统计学习：第五章

1 根据表 5.1 所给的训练数据集，利用信息增益比（C4.5 算法）生成决策树。
答：见代码

2 已知如表 5.2 所示的训练数据，试用平方误差损失准则生成一个二叉回归树。
答：见代码

3 证明 CART 剪枝算法中，当 α 确定时，存在唯一的最小子树方使损失函数 $C_\alpha(T)$ 最小。
答：若 T_1 和 T_2 都是满足条件的不同最小子树，故 $|T_1| = |T_2|$，则存在 $t \in T_1 \cap T_2$ 非 T_1 中的叶节点且为 T_2 中的叶节点，考察 $C_\alpha(t)$ 和 $C_\alpha(T_t)$ 必有 $C_\alpha(t) = C_\alpha(T_t)$，考虑 $C_\alpha(T_1) = C_\alpha(T_2)$。则考察 $C_\alpha(T_1 - t)$ 为更小的最优子树，矛盾，故存在唯一最小子树。

4 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，$i = 0, 1, \dots, n$, $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_n < +\infty$。
答：即证明，再在剪掉某最小 α 节点后, 所有其余节点 α 不会更小。首先看被剪枝节点的父节点 t，不妨设左子节点被剪枝，则原来有

$$C_\alpha(t) = \frac{C(t) - c(T_{t_L}) - C(T_{t_R})}{|T_L| + |T_R| - 1} \geq \frac{C(t_L) - C(T_{t_L}) + C(t_R) - C(T_{t_R})}{|T_L| + |T_R| - 1}, \tag{1}$$

现在为

$$C_\alpha(t)' = \frac{C(t) - c(t_L) - C(T_{t_R})}{1 + |T_R| - 1} \tag{2}$$

$$= \frac{C(t) - c(T_{t_L}) - C(T_{t_R} + C(T_{t_L}) - C(t_L)}{|T_L| + |T_R| - 1 - \ |T_L| - 1} \geq C_\alpha(t), \tag{3}$$

不等号由糖水不等式得到，故所有节点的 α 单调递增，因此 T_α 始终为最优子树。