机器学习第六次作业

1. 分析Bagging通常为何难以提升朴素贝叶斯分类器的性能.

2. 分析随机森林为何比决策树Bagging集成的训练速度更快.

3. 假设抛硬币正面朝上的概率为p，反面朝上的概率为$1-p$. 令$H(n)$代表抛n次硬币所得的正面朝上的次数，则最多k次正面朝上的概率为

$$P(H(n) \leq k) = \sum_{i=1}^{k} \binom{n}{i} p^i (1-p)^{n-i} .$$

对$\delta > 0, k = (p-\delta)n$, 有Hoeffding不等式：

$$P(H(n) \leq (p-\delta)n) \leq e^{-2\delta^2 n} ,$$

请推导出教材P173页8.3式（集成的错误率）：

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

其中，f为真实函数，ϵ为基分类器（单次掷硬币判定）的错误率.

4. 编程实现Adaboost和Bagging，分别以决策树桩（单层决策树）、朴素贝叶斯为基分类器，在breast-cancer, glass数据集上做5次交叉验证，比较性能.