$$egin{aligned} P(H(oldsymbol{x}) 
eq f(oldsymbol{x})) &= \sum_{k=0}^{\lfloor T/2 
floor} inom{T}{k} (1-\epsilon)^k \epsilon^{T-k} \ &\leqslant \expigg(-rac{1}{2}T(1-2\epsilon)^2igg) \end{aligned}$$

[推导]:由基分类器相互独立,设X为T个基分类器分类正确的次数,因此 $X \sim B(T, 1-\epsilon)$ 

$$\begin{split} P(H(x) \neq f(x)) = & P(X \leq \lfloor T/2 \rfloor) \\ & \leqslant P(X \leq T/2) \\ & = P \left[ X - (1 - \varepsilon)T \leqslant \frac{T}{2} - (1 - \varepsilon)T \right] \\ & = P \left[ X - (1 - \varepsilon)T \leqslant -\frac{T}{2}(1 - 2\varepsilon) \right] \end{split}$$

根据Hoeffding不等式 $P(X-(1-\epsilon)T\leqslant -kT)\leq \exp(-2k^2T)$  令 $k=\frac{(1-2\epsilon)}{2}$ 得

$$egin{aligned} P(H(oldsymbol{x}) 
eq \int_{k=0}^{\lfloor T/2 
floor} \left( egin{aligned} T \ k \end{aligned} 
ight) (1-\epsilon)^k \epsilon^{T-k} \ &\leqslant \expigg( -rac{1}{2}T(1-2\epsilon)^2 igg) \end{aligned}$$

## 8.5-8.8

[推导]:由式(8.4)可知  $H(oldsymbol{x}) = \sum_{t=1}^T lpha_t h_t(oldsymbol{x})$ 

又由式(8.11)可知

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

该分类器的权重只与分类器的错误率负相关(即错误率越大,权重越低)

(1)先考虑指数损失函数 $e^{-f(x)H(x)}$ 的含义:f为真实函数,对于样本x来说, $f(x) \in \{-1,+1\}$ 只能取和两个值,而 H(x)是一个实数;当H(x)的符号与f(x)一致时,f(x)H(x)>0,因此 $e^{-f(x)H(x)}=e^{-|H(x)|}<1$ ,且|H(x)|越大指数损失函数 $e^{-f(x)H(x)}$ 越小(这很合理:此时|H(x)|越大意味着分类器本身对预测结果的信心越大,损失应该越小;若|H(x)|在零附近,虽然预测正确,但表示分类器本身对预测结果信心很小,损失应该较大); 当H(x)的符号与f(x)不一致时,f(x)H(x)<0,因此 $e^{-f(x)H(x)}=e^{|H(x)|}>1$ ,且|H(x)|越大指数损失函数越大(这很合理:此时|H(x)|越大意味着分类器本身对预测结果的信心越大,但预测结果是错的,因此损失应该越大;若|H(x)|在零附近,虽然预测错误,但表示分类器本身对预测结果信心很小,虽然错了,损失应该较小); (2)符号 $\mathbb{E}_{x\sim\mathcal{D}}[\cdot]$ 的含义: $\mathcal{D}$ 为概率分布,可简单理解为在数据集 $\mathcal{D}$ 中进行一次随机抽样,每个样本被取到的概率; $\mathbb{E}[\cdot]$ 为经典的期望,则综合起来 $\mathbb{E}_{x\sim\mathcal{D}}[\cdot]$ 表示在概率分布 $\mathcal{D}$ 上的期望,可简单理解为对数据集 $\mathcal{D}$ 以概率 $\mathcal{D}$ 进行加权后的期望。

$$egin{aligned} \ell_{ ext{exp}}(H|\mathcal{D}) = & \mathbb{E}_{oldsymbol{x}\sim\mathcal{D}}\left[e^{-f(oldsymbol{x})H(oldsymbol{x})}
ight] \ = & P(f(x) = 1|x) * e^{-H(x)} + P(f(x) = -1|x) * e^{H(x)} \end{aligned}$$

由于P(f(x) = 1|x)和P(f(x) = -1|x)为常数

故式(8.6)可轻易推知

$$rac{\partial \ell_{\mathrm{exp}}(H|\mathcal{D})}{\partial H(oldsymbol{x})} = -e^{-H(oldsymbol{x})}P(f(oldsymbol{x}) = 1|oldsymbol{x}) + e^{H(oldsymbol{x})}P(f(oldsymbol{x}) = -1|oldsymbol{x})$$

令式(8.6)等于0可得

式(8.7)

$$H(oldsymbol{x}) = rac{1}{2} \mathrm{ln} \, rac{P(f(x) = 1 | oldsymbol{x})}{P(f(x) = -1 | oldsymbol{x})}$$

式(8.8)显然成立

$$\begin{split} \operatorname{sign}(H(\boldsymbol{x})) &= \operatorname{sign} \left( \frac{1}{2} \ln \frac{P(f(x) = 1 | \boldsymbol{x})}{P(f(x) = -1 | \boldsymbol{x})} \right) \\ &= \begin{cases} 1, & P(f(x) = 1 | \boldsymbol{x}) > P(f(x) = -1 | \boldsymbol{x}) \\ -1, & P(f(x) = 1 | \boldsymbol{x}) < P(f(x) = -1 | \boldsymbol{x}) \end{cases} \\ &= \underset{y \in \{-1,1\}}{\operatorname{arg max}} P(f(x) = y | \boldsymbol{x}) \end{split}$$

## 8.16

$$egin{aligned} h_t(oldsymbol{x}) &= rg\max_h & \mathbb{E}_{oldsymbol{x} \sim \mathcal{D}} \left[ rac{e^{-f(oldsymbol{x})H_{t-1}(oldsymbol{x})}}{\mathbb{E}_{oldsymbol{x} \sim \mathcal{D}} \left[ e^{-f(oldsymbol{x})H_{t-1}(oldsymbol{x})} 
ight]} f(oldsymbol{x}) h(oldsymbol{x}) 
ight] \ &= rg\max_{oldsymbol{h}} & \mathbb{E}_{oldsymbol{x} \sim \mathcal{D}_t} [f(oldsymbol{x}) h(oldsymbol{x})] \end{aligned}$$

[推导]: 假设x的概率分布是f(x) (注:本书中概率分布全都是 $\mathcal{D}(x)$ )

$$\mathbb{E}(\mathrm{g}(\mathrm{x})) = \sum_{i=1}^{|D|} f(x) g(x)$$

故可得

$$\mathbb{E}_{oldsymbol{x} \sim \mathcal{D}}\left[e^{-f(oldsymbol{x})H(oldsymbol{x})}
ight] = \sum_{i=1}^{|D|} \mathcal{D}\left(oldsymbol{x}_i
ight)e^{-f(oldsymbol{x}_i)H(oldsymbol{x}_i)}$$

由式(8.15)可知

$$\mathcal{D}_{t}\left(oldsymbol{x}_{i}
ight) = \mathcal{D}\left(oldsymbol{x}_{i}
ight) rac{e^{-f\left(oldsymbol{x}_{i}
ight)H_{t-1}\left(oldsymbol{x}_{i}
ight)}}{\mathbb{E}_{oldsymbol{x}\sim\mathcal{D}}\left[e^{-f\left(oldsymbol{x}
ight)H_{t-1}\left(oldsymbol{x}
ight)}
ight]}$$

所以式(8.16)可以表示为

$$egin{aligned} \mathbb{E}_{oldsymbol{x}\sim\mathcal{D}}\left[rac{e^{-f(oldsymbol{x})H_{t-1}(oldsymbol{x})}}{\mathbb{E}_{oldsymbol{x}\sim\mathcal{D}}\left[e^{-f(oldsymbol{x})H_{t-1}(oldsymbol{x})}
ight]}f(oldsymbol{x})h(oldsymbol{x})
ight] \ =&\sum_{i=1}^{|D|}\mathcal{D}\left(oldsymbol{x}_{i}
ight)rac{e^{-f(oldsymbol{x}_{i})H_{t-1}(oldsymbol{x}_{i})}}{\mathbb{E}_{oldsymbol{x}\sim\mathcal{D}}\left[e^{-f(oldsymbol{x})H_{t-1}(oldsymbol{x})}
ight]}f(oldsymbol{x}_{i})h(oldsymbol{x}_{i}) \ =&\sum_{i=1}^{|D|}\mathcal{D}_{t}\left(oldsymbol{x}_{i}
ight)f\left(oldsymbol{x}_{i}
ight)h\left(oldsymbol{x}_{i}
ight) \end{array} =& \mathbb{E}_{oldsymbol{x}\sim\mathcal{D}_{t}}\left[f(oldsymbol{x})h(oldsymbol{x})
ight] \end{aligned}$$

【注】:由下式(\*)也可推至式(8.16)

$$P(f(x) = 1|x)e^{-H(x)} + P(f(x) = -1|x)e^{H(x)}(*)$$

首先式(\*)可以拆成n个式子,n的个数为x的取值个数

$$P(f(x_i) = 1 | x_i) e^{-H(x_i)} + P(f(x_i) = -1 | x_i) e^{H(x_i)} (i = 1, 2, \dots, n) (**)$$

当 $x_i$ 确定的时候  $P(f(x_i=1|x_i))$ 与 $P(f(x_i=-1|x_i))$  其中有一个为0 , 另一个为1

则式(\*\*)可以化简成

$$e^{-f(x_i)H(x_i)} (i=1,2,\ldots,n) (***)$$

拆成n个式子是根据不同的x来拆分的,可以把 $x = x_i$ 看成一个事件,设为事件 $A_i$ 。

当事件 $A_i$ 发生时,事件 $A_j$ 一定不发生,即各事件互斥,而且各个事件发生的概率是 $P(A_i)=\mathcal{D}(x_i)$ 

此时可以考虑成原来的x被分成了n叉树,每个路径的概率是 $\mathcal{D}(x_i)$ ,叶子结点的值是 $e^{-f(x_i)H(x_i)}$ 相乘再相加即为期望,同式(8.16)