HOMEWORK 5

一、在 CART 剪枝过程中, 假设第 k 步, 对每个内部节点 t 计算 $C\left(T_t\right)$ 、$|T_t|$ 以及

$$g_k(t)=\frac{C(t)-C\left(T_t\right)}{|T_t|-1}$$

记第 k 步所有内部节点的集合为 $\mathcal{M}_k$, 记 $\alpha_k=g_k(a)=\min_{t\in\mathcal{M}_k} g_k(t)$, 即节点 a 是使函数 $g_k(t)$ 取值最小的内部节点（假设此内部节点唯一), 则将 a 剪枝。记剪枝后内部节点的集合是 $\mathcal{M}_{k+1}$, 定义 $\alpha_{k+1}=g_{k+1}(b)=\min_{t\in\mathcal{M}_{k+1}} g_{k+1}(t)$ 。请证明 $\alpha_{k+1}>\alpha_k$。

二、Hoeffding 不等式：设有独立的一系列随机变量 $X_1,\ldots,X_n$ ，且对于所有 $i=1,2,\ldots,n$，有 $P\left(X_i\in[a_i,b_i]\right)=1$，记均值为 $\bar{X}=n^{-1}\sum_{i=1}^n X_i$，则均值 $\bar{X}$ 满足以下不等式:

$$P(\bar{X}-E(\bar{X})\geq t)\leq\exp\left(-\frac{2t^2n^2}{\sum_{i=1}^n\left(b_i-a_i\right)^2}\right)。$$

考虑二分类问题 $y\in\{-1,1\}$ ，假设基分类器的错误率为 ϵ ，即 $P\left(h_i(x)\neq g(x)\right)=\epsilon$。若利用投票法集成 T 个基分类器：$H(x)=\operatorname{sign}\left(\sum_t h_i(x)\right)$，请利用 Hoeffding 不等式证明该分类器的错误概率小于 $\exp\left(-\frac{1}{2}T(1-2\epsilon)^2\right)$，即：

$$P(H(x)\neq g(x))\leq\exp\left(-\frac{1}{2}T(1-2\epsilon)^2\right)。$$

三、奥运视频热议度数据分析。

编程语言可以使用 R/python 等一切能满足题目要求的语言。具体任务见“奥运视频热议度”任务文档。

最后以 HTML/PDF 的形式提交报告。报告中需包括题目内容中涉及的代码和相关文字解释、结果分析。(提示：R 语言可用 Rmarkdown 输出分析报告；Python 可用 Jupyter 输出 HTML 报告；也自己进行格式调整后输出 PDF 格式的报告)

提交时间：5 月 1 日，晚 20:00 之前。请预留一定的时间，迟交作业扣 3 分，作业抄袭 0 分。