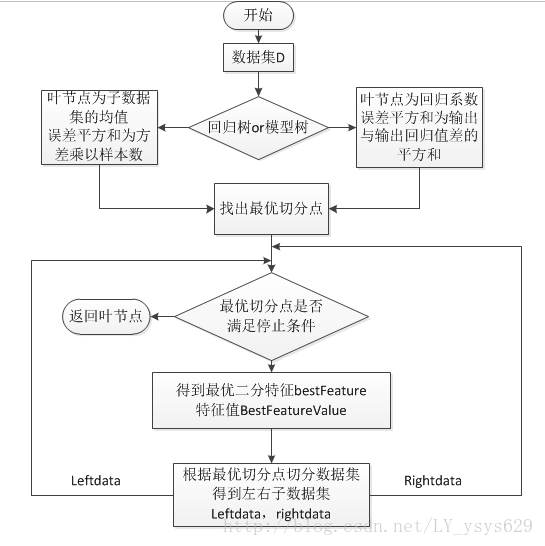
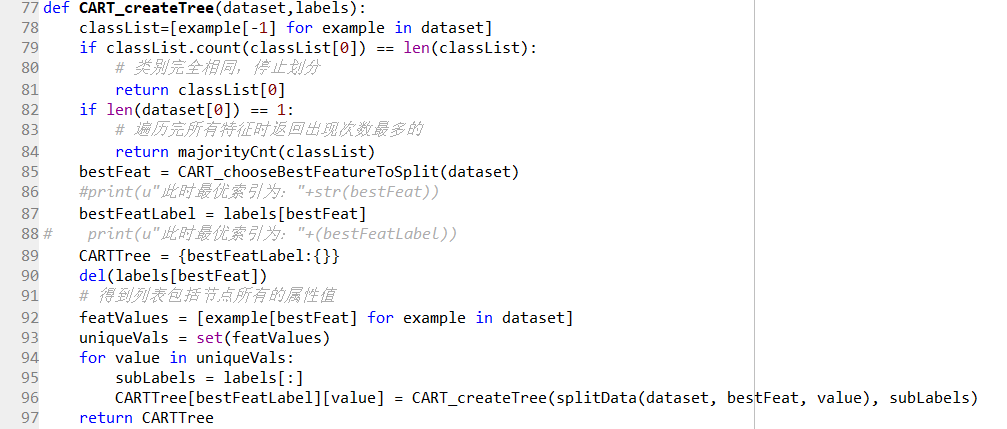
《机器学习》读书笔记

1. 算法概述（算法名称及原理）

CART：决策树的一种，CART既是分类树又是回归树。当CART是分类树的时候，采用GINI值作为分裂节点的依据，当CART作为回归树的时候，使用样本的最小方差作为分裂节点的依据

1. 算法设计（流程图及主要分段代码，附详细代码注释）







1. 选用数据（数据集描述，包括来源，行数，列数，格式等）

数据集是sklearn.datasets库中的鸢尾花数据

描述信息如下：

* 共有150行5列，
* 其中属性信息如下：

1. 萼片长度cm，
2. 萼片宽度cm，
3. 花瓣长度cm，
4. 花瓣宽度cm
5. 类：   
   - Iris Setosa   
   - Iris Versicolour   
   - Iris Virginica
6. 评价方法（说明训练集和测试集分配方法及评价指标）

十折[交叉验证](https://baike.baidu.com/item/%E4%BA%A4%E5%8F%89%E9%AA%8C%E8%AF%81/8543100" \t "https://baike.baidu.com/item/%E5%8D%81%E6%8A%98%E4%BA%A4%E5%8F%89%E9%AA%8C%E8%AF%81/_blank)。将数据集分成十份，轮流将其中9份作为训练数据，1份作为测试数据，进行试验。每次试验都会得出相应的正确率（或差错率）。10次的结果的正确率（或差错率）的平均值作为对算法精度的估计，一般还需要进行多次10折交叉验证（例如10次10折交叉验证），再求其均值，作为对算法准确性的估计。

1. 实验结果截图



1. 实验结果分析及比较

对于鸢尾花数据集，与c4.5相比cart算法的验证结果要更优



1. 遇到的问题及解决方法，实践心得

鸢尾花有三类结果，而cart是二叉树，需要将目标变量合并为超类，经过实践发现第一类和第三类合并效果最佳，另外关于数值型变量的离散化需要使用选择差异损失值最大的，目前感觉还有点不大会，就直接使用了等宽法。但是这样可能违背了cart算法，最近在忙着准备毕业答辩，时间上有点不够，我想在继续想想，之后一定弥补这个问题。

