HW7

8.2 $H(x)=\sum_{t=1}^{T}\alpha_t h_t(x)$ $\alpha_t=\frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$

对任意损失函数 $l(-f(x)H(x))$，整体损失 Loss 为

$Loss=E_x(l(-f(x)H(x)))=l(-H(x))P(f(x)=1|x)+l(H(x))P(f(x)=-1|x)$

当 $P(f(x)=1|x)>P(f(x)=-1|x)$ 时，即 x 分类为1的概率更大时，

为了使得损失函数更小，希望有 $l(-H(x))<l(H(x))$

由于 $l(-f(x)H(x))$ 对 $H(x)$ 是单调递减的，只有当 $H(x)>-H(x)$ 时，

才有 $l(-H(x))<l(H(x))$

又∵ $H(x)$ 只能取1或 -1，所以只有 $H(x)=+1$ 满足

同理，当 x 分类为 (-1) 的概率更大时，$H(x)$ 取值为 -1

由此可知，在最小化损失函数的过程中，已经达到了贝叶斯最优错误率，

使预测的标签与实际标签尽可能一致。l 可以为 0/1 损失函数的一致替代函数

8.6. 朴素贝叶斯分类器是通过使所有训练样本的后验概率达到最大而进行的，其误差主要来自于偏差。而Bagging主要关注于降低方差，难以提升朴素贝叶斯分类器的性能。

8.8. MultiBoosting 由于集合了Bagging，Wagging，AdaBoost，可以有效地降低误差和方差，特别是误差。但是训练成本和预测成本都会显著增加；Iterative Bagging 相比Bagging 会降低误差，但是方差上升。由于Bagging 本身是一种降低方差的算法，所以Iterative Bagging 相当于Bagging 和单分类器的折中。

HW 8.

1. ① 举反例:

$A=(1,0)$ $B=(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$; $C=(0,1)$

$D_{A,B}=1-\frac{\sqrt{2}}{2}$ $D_{B,C}=1-\frac{\sqrt{2}}{2}$ $D_{A,C}=1$

$D_{A,B}+D_{B,C}<D_{A,C}$ 不具有传递性

② $\arccos\left(\frac{x^Ty}{|x||y|}\right)$

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

令 $|x|=|y|=|z|=1$ 则 $(x,y,z)^T(x,y,z)\geq 0$.

$$\left|(x,y,z)^T(x,y,z)\right| = \begin{bmatrix} 1 & x^Ty & x^Tz \\ y^Tx & 1 & y^Tz \\ z^Tx & z^Ty & 1 \end{bmatrix} = 1+2(x^Ty)(y^Tz)(z^Tx)-(x^Ty)^2-(x^Tz)^2-(y^Tz)^2\geq 0$$

$$1+2\cos\theta_1\cos\theta_2\cos\theta_3-\cos^2\theta_1-\cos^2\theta_3-\cos^2\theta_2\geq 0.$$

$$(\cos\theta_1\cos\theta_2-\cos\theta_3)^2\leq(1-\cos^2\theta_1)(1-\cos^2\theta_2)$$

$$(\cos\theta_1\cos\theta_2-\cos\theta_3)^2\leq\sin^2\theta_1\sin^2\theta_2$$

$$|\sin\theta_1\sin\theta_2|\geq|\cos\theta_1\cos\theta_2-\cos\theta_3|$$

若 $\theta_1+\theta_2<\pi$. $\cos\theta_3\geq\sin\theta_1\sin\theta_2-\cos\theta_1\cos\theta_2$

$\cos\theta_3\geq\cos(\theta_1+\theta_2)$

又∵ $\cos\theta$ 在 $[0,\pi]$ 上单调递减, ∴ $\theta_1+\theta_2>\theta_3$. (看出传递性)

若 $\theta_1+\theta_2>\pi$, 则由于 $\theta_1+\theta_2+\theta_3<2\pi$

∴ $\theta_1+\theta_2>\theta_3$

三、余弦夹角满足 ~~传递性~~

2. 证明 k-means 算法的收敛性

考虑样本集 $D=\{x_1, x_2, \cdots, x_m\}$. 其中 x_t 为 n 维向量: $x_t=(x_1(t), \cdots, x_n(t))$ $1 \le t \le m$

对于 D 的一个非空子集 C_i，定义函数: $P(C_i)=\sum_{x_t \in C_i}(x_t-e_{C_i})'(x_t-e_{C_i})$

其中 $e_{C_i}=\frac{1}{|C_i|}\sum_{x_t \in C_i} x_t$. 对于 D 的一个任意划分, 目标函数可理解为

$C_1 \sim C_K$ 的目标函数求和, 即: $\sum_{i=1}^{k} P(C_i)$

$\sum_{i=1}^{k} P(C_i)=\sum_{i=1}^{m}(x_i-x^*)'(x_i-x^*)$ x^* 表示 x_i 所在类的中心

若 $\sum_{i=1}^{k} P(C_i)$ 在 k-means 算法迭代更新均值向量过程中不断减小，单调递减且有下界 0. 则说明此算法收敛。

① 选定 K 个数据向量 $\mu_1 \cdots \mu_K$.

计算每个数据到 K 个中心的距离, 并将数据划分到最近的类 C_j

计算新均值向量: $\mu_i'=\frac{1}{|C_i|}\sum_{x \in C_i} x$

$\sum_{i=1}^{k} P(C_i)=\sum_{i=1}^{n}(x_i-x^*)'(x_i-x^*)$

② 计算每个数据到 $C_1', C_2', \cdots C_K'$ K 个中心的距离，再将数据划入最近的类

$\sum_{i=1}^{k} P(C_i')=\sum_{i=1}^{n}(x_i-x')'(x_i-x')$

对于这其中任何一个 x_i 来说，两次划分时可选中心点位置一样，但第二次划分时都选择更近的中心点

所以每个 $(x_i-x')'(x_i-x')$ 必然不大于 $(x_i-x^*)'(x_i-x^*)$

$$\sum_{i=1}^{n}(x_i-x')'(x_i-x') \le \sum_{i=1}^{n}(x_i-x^*)'(x_i-x^*) = \sum_{i=1}^{k}\rho(C_i)$$

而新划分下新的中心位置 C_1^{**}、C_2^{**}、……C_k^{**} 是根据新的数据划分后计算的均值向量确定，下面证明：对于归属于同一类的数据 A_1、A_2、……A_j，要使 $\sum_{i=1}^{j}(A_i-x)'(A_i-x)$ 取到最小值当且仅当 $x=\frac{1}{j}\sum_{i=1}^{j}A_i$

$$\frac{\partial\sum_{i=1}^{j}(A_i-x)'(A_i-x)}{\partial x}=-2\sum_{i=1}^{j}(A_i-x)=0.$$

$x=\frac{1}{j}\sum_{i=1}^{n}A_i$ 是上式的一个驻点，又由于目标函数为严格凸函数，故有：$x=\frac{1}{j}\sum_{i=1}^{k}A_i$ 为目标函数唯一最小值点

因此有 $\sum_{i=1}^{k}\rho(C_i') \le \sum_{i=1}^{n}(x_i-x')'(x_i-x') \le \sum_{i=1}^{k}\rho(C_i)$

若均值向量有更新，则 $\sum_{i=1}^{k}\rho(C_i') \le \sum_{i=1}^{k}\rho(C_i)$

否则 $\sum_{i=1}^{k}\rho(C_i') = \sum_{i=1}^{k}\rho(C_i)$，但算法终止。

不妨设从第1次到最后一次划分，每次划分情况为 T_1、T_2、……

由于数据集 D 为有限数据集，所以对 D 的划分情况也是有限的，算法不可能总能找到与之前不同且更优的划分，算法最后一定会终止。且在算法执行过程中 $0<f(T_{p+1})<f(T_p)$　$f(T_p)$ 表示 T_p 对应的目标函数。　数列 $\{f(T_p)\}$ 单调递减且有下界。

由单调有界数列的收敛定理知，$\{f(T_p)\}$ 收敛.

即 $\lim_{p\to\infty}f(T_p)$ 存在，即有算法收敛

3. ① 曼哈顿距离：计算每个维度上所有数据点坐标的中位数。

② 余弦相似度：可以采用簇内所有数据点的平均向量