phi(-g_\phi^*(X))] \\
&\leq E_X[\eta(X)\phi(g(X)) + (1-\eta(X))\phi(-g(X)) - \eta(X)\phi(g_\phi^*(X)) - (1-\eta(X))\phi(-g_\phi^*(X))] \\
&\leq R_\phi(g) - R_\phi^*
\end{aligned}$$

为了简单起见, 记

$$f(t) = \eta(X)\phi(t) + (1-\eta(X))\phi(-t)$$

由于$\phi(t)$是凸函数, 从而得到$f(t)$也是凸函数, 有凸函数的性质可得到: 如果$0 \in [a,b]$,则有$f(0) \leq \max\{f(a), f(b)\}$成立. 下面分三种情况研究

- 如果$\eta(X) > 0.5$, 那么有$g(X) < 0$和$g_\phi^*(X) > 0$. 由此可得$0 \in [g(X), g_\phi^*(X)]$, 进一步有

$$\phi(0) = f(0) \leq \max\{f(g(X)), f(g_\phi^*(X))\} = f(g(X)) = \eta(X)\phi(g(X)) + (1-\eta(X))\phi(-g(X)).$$

- 如果$\eta(X) < 0.5$, 那么有$g(X) > 0$和$g_\phi^*(X) < 0$, 同理可证.
- 如果$\eta(X) = 0.5$, 由ϕ是凸函数可得

$$\phi(0) = \phi(g(X)/2 - g(X)/2) \leq 1/2\phi(g(X)) + 1/2\phi(-g(X)) = \eta(X)\phi(g(X)) + (1-\eta(X))\phi(-g(X)).$$

定理得证. □

2.3 实例

本节将对最小二乘替代函数, Hinge替代函数, 平方Hinge替代函数, 指数替代函数, 对数替代函数一致性进行研究. 根据定理 2.1可知, 我们需要验证两个条件, 即优化替代函数所得最优分类器$g_\phi^* \in \mathcal{G}$, 以及存在$s \geq 1$和$c > 0$, 使得对每个$X \in \mathcal{X}$, 有

$$|0.5 - \eta(X)|^s \leq c^s(\phi(0) - \eta(X)\phi(g_\phi^*(X)) - (1-\eta(X))\phi(-g_\phi^*(X))).$$

2.3.1 最小二乘替代函数

对最小二乘替代函数$\phi(t) = (1-t)^2$, 由引理 2.1可知$g_\phi^*(X) = 2\eta(X) - 1 \in \mathcal{G}$, 而对每个$X \in \mathcal{X}$有

$$\phi(0) - \eta(X)\phi(g_\phi^*(X)) - (1-\eta(X))\phi(-g_\phi^*(X)) = 1 - 4\eta(X)(1-\eta(X)) = (1-2\eta(X))^2 = 4|0.5-\eta(X)|^2$$

由此可得$c = 1/2$和$s = 2$. 对最小二乘替代函数有

$$R(g) - R^* \leq \sqrt{R_\phi(g) - R_\phi^*}.$$

2.3.2 Hinge替代函数

对于支持向量机(SVMs)的Hinge替代函数$\phi(t)=\max(0,1-t)$, 由引理 2.1可知$g_\phi^*(X)=\mathrm{sgn}(2\eta(X)-1)\in\mathcal{G}$. 对$X\in\mathcal{X}$ 且满足$\eta(X)\geq 1/2$, 则$g_\phi^*(X)=1$, 并且有

$$\phi(0)-\eta(X)\phi(g_\phi^*(X))-(1-\eta(X))\phi(-g_\phi^*(X))=1-2(1-\eta(X))=2|\eta(X)-0.5|;$$

对$X\in\mathcal{X}$ 且满足$\eta(X)<1/2$, 则$g_\phi^*(X)=-1$, 并且有

$$\phi(0)-\eta(X)\phi(g_\phi^*(X))-(1-\eta(X))\phi(-g_\phi^*(X))=1-2\eta(X)=2|\eta(X)-0.5|;$$

由此可得$c=1/2$和$s=1$. 对Hinge替代函数有

$$R(g)-R^*\leq R_\phi(g)-R_\phi^*.$$

2.3.3 平方Hinge替代函数

平方Hinge替代函数$\phi(t)=(\max(0,1-t))^2$与最小二乘替代函数类似, 首先由引理 2.1可知$g_\phi^*(X)=2\eta(X)-1\in\mathcal{G}$, 而对每个$X\in\mathcal{X}$ 有

$$\phi(0)-\eta(X)\phi(g_\phi^*(X))-(1-\eta(X))\phi(-g_\phi^*(X))=1-4\eta(X)(1-\eta(X))=(1-2\eta(X))^2=4|0.5-\eta(X)|^2$$

由此可得$c=1/2$和$s=2$. 对平方Hinge替代函数有

$$R(g)-R^*\leq\sqrt{R_\phi(g)-R_\phi^*}.$$

2.3.4 指数替代函数

对指数替代函数$\phi(t)=e^{-t}$, 由引理 2.1可知$g_\phi^*(X)=\frac{1}{2}\ln\frac{\eta(X)}{1-\eta(X)}\in\mathcal{G}$, 而对每个$X\in\mathcal{X}$有

$$\begin{aligned}
&\phi(0)-\eta(X)\phi(g_\phi^*(X))-(1-\eta(X))\phi(-g_\phi^*(X))\\
&= 1-2\sqrt{\eta(X)(1-\eta(X))}=(\sqrt{\eta(X)}-\sqrt{1-\eta(X)})^2\\
&= \frac{(\eta(X)-(1-\eta(X)))^2}{(\sqrt{\eta(X)}+\sqrt{1-\eta(X)})^2}=\frac{2|\eta(X)-0.5|^2}{1+2\sqrt{\eta(X)(1-\eta(X))}}\\
&\geq |\eta(X)-0.5|^2
\end{aligned}$$

由此可得$c=1$和$s=2$. 对最小二乘替代函数有

$$R(g)-R^*\leq 2\sqrt{R_\phi(g)-R_\phi^*}.$$

2.3.5 对数替代函数

对对数替代函数$\phi(t)=\ln(1+e^{-t})$, 由引理 2.1可知$g_\phi^*(X)=\ln\frac{\eta(X)}{1-\eta(X)}\in\mathcal{G}$, 而对每个$X\in\mathcal{X}$ 有

$$\phi(0)-\eta(X)\phi(g_\phi^*(X))-(1-\eta(X))\phi(-g_\phi^*(X))=\ln 2+\eta(X)\ln\eta(X)+(1-\eta(X))\ln(1-\eta(X)).$$

设

$$f(t)=\ln 2+t\ln t+(1-t)\ln(1-t) \quad \text{for} \quad t\in(0,1).$$

有$f(1/2)=0$, 求一阶导数得到

$$f'(t)=\ln t-\ln(1-t) \quad \text{以及} \quad f'(1/2)=0.$$

对函数$f(t)$求二阶导数可得

$$f''(t)=1/t(1-t)\geq 4 \quad \text{for} \quad t\in(0,1)$$

根据泰勒展开式可知存在$t_0\in(0,1)$使得下式成立

$$f(t)=f(1/2)+f'(1/2)(t-1/2)+f''(t_0)(t-1/2)^2/2\geq 2(t-0.5)^2,$$

即

$$\phi(0)-\eta(X)\phi(g_\phi^*(X))-(1-\eta(X))\phi(-g_\phi^*(X))\geq 2(\eta(X)-0.5)^2.$$

于是得到$s=2$和$c=1/\sqrt{2}$, 对对数替代损失函数有

$$R(g)-R^*\leq\sqrt{2(R_\phi(g)-R_\phi^*)}.$$

2.4 优化AUC替代函数一致性

给定示例空间$\mathcal{X}\subseteq\mathbb{R}^d$以及标记空间$\mathcal{Y}=\{0,1\}$, 假设$\mathcal{D}$是空间$\mathcal{X}\times\mathcal{Y}$的一个联合分布. 分布$\mathcal{D}$可分解为在示例空间$\mathcal{X}$的边缘分布$\mathcal{D}_\mathcal{X}$ 和条件概率$\eta(x)=\Pr[Y=1|X=x]$. 给定一个实值函数$g\colon\mathcal{X}\to\mathbb{R}$, 其在分布$\mathcal{D}$上的AUC定义为

$$\mathrm{AUC}(g)=E_{(X,Y),(X',Y')\sim\mathcal{D}}[I[(Y-Y')g(X)-g(X')>0]+\tfrac{1}{2}I[g(X)=g(X')]|Y\neq Y'].$$

最大化AUC等价于最小化如下期望排序错误率

$$\begin{aligned} R(g) &= E_{(X,Y),(X',Y')\sim\mathcal{D}}[I[(Y-Y')g(X)-g(X')<0]+\tfrac{1}{2}I[g(X)=g(X')]|Y\neq Y'] \\ &= E_{X,X'}[\eta(X)(1-\eta(X'))\ell(g,X,X')+\eta(X')(1-\eta(X))\ell(g,X',X)] \end{aligned} \tag{2}$$

其中$\ell(g,X,X')=I[g(X)<g(X')]+I[g(X)=g(X')]/2$. 记$R^*=\inf_g[R(g)]$, 有

习题2.1. 试证明最优排序函数所构成的函数集合为

$$\begin{aligned} \mathcal{B} &= \{g\colon R(g)=R^*\} \\ &= \{g\colon (g(X)-g(X'))(\eta(X)-\eta(X'))>0 \ \mathit{if}\ \eta(X)\neq\eta(X')\}. \end{aligned} \tag{3}$$

相同的道理, 上述$0-1$损失函数是非凸不连续的, 在实际应用中优化如下替代损失函数

$$\Psi(g,X,X')=\phi(g(X)-g(X'))$$

其中ϕ是连续的凸函数, 例如指数替代函数$\phi(t) = e^{-t}$(RankBoost算法), Hinge替代函数$\phi(t) = \max(0, 1-t)$(RankSVM算法), 等. 给定实值函数$g\colon \mathcal{X} \to \mathbb{R}$, 其替代损失函数的期望为

$$R_\phi(g) = E_{X,X'}[\eta(X)(1-\eta(X'))\phi(g(X)-g(X')) + \eta(X')(1-\eta(X))\phi(g(X')-g(X))]$$

以及定义最优替代函数期望$R_\phi^* = \inf_g R_\phi(g)$. 下面给出一致性定义

定义2.2. 对任意给定分布$\mathcal{D}$以及任何函数序列$\{g^{\langle n\rangle}(X)\}_{n\geq 1}$, 如果有

$$R_\phi(g^{\langle n\rangle}) \to R_\phi^*, \text{ 那么} R(g^{\langle n\rangle}) \to R^*,$$

满足这样性质的替代损失函数ϕ称为对AUC具有一致性.

下面给出满足AUC一致性的一个充分条件:

定理2.2. 如果函数$\phi\colon \mathbb{R} \to \mathbb{R}$是可导的, 非单调递减的凸函数, 并且满足$\phi'(0) < 0$, 那么替代函数$\Psi(g, X, X') = \phi(g(X) - g(X'))$与$AUC$具有一致性.

由此定理可知, 指数替代函数$\phi(t) = e^{-t}$, 对数替代函数$\phi(t) = \ln(1+e^{-t})$, 平方Hinge替代函数$\phi(t) = (\max(0, 1-t))^2$与AUC具有一致性, 但Hinge替代函数$\phi(t) = \max(0, 1-t)$不能使用满足上述定理, 因为在$t = 1$不可导.

在证明定理 2.2之前, 先引入一个重要的引理:

引理2.2. 如果函数$\phi\colon \mathbb{R} \to \mathbb{R}$是可导的, 非单调递减的凸函数, 并且满足$\phi'(0) < 0$, 那么有

$$\inf_{g \notin \mathcal{B}} R_\phi(g) > \inf_g R_\phi(g).$$

Proof. 由定义可知

$$R_\phi(g) = \int_{\mathcal{X}}\int_{\mathcal{X}} \eta(X)(1-\eta(X'))\phi(g(X)-g(X')) + \eta(X')(1-\eta(X))\phi(g(X')-g(X)) d\Pr(X) d\Pr(X')$$

我们将通过反正法证明该引理, 假设

$$\inf\nolimits_{g\notin\mathcal{B}} R_\phi(g) = \inf\nolimits_g R_\phi(g),$$

那么将存在函数g^*满足$R_\phi(g^*) = \inf_g R_\phi(g)$但$g^* \notin \mathcal{B}$, 即存在$X_1, X_2 \in \mathcal{X}$, 有$g^*(X_1) \leq g^*(X_2)$但$\eta(X_1) > \eta(X_2)$.

我们引入一个新的函数$h_1\colon \mathcal{X} \to 0, 1$, 满足$h_1(X_1) = 1$, 以及当$X \neq X_1$时$h_1(X) = 0$. 同时令$f(\gamma) = R_\phi(g^* + \gamma h_1)$, 易知函数$f$ 是凸的, 且由g^*的最优性, 我们有$f'(0) = 0$成立, 即

$$\int_{\mathcal{X}\setminus X_1} \eta(X_1)(1-\eta(X))\phi'(f^*(X_1) - f^*(X)) - \eta(X)(1-\eta(X_1))\phi'(f^*(X) - f^*(X_1)) d\Pr(X) = 0. \quad (4)$$

同理引入另一个函数h_2, 满足$h_2(X_2) = 1$, 以及当$X \neq X_2$时$h_2(X) = 0$. 同理有

$$\int_{\mathcal{X}\setminus X_2} \eta(X_2)(1-\eta(X))\phi'(f^*(X_2) - f^*(X)) - \eta(X)(1-\eta(X_2))\phi'(f^*(X) - f^*(X_2)) d\Pr(X) = 0. \quad (5)$$

结合式 (4)和式 (5)可得

$$\begin{aligned}\int_{\mathcal{X}\backslash\{X_1,X_2\}} &\eta(X)\big((1-\eta(X_2))\phi'(f^*(X)-f^*(X_2))-(1-\eta(X_1))\phi'(f^*(X)-f^*(X_1))\big)\\ &+(1-\eta(X))\big(\eta(X_1)\phi'(f^*(X_1)-f^*(X))-\eta(X_2)\phi'(f^*(X_2)-f^*(X))\big)d\Pr(X)+(\Pr(X_1)+\Pr(X_2))\\ &\times\big(\eta(X_1)(1-\eta(X_2))\phi'(f^*(X_1)-f^*(X_2))-\eta(X_2)(1-\eta(X_1))\phi'(f^*(X_2)-f^*(X_1))\big)=0. \quad (6)\end{aligned}$$

对非单调递减, 可导的凸函数ϕ, 当$t_1 \le t_2$时有$\phi'(t_1) \le \phi'(t_2) \le 0$成立. 因此, 当$f^*(X_1) \le f^*(X_2)$时有$\phi'(f^*(X_1)-f^*(X)) \le \phi'(f^*(X_2)-f^*(X)) \le 0$. 当$\eta(X_1)>\eta(X_2)$有

$$\eta(X_1)\phi'(f^*(X_1)-f^*(X))-\eta(X_2)\phi'(f^*(X_2)-f^*(X)) \le 0. \quad (7)$$

同理可得

$$(1-\eta(X_2))\phi'(f^*(X)-f^*(X_2))-(1-\eta(X_1))\phi'(f^*(X)-f^*(X_1)) \le 0. \quad (8)$$

如果$f^*(X_1)=f^*(X_2)$, 由$\phi'(0)<0$和$\eta(X_1)>\eta(X_2)$, 我们有

$$\begin{aligned}&\eta(X_1)(1-\eta(X_2))\phi'(f^*(X_1)-f^*(X_2))\\ &\qquad -\eta(X_2)(1-\eta(X_1))\phi'(f^*(X_2)-f^*(X_1))=(\eta(X_1)-\eta(X_2))\phi'(0)<0\end{aligned}$$

但联系式 (7)和式 (8)可以得到与式 (6)矛盾.

如果$f^*(X_1)<f^*(X_2)$, 我们有$\phi'(f^*(X_1)-f^*(X_2)) \le \phi'(0)<0$, 以及$\phi'(f^*(X_1)-f^*(X_2)) \le \phi'(f^*(X_2)-f^*(X_1)) \le 0$, 从而得到

$$\eta(X_1)(1-\eta(X_2))\phi'(f^*(X_1)-f^*(X_2))-\eta(X_2)(1-\eta(X_1))\phi'(f^*(X_2)-f^*(X_1))<0$$

同理与式 (6)矛盾. 引理得证. □

证明定理 2.2 由引理 2.2, 设

$$\delta=\inf_{g\notin\mathcal{B}} R_\phi(g)-\inf_g R_\phi(g)>0.$$

假设$\{g^{\langle n\rangle}\}_{n\ge 0}$是任意一个序列, 满足$R_\phi(g^{\langle n\rangle}) \to R_\phi^*$. 那么存在$N_0>0$使得

$$R_\phi(g^{\langle n\rangle})-R_\phi^*<\delta/2 \text{ for } n\ge N_0.$$

从而立即得到当$n\ge N_0$时有$g^{\langle n\rangle}\in\mathcal{B}$, 这是由下面的矛盾所致

$$R_\phi(g^{\langle n\rangle})-R_\phi^*=R_\phi(g^{\langle n\rangle})-\inf_{g'\notin\mathcal{B}}R_\phi(g')+\inf_{g'\notin\mathcal{B}}R_\phi(g')-R_\phi^*>\delta \text{ if } f\notin\mathcal{B}.$$

因此当$n\ge N_0$时有$R(f^{\langle n\rangle})=R^*$成立, 定理得证. □