IML 第四次作业

习题 8.2

损失函数 $\ell(-f(x)H(x)) = \ell(-H(x))P(f(x)=1|x) + \ell(H(x))P(f(x)=-1|x)$

当 $P(f(x)=1|x) > P(f(x)=-1|x)$ 时，要使 ℓ 最小，必须 $\ell(-H(x)) < \ell(H(x))$，若 ℓ 关于 $H(x)$ 在 $[-\infty,\delta]$ 上递减，则 ℓ 关于 $-H(x)$ 在 $[-\delta,\infty]$ 上递增，则 $\ell(-H(x)) < \ell(H(x)) \Rightarrow \ell(-H(x)) < \ell(-(-H(x))) \Rightarrow -H(x) < H(x) \Rightarrow H(x) > 0$，这与 $P(f(x)=1|x) > P(f(x)=-1|x)$ 一致。

当 $P(f(x)=-1|x) > P(f(x)=1|x)$ 时，要使 ℓ 最小，必须 $\ell(-H(x))\ell(H(x))$，若 ℓ 关于 $H(x)$ 在 $[-\infty,\delta]$ 上递减，则 $\ell(-H(x)) < \ell(H(x)) \Rightarrow -H(x) > H(x) \Rightarrow H(x) < 0$，这与 $P(f(x)=-1|x) > P(f(x)=1|x)$ 一致。

所以当 ℓ 最小化时，分类错误率也最小化，说明 ℓ 是分类任务原本 0/1 损失函数的一致替代损失函数。

以下是按照习题课讲义的修改：

$$
\begin{aligned}
L(H \mid D) &= \mathbb{E}_{\boldsymbol{x},y}[\ell(-yH(\boldsymbol{x}))] \\
&= \mathbb{E}_{\boldsymbol{x},y}[\mathbb{I}(y=1)\ell(-H(\boldsymbol{x})) + \mathbb{I}(y=-1)\ell(H(\boldsymbol{x}))] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{y|\boldsymbol{x}}\mathbb{I}(y=1)\ell(-H(\boldsymbol{x})) + \mathbb{E}_{y}\mathbb{I}(y=-1)\ell(H(\boldsymbol{x}))\right] \\
&= \mathbb{E}_{\boldsymbol{x}}[P(y=1 \mid \boldsymbol{x})\ell(-H(\boldsymbol{x})) + P(y=-1 \mid \boldsymbol{x})\ell(H(\boldsymbol{x}))] \\
&= \int_x P(x)[P(y=1 \mid \boldsymbol{x})\ell(-H(\boldsymbol{x})) + P(y=-1 \mid \boldsymbol{x})\ell(H(\boldsymbol{x}))]
\end{aligned}
$$

因此

$$
\frac{\partial L(H \mid D)}{\partial H(\boldsymbol{x})} = P(\boldsymbol{x})(-P(y=1 \mid \boldsymbol{x})\ell(-H(\boldsymbol{x})) + P(y=-1 \mid \boldsymbol{x})\ell(H(\boldsymbol{x}))) = 0
$$

从而可得

$$
P(y=1 \mid \boldsymbol{x})\ell(-H(\boldsymbol{x})) = P(y=-1 \mid \boldsymbol{x})\ell(H(\boldsymbol{x}))
$$

若 $P(y=1 \mid \boldsymbol{x}) > P(y=-1 \mid \boldsymbol{x})$, 则有 $\ell(-H(\boldsymbol{x})) < \ell(H(\boldsymbol{x}))$ 由于单调性可知 $-H(\boldsymbol{x}) < H(\boldsymbol{x})$，从而 $\operatorname{sign}(H(x)) = 1$.

习题 8.8

MultiBoosting 优点：有效降低误差和方差。缺点：训练成本和预测成本高。

Iterative Bagging 优点：降低误差。缺点：增大方差。由于 Bagging 本身就是一种降低方差的算法，所以 Iterative Bagging 相当于 Bagging 与单分类器的折中。

作业 9.1

假设取二维向量 $e_1=(1{,}0)$,$e_2=(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$,$e_3=(0{,}1)$，则

$$D_1(e_1,e_2)=1-\frac{1}{\sqrt{2}}$$
$$D_1(e_2,e_3)=1-\frac{1}{\sqrt{2}}$$
$$D_1(e_1,e_3)=1$$

但是 $D_1(e_1,e_2)+D_1(e_2,e_3)<D_1(e_1,e_3)$，说明 D_1（余弦距离）不具有传递性。

任取三个不同的单位向量 x,y,z，即余弦夹角函数为 D_2，则

$$\theta_1:=D_2(x,y)=\arccos\left(\frac{x^Ty}{|x||y|}\right)=\arccos\left(x^Ty\right)$$
$$\theta_2:=D_2(y,z)=\arccos\left(\frac{y^Tz}{|y||z|}\right)=\arccos\left(y^Tz\right)$$
$$\theta_3:=D_2(x,z)=\arccos\left(\frac{x^Tz}{|x||z|}\right)=\arccos\left(x^Tz\right)$$

注意到 $A=(x,y,z)^T(x,y,z)\geq 0$，所以其行列式

$$\det(A)=1+2(x^Ty)(y^Tz)(z^Tx)-(x^Ty)^2-(y^Tz)^2-(x^Tz)^2\geq 0$$

即

$$\begin{aligned}
&1+2\cos\theta_1\cos\theta_2\cos\theta_3-\cos^2\theta_1-\cos^2\theta_2-\cos^2\theta_1\geq 0\\
&\Rightarrow\left(\cos\theta_1\cos\theta_2-\cos\theta_3\right)^2\leq\left(1-\cos^2\theta_1\right)\left(1-\cos^2\theta_2\right)\\
&\Rightarrow\sin^2\theta_1\sin^2\theta_2\geq\left(\cos\theta_1\cos\theta_2-\cos\theta_3\right)^2\\
&\Rightarrow\sin\theta_1\sin\theta_2\geq|\cos\theta_1\cos\theta_2-\cos\theta_3|\geq\cos\theta_1\cos\theta_2-\cos\theta_3\\
&\Rightarrow\cos\theta_3\geq\cos\theta_1\cos\theta_2-\sin\theta_1\sin\theta_2\\
&\Rightarrow\cos\theta_3\geq\cos(\theta_1+\theta_2)
\end{aligned}$$

若 $\theta_1+\theta_2\leq\pi$，因为 $y=\cos x$ 在 $[0,\pi]$ 单调递减，所以 $\theta_3\leq\theta_1+\theta_2$；若 $\pi<\theta_1+\theta_2\leq 2\pi$，因为 $\theta_3<\pi$，显然 $\theta_3\leq\theta_1+\theta_2$ 也成立，所以传递性成立。

作业 9.2

损失函数 $E=\sum_{i=1}^k\sum_{x\in C_i}||x-\mu_i||_2^2$，则

$$\frac{\partial E}{\partial\mu_i}=2\sum_{x\in C_i}(x-\mu_i)=0\Rightarrow\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x=\mu_i'$$

说明 μ_i' 是损失函数的极值点，说明每次更新中心点为均值向量时，都会让 E 严格减小，由 $E\geq 0$ 有界，所以最后一定会收敛。

作业 9.3

首先样本 x_j 与各均值向量 μ_i 的距离 d_{ji} 要用新的度量 $dist(x,\mu_j)$ 重新计算。

其次对于损失函数 $E=\sum_{i=1}^{k}\sum_{x\in C_i} dist(x,\mu_i)$，由 $\frac{\partial E}{\partial \mu_i}(i=1,...,k)$ 给出每个 μ_i' 的更新公式，即由 $\sum_{x\in C_i} dist'(x,\mu_i)=0$ 反解出的 μ_i。