

HOMWORK 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 Table1 所示的训练数据，试用平方损失准则生成一个二叉回归树。

Table 1: 训练数据表

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

(3) 证明 CART 剪枝算法中求出的子树序列 $\{T_0, T_1, \dots, T_n\}$ 分别是区间 $\alpha \in [\alpha_i, \alpha_{i+1})$ 的最优子树 T_α , 这里 $i = 0, 1, \dots, n$, $0 = \alpha_0 < \alpha_1 < \dots < \alpha_n < +\infty$.

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 27 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。