## 2020.4.5 第十次读书报告

09018330 孙毅远

### 一、自己提出的问题的理解

#### 1.决策树剪枝时的值如何确定

理解：书上只给出大概定义，用于限制模型复杂度，根据具体要求具体确定

### 二、别人提出的问题的理解

#### 1. 如何证明CART剪枝算法中当确定时，存在唯一子树使得最小

理解：

#### 2.为什么较大的促使选择较简单的模型，较小的促使选择较复杂的模型？

理解：因为要使得损失函数尽量小，与结点数|T|相对应，是防止过拟合的一个罚项

#### 3. ID3算法和C4.5算法的区别

理解：C4.5算法是为了解决ID3决策树的一个缺点，当一个属性的可取值数目较多的时候，可能在这个属性对应的可取值下的样本只有一个或很少，那么这时候信息增益非常高，ID3算法会认为这个属性适合划分，但会带来泛化能力较弱的问题，不能对新样本进行有效检测

### 三、读书计划

#### 本周 学完第五章

#### 下周 学完第六章

### 四、读书笔记

#### 5.1 决策树模型与学习

* 分类决策树模型是一种描述对实例进行分类的树形结构，可以认为是定义在特征空间与类空间上的条件概率分布
* 也可看成一个if-then规则的集合，主要优点：模型具有可读性，分类速度快
* 决策树与条件概率分布
* 条件概率分布定义在特征空间的一个划分上，决策树的一条路径对应于划分中的一个单元，决策树所表示的条件概率分布由各个单元给定条件下类的条件概率分布组成
* 决策树学习
  + 决策树学习本质上是从训练数据集中归纳出一组分类规则
  + 决策树学习的损失函数通常为正则化的极大似然函数，学习的策略是以损失函数为目标函数的最小化
  + 当损失函数确定以后，学习问题就变为在损失函数意义下选择最优决策树的问题，通常采用启发式方法，得到的是次最优的（）
  + 算法：递归地选择最优特征并根据该特征对训练数据进行分割，使得对各个子数据集有一个最好的分类的过程
  + 这样产生的决策树可能产生过拟合，使用需要进行剪枝使得具有更好的泛化能力，或者对特征进行选择减少过于细分的过程
* 特征选择（全局选择）、决策树的生成（局部最优）、决策树的剪枝（全局最优）

#### 5.2 特征选择

* 特征选择问题
  + 准则通常是信息增益或信息增益率
* 信息增益
  + 设X是取有限个值的离散随机变量，概率分布为 则X的熵定义为
  + 熵越大，随机变量的不确定性越大
  + 联合分布 X给定条件下Y的条件概率分布的熵对X的数学期望
  + 信息增益
  + 特征A对训练数据集D的信息增益,定义为 也称为互信息（）
  + 特征选择方法：计算每个特征的信息增益，选择信息增益最大的特征
  + 算法：
* 信息增益比

#### 5.3 决策树的生成

* ID3算法
  + 相当于用极大似然估计法进行概率模型的选择
  + 选择信息增益最大的作为根节点，建立子节点再对子结点递归调用
  + 只有树的生成，容易产生过拟合
* C4.5生成算法
  + 对ID3算法进行改进，在生成的过程中用信息增益比来选择特征

#### 5.4 决策树的剪枝

* 表示模型对训练数据的预测误差，较大的促使选择较简单的模型，较小的选择较复杂的
* 剪枝，是当确定时选择损失函数最小的子树

#### 5.5 CART算法

* CART时在给定输入随机变量X条件下输出随机变量Y的条件概率分布的方法，假设决策树是二叉树，内部结点特征的取值为是和否，等价于递归地二分每个特征，将特征空间划分为有限个单元，在这些单元熵确定预测的概率分布
* CART生成
  + 递归地构建二叉决策树的过程，对回归树用平方误差最小化准则，对分类树用基尼指数最小化准则进行特征选择
  + 基尼指数
  + CART生成算法
  + 对结点计算现有特征对该数据集的基尼指数，对每个特征A对其可能取的每个值a，将D分割为两部分，选择基尼指数最小的特征及其切分点作为最优特征和最优切分点，再分配到子结点上，递归调用
* CART剪枝
  + 从底端剪去一些子树使得决策树变简单
  + 利用剪枝得到的子树序列通过交叉验证选取最优子树