HOMEWORK 6

(1) 根据《统计学习方法》中表5.1所给的训练集数据，利用信息增益比算法（C4.5算法）生成决策树。

(2) 已知表 1所示的训练数据，试用平方损失准则生成一个二叉回归树。（提示：写出计算步骤）

x_i	1	2	3	4	5	6	7	8	9	10
y_i	4.50	4.75	4.91	5.34	5.80	7.05	7.90	8.23	8.70	9.00

表 1: 训练数据表

(3) 在CART剪枝过程中，假设第k步，对每个内部节点t计算$C(T_t)$、$|T_t|$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$\alpha_{k+1} > \alpha_k$.

以上证明题请以PDF格式提交。

(4) 数据分析及算法实现。

数据集介绍：请使用征信数据集完成实训题目，具体数据描述及题目要求见WORD文档。

注意：要求代码简洁、高效、可读性强；结果正确无误。提交HTML格式的代码文件。

提交时间：11月29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。