8.3

$$
\begin{aligned}
P(H(\boldsymbol{x}) \neq f(\boldsymbol{x})) &= \sum_{k=0}^{\lfloor T/2 \rfloor} \binom{T}{k} (1-\epsilon)^k \epsilon^{T-k} \\
&\leqslant \exp\left(-\frac{1}{2} T(1-2\epsilon)^2\right)
\end{aligned}
$$

[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$\mathrm{X} \sim \mathrm{B}(\mathrm{T}, 1-\epsilon)$

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

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

$$
\begin{aligned}
P(H(\boldsymbol{x}) \neq f(\boldsymbol{x})) &= \sum_{k=0}^{\lfloor T/2 \rfloor} \binom{T}{k} (1-\epsilon)^k \epsilon^{T-k} \\
&\leqslant \exp\left(-\frac{1}{2} T(1-2\epsilon)^2\right)
\end{aligned}
$$

8.5-8.8

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

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，$f(\boldsymbol{x}) \in \{-1,+1\}$只能取和两个值，而$H(\boldsymbol{x})$是一个实数； 当$H(\boldsymbol{x})$的符号与$f(x)$一致时，$f(\boldsymbol{x})H(\boldsymbol{x})>0$，因此$e^{-f(\boldsymbol{x})H(\boldsymbol{x})}=e^{-|H(\boldsymbol{x})|}<1$，且$|H(\boldsymbol{x})|$越大指数损失函数$e^{-f(\boldsymbol{x})H(\boldsymbol{x})}$越小（这很合理：此时$|H(\boldsymbol{x})|$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$|H(\boldsymbol{x})|$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$H(\boldsymbol{x})$的符号与$f(\boldsymbol{x})$不一致时，$f(\boldsymbol{x})H(\boldsymbol{x})<0$，因此$e^{-f(\boldsymbol{x})H(\boldsymbol{x})}=e^{|H(\boldsymbol{x})|}>1$，且$|H(\boldsymbol{x})|$越大指数损失函数越大（这很合理：此时$|H(\boldsymbol{x})|$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$|H(\boldsymbol{x})|$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$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\cdot]$的含义：$\mathcal{D}$为概率分布，可简单理解为在数据集$D$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$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\cdot]$表示在概率分布$\mathcal{D}$上的期望，可简单理解为对数据集$D$以概率$\mathcal{D}$进行加权后的期望。

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

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

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

令式(8.6)等于0可得

式(8.7)

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

式(8.8)显然成立

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

8.16

$$\begin{aligned}
h_t(\boldsymbol{x})&=\underset{h}{\arg\max}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\frac{e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}f(\boldsymbol{x})h(\boldsymbol{x})\right]\\
&=\underset{\boldsymbol{h}}{\arg\max}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[f(\boldsymbol{x})h(\boldsymbol{x})]
\end{aligned}$$

[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

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

由式(8.15)可知

$$\mathcal{D}_t\left(\boldsymbol{x}_i\right)=\mathcal{D}\left(\boldsymbol{x}_i\right)\frac{e^{-f(\boldsymbol{x}_i)H_{t-1}(\boldsymbol{x}_i)}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}$$

所以式(8.16)可以表示为

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\frac{e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}f(\boldsymbol{x})h(\boldsymbol{x})\right]\\
=&\sum_{i=1}^{|D|}\mathcal{D}\left(\boldsymbol{x}_i\right)\frac{e^{-f(\boldsymbol{x}_i)H_{t-1}(\boldsymbol{x}_i)}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}f(x_i)h(x_i)\\
=&\sum_{i=1}^{|D|}\mathcal{D}_t\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_i\right)h\left(\boldsymbol{x}_i\right)\\
=&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[f(\boldsymbol{x})h(\boldsymbol{x})]
\end{aligned}$$

【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

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

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)