**2019/04/06读书报告**

学号 71117228 姓名 李泓烨

读书进度：统计机器学习第5章

1. **读书报告内容**

**1. 自己提出问题的理解**

① 在5.4决策树的剪枝中，对于式子(5.14)而言，为什么较大的α促使选择较简单的模型，较小的α促使选择较复杂的模型？

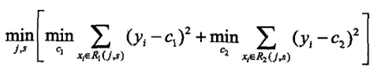
**讨论后的理解**：节点数|T|越大的话，也就是复杂度越高，所以如果增加α的值的话，就会让损失函数更大，更倾向于选择复杂度低的树

② 怎么理解(5.19)这个式子？

**讨论后的理解**：最优的c值就是均值，应该就是分成两个集合之后的平均值

**2. 别人提出问题的理解**

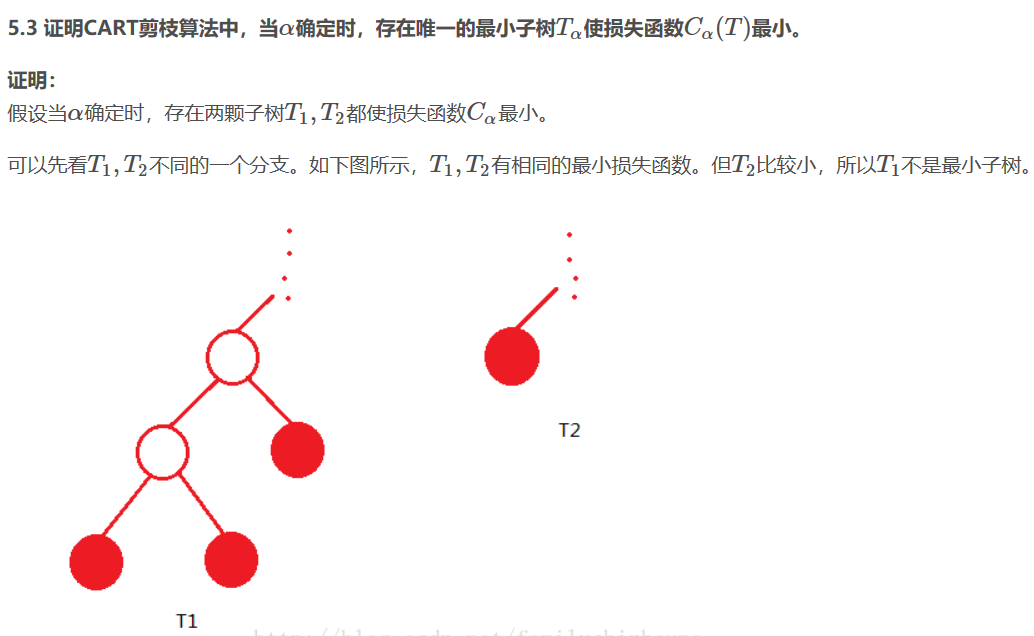
1. **提出的问题1**： CART的回归决策树生成算法中：这里的求最小值中的c1和c2也是平均值吗？



**讨论后的理解**：和我的第二个问题一样

2. **提出的问题2**：证明CART剪枝算法中，当α确定时，存在唯一的最小子树Tα使损失函数Cα(T)最小？

**小组成员的解释**：



3. **提出的问题3**：ID3算法和C4.5算法的区别，我认为信息熵增益最大不是应该就意味着相对信息熵增益最大吗，毕竟信息熵是相同的

**讨论后的理解**：我觉得之所以提出第二个算法的关键原因是这个

(1)信息增益的计算依赖于特征数目较多的特征，而属性取值最多的属性并不一定最优。

4. **提出的问题4**：5.5中，CART算法在决策树生成中，基于训练数据集生成决策树生成的决策树要尽量大是为什么？

**讨论后的理解**：考虑更多的情况吧。毕竟是二叉树，和多叉树的分类功能还是有区别的

5. **提出的问题5**：式子5.31表示了剪枝后整体损失函数减少的程度，但为什么说是在T0中剪去g(t)最小的T，不应该剪去损失函数减少的多的吗，怎么理解？

**讨论后的理解**：应该是对于不同的α值有不同的最优子树，但是这个α是不确定哪个最好，所以从最小的开始，一步步到最大，然后从这里面挑

**3. 读书计划**

① 本周完成的内容章节：看完了第5章

② 下周计划：看完第6章（视情况而定，暂时不知道第5章有多少）

**4. 读书摘要、理解&伪代码的具体实现**

1. 读书摘要

这是我的一些读书笔记，里面是对决策树的部分理解

