# 决策树和集成方法

## 简述：

决策树大致会讲四种：ID3、C4.5、CART和C5.0。每一种按照需求不同侧重点不一样

决策树：

(从一个例子入手，讲解基于标签不纯度的计算方法)

## ID3、C4.5：

+数学基础(熵、信息增益(绝对增益)、信息增益比(相对增益))

+利用小数据集计算熵和增益(有**代码**)

+根据特征索引进行样本的划分

+递归建树(重难点：如何**控制结束条件**、设计**树的结构**

(每个节点处需要保存的内容(特征索引，分好类的数据，各个子树)))

+？？

## CART(分类回归树)：

即可以处理分类形式的特征空间，可以处理连续的特征空间

+CART的改进介绍(**大致介绍**)与伪代码,**解释分类回归的含义**

+数学基础(基尼不纯度、最小平方误差)

+基尼系数的计算

+选特征、选取值的过程(**代码**)

+讲树节点结构1 (分叉处的需要保存的内容(index, value, groups, left, right))

+讲树节点结构2 (为什么带着groups，方便后剪枝，当使用随机森林，

boosting等方法时，不需要后剪枝，可以从结构中去掉)

+递归画树(**停止条件**)、**先剪枝**的使用(设置树深度，设置树枝最小样本数量)

+树使用(包括两个函数：**对一个样本验证分类**；**对验证样本集给出正确率**)

+ CART的使用与**C5.0的引入**(有**实例**——岩石水雷数据，还会加一些**数据预处理**，**数据可视化**的代码)(把基尼系数改成**信息增益比**)(非重点，用来练习CART的使用)

+CART的改进(面向过程的编程——面向对象的编程)分两个主要步骤

+对象1：将节点从字典形式变成节点类(使用过程中，**判断建树终止条件处要注意**)

+对象2：创建CART类：建树(训练)过程，使用树(验证)过程

(类中的**每个函数的作用**，哪些是**类外最需要调用的**)

+使用创建的对象进行数据分类(岩石水雷数据集，鸢(yuan)尾花数据集，

适当介绍sklearn中的数据集和sklearn中 CART方法的使用)

+

集成方法：

(集成方法的定义、使用方法简述)

+主要分为三类bagging、RandomForest、boosting(AdaBoosting、GBDT、XGBoost)

+bagging，原理(多次有放回的取样， bagging大小和数目的确定)，

代码(改进CART代码)(可以加使用，可以不加，讲完RandomForest之后再使用)

+bagging进阶，怎么确定最终分类

(设计一个**函数记录最好的树**，**面向过程**，**面向对象**编程，哪个更好)

+RF，bagging再加上——随机提取d个特征(**特征数目的选取**，伪代码和**代码**)

+RF进阶，RF使用时一般外层还会加上n-折交叉验证(**n-折交叉验证**的使用)，

RF的使用(利用数据集进行训练，使用)

## Boosting

弱分类器和强分类器

+AdaBoosting-自适应提升:基于树桩的AdaBoosting(原理、公式、伪代码、sklearn使用)

+GBDT和XGBoost(只讲原理、使用，不对过程进行编程)

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 教学内容 | 学时 | 备注 |
| 1 | 课程简介+决策树简介 | 1 |  |
| 2 | ID3信息熵的计算(熵+条件熵+信息增益+特征熵)  编程实现(熵计算，分割，最优特征选取三兄弟) | 3 | 好了 |
| 3 | ID3建树+树结构分析+树的可视化+ID3树分类预测 | 2 | 好了 |
| 4 | ID3算法的不足(两方面)，提出C4.5(特征熵、相对信息增益)与离散化连续型特征 | 2 |  |
| 5 | C4.5建树与离散化方法代码 | 4 |  |
| 6 | 二分类与决策树 | 好了 |
| 7 | 基尼系数(编程)与最小均方误差 | 2 | 好了 |
| 8- | 重新规划(增益计算，分割，最优特征选取三兄弟) | 好了 |
| 8 | CART建树+两种递归的条件+先剪枝 | 2 | 好了 |
| 9 | CART分类树预测 | 2 | 好了后剪枝 |
| 9+ | CART回归建树与回归预测 | 2 | 明天逐步 |
| 10 | 数据可视化(箱线图|相关性|折线图|散点图|柱状图) | 2 | 明天准备 |
| 11 | CART 使用实例(使用CSV导入的数据集) |  |
| 12 | 决策树剪枝(后剪枝)+验证剪枝的方法 | 2 | 7个半天 |
| 13 | CART算法封装+使用 | 4 |
| 14 | 集成算法简介与Bagging和随机森林基础 | 2 | 理论基础又来 |
| 15 | Bagging和随机森林程序设计 | 2 | 理论有了 |
| 16 | N折交叉验证与随机森林 | 2 | 有代码 |
| 17 | 分类随机森林算法的使用（+sklearn） | 2 |  |
| 18 | 回归随机森林算法的使用(+sklearn) | 2 |  |
| 19 | Boosting与Adaboost | 2 |  |
| 20 | Adaboost程序设计 | 2 |  |
| 21 | GBDT和XGBoost理论基础 | 2 |  |
| 22 | GBDT使用 | 2 | 5个半天 |
| 23 | XGBoost使用 | 2 |
| 24 |  |  |  |
| 合计 | | 40 |  |
| 序号 | 教学内容 | 学时 | 备注 |
| 1 | 数据可视化 | 2 |  |
| 2 | CART算法应用 | 1 |  |
| 3 | Bagging算法的使用 | 1 |  |
| 4 | 随机森林算法的使用 | 1 |  |
| 5 | GBDT算法使用 | 2 |  |
| 6 | XGboost算法使用 | 1 |  |
| 7 |  |  |  |
| 8 |  |  |  |
| 9 |  |  |  |
| 10 |  |  |  |
| 11 |  |  |  |
| 合计 | | 8 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 教学内容 | 学时 | 备注 |
| 1 | 决策树简介 | 2 |  |
| 2 | 信息熵与ID3 | 4 |  |
| 3 | 离散化连续型特征 | 2 |  |
| 4 | 相对信息增益与C4.5 | 2 |  |
| 5 | 二分类与决策树 | 2 |  |
| 6 | 基尼系数与最小均方误差 | 1 |  |
| 7 | CART建树与预测 | 4 |  |
| 8 | CART算法封装 | 2 |  |
| 9 | 决策树剪枝 | 2 |  |
| 10 | CART做回归 | 2 |  |
| 11 | 集成算法简介 | 1 |  |
| 12 | Bagging和随机森林 | 2 |  |
| 13 | Bagging程序设计 | 2 |  |
| 14 | 随机森林程序设计 | 2 |  |
| 15 | N折交叉验证与随机森林 | 2 |  |
| 16 | Boosting与Adaboost | 2 |  |
| 17 | GBDT和XGBoost理论基础 | 2 |  |
| 合计 | | 36 |  |