**8.2** 对于 0/1 损失函数来说, 指数损失函数并非仅有的一致替代函数. 考虑式(8.5), 试证明: 任意损失函数 $\ell(-f(\boldsymbol{x})H(\boldsymbol{x}))$, 若对于 $H(\boldsymbol{x})$ 在区间 $[-\infty, \delta]$ $(\delta > 0)$ 上单调递减, 则 $\ell$ 是 0/1 损失函数的一致替代函数.

$\because$ $L(-f(x)H(x))$ 关于 $f(x)H(x)$ 在 $[-\infty, \delta]$ $(\delta > 0)$ 上单减，

$\therefore$ $L$ 的最小值所对应的横坐标位置在 $\delta$ 右侧，即最小值点 $u^* = f(x^*)H(x^*) \geq \delta > 0$

即 $f(x^*)$ $H(x^*)$ 同正负，$\text{sign}(H(x^*)) = f(x^*) = \underset{y \in \{-1,1\}}{\arg\max} P(f(x) = y \mid x)$

即当指数损失函数最小化，则分类错误率也将最小化

**8.8** MultiBoosting 算法 [Webb, 2000] 将 AdaBoost 作为 Bagging 的基学习器, Iterative Bagging 算法 [Breiman, 2001b] 则是将 Bagging 作为 AdaBoost 的基学习器. 试比较二者的优缺点.

MultiBoosting 先降低模型的偏差再降低方差. 由于综合了 Bagging、AdaBoost，可以有效降低偏差和方差.
但训练成本和预测成本都会显著增加.

Iterative Bagging 先降低方差再降低偏差. 相比于 Bagging 偏差下降, 但方差上升.
训练成本和预测成本也会显著增加.

• 给定任意的两个相同长度向量$\boldsymbol{x}, \boldsymbol{y}$，其余弦距离为$1-\frac{\boldsymbol{x}^{\top}\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}$，证明余弦距离不满足传递性，而余弦夹角$\arccos\left(\frac{\boldsymbol{x}^{\top}\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}\right)$满足

(1) 取 $x=(1,0)^T$, $y=(1,1)^T$, $z=(0,1)^T$

使用余弦距离，有 $dist(x,z) = 1-\frac{x^Tz}{|x||z|} = 1$

$dist(x,y) = 1-\frac{x^Ty}{|x||y|} = 1-\frac{1}{\sqrt{2}}$

$dist(y,z) = 1-\frac{y^Tz}{|y||z|} = 1-\frac{1}{\sqrt{2}}$

$dist(x,y) + dist(y,z) = 2-\sqrt{2} < 1 = dist(x,z)$

∴ 余弦距离不满足传递性.

(2) 不妨设 $x$、$y$、$z$ 均为单位向量，即 $|x|=|y|=|z|=1$

要证 $\arccos x^Tz \le \arccos x^Ty + \arccos y^Tz$

① $\arccos x^Ty + \arccos y^Tz > \pi$ 则不等式显然成立

② $0 \le \arccos x^Ty + \arccos y^Tz \le \pi$

又 $\arccos x \in [0,\pi]$，$\forall x \in R$，$\cos x$ 在 $[0,\pi]$ 上单调递减

记 $x^Tz=a$，$x^Ty=b$，$y^Tz=c$

故即证 $\cos(\arccos a) \geqslant \cos(\arccos b + \arccos c)$

$a \geqslant \cos(\arccos b)\cos(\arccos c) - \sin(\arccos b)\sin(\arccos c)$

$\sin(\arccos b)\sin(\arccos c) \geqslant bc - a$

$(1-\cos^2(\arccos b))(1-\cos^2(\arccos c)) \geqslant (bc-a)^2$

$(1-b^2)(1-c^2) \geqslant b^2c^2 - 2abc + a^2$

$1-b^2-c^2+\cancel{b^2c^2} \geqslant \cancel{b^2c^2} - 2abc + a^2$

即证 $a^2+b^2+c^2-2abc-1 \le 0$

设 $m=(x,y,z)$

设 $A = m^Tm = \begin{pmatrix} 1 & x^Ty & x^Tz \\ y^Tx & 1 & y^Tz \\ z^Tx & z^Ty & 1 \end{pmatrix}$

则 $|A| = |m|^2 = 1+2abc-a^2-b^2-c^2 \geqslant 0$，证毕.

- 证明k-means算法的收敛性

对于每个数据点 $x_n$，引入一组对应的二值指示变量 $r_{nk} \in \{0,1\}$，
其中 $k=1,\cdots,K$，表示数据点 $x_n$ 属于 K 个聚类中心的哪一个.

定义目标函数 $J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n-\mu_k\|^2$，
表示每个数据点与它被分配到的向量 $\mu_k$ 之间的距离的平方和.

目标是找到 $\{r_{nk}\}$ 和 $\{\mu_k\}$ 的值，使得 J 达最小值. 可以使用一种迭代的方法.

1. 初始化 $\mu_k$.
2. 关于 $r_{nk}$ 最小化 J，保持 $\mu_k$ 固定.
3. 关于 $\mu_k$ 最小化 J，保持 $r_{nk}$ 固定.
4. 不断重复步骤2和3，直至收敛.

可见更新 $r_{nk}$ 和更新 $\mu_k$ 的两个阶段分别对应于 EM算法中的 E 和 M.
首先考虑 $r_{nk}$，J 是 $r_{nk}$ 的一个线性函数，因此最优化过程可以得到解析解.
将数据点的聚类设置为最近的聚类中心

$$r_{nk} = \begin{cases} 1, & \text{if } k = \arg\min_j \|x_n-\mu_j\|^2 \\ 0, & \text{else} \end{cases}$$

然后考虑 $r_{nk}$ 固定时，关于 $\mu_k$ 的优化. J是 $\mu_k$ 的一个二次函数，令它关于 $\mu_k$ 的导数为0，即可达到最小值，即

$$2\sum_{n=1}^{N} r_{nk}(x_n-\mu_k) = 0$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

，即令 $\mu_k$ 为类别 k 的所有数据点的均值.

重新为数据点分配聚类的步骤以及重新计算聚类均值的步骤重复进行，
直至聚类的分配不改变或迭代次数超过了某个最大值.

由于每个阶段都减小了目标函数 J 的值，因此算法的收敛性得到了保证.
但算法可能收敛到 J 的一个局部最小值而不是全局最小值.

- 在k-means算法中替换欧式距离为其他任意的度量，请问“聚类
  簇”中心如何计算？

$$J = \sum_{i=1}^{k}\sum_{x\in C_i} dist(x,\mu_i)$$

，若 J 可导，令 $\frac{\partial J}{\partial \mu_i} = 0$ $(i=1,\cdots,k)$

1. 随机选取 k 个样本作为初始均值向量 $\{\mu_1, \mu_2, \cdots, \mu_k\}$
2. 求 $\frac{\partial J}{\partial \mu_i} = 0$，更新 $\mu_i$ $(i=1,\cdots,k)$，若能求出解析解，则使用解析解作为更新值，否则使用梯度下降法.