决策树

决策树（decision tree）是一个树结构（可以是二叉树或非二叉树）。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果。

递归结束条件

1. 当前节点包含的样本全属于同一类别，将node标记为该类叶节点
2. 当前属性集为空，或是所以样本在所有属性上取值相同，将node标记为该节点包含最多的类别叶节点
3. 当前节点包含的样本集合为空，将node标记为其父节点包含最多的类别叶节点

划分属性

构造决策树的关键步骤是分裂属性。所谓分裂属性就是在某个节点处按照某一特征属性的不同划分构造不同的分支，其目标是让各个分裂子集尽可能地“纯”。尽可能“纯”就是尽量让一个分裂子集中待分类项属于同一类别。分裂属性分为三种不同的情况：

1、属性是离散值且不要求生成二叉决策树。此时用属性的每一个划分作为一个分支。

2、属性是离散值且要求生成二叉决策树。此时使用属性划分的一个子集进行测试，按照“属于此子集”和“不属于此子集”分成两个分支。

3、属性是连续值。此时确定一个值作为分裂点split_point，按照>split_point和<=split_point生成两个分支。

构造决策树的关键性内容是进行属性选择度量，属性选择度量是一种选择分裂准则，是将给定的类标记的训练集合的数据划分D“最好”地分成个体类的启发式方法，它决定了拓扑结构及分裂点split_point的选择。

根据不同的目标函数，决策树算法有不同的版本

ID3算法

ID3算法使用信息增益划分属性。“信息熵”是度量样本集合纯度最常用的一种指标，假设

$$Ent(D) = -\Sigma_{k=1}^{|Y|} p_k \log_2 p_k$$

其中pk代表了第k类样本在D中占有的比例。

假设离散属性a有V个可能的取值$\{a_1, a_2, \ldots, a_V\}$，若使用$a$对数据集$D$进行划分，则产生$D$个分支节点，记为$D_v$。则使用$a$对数据集进行划分所带来的信息增益被定义为：

$$Gain(D, a) = Ent(D) - \Sigma_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D_v)$$

一般的信息增益越大，则意味着使用特征a来进行划分的效果越好。

缺点：对取值数目较多的属性有所偏好

C4.5算法

信息增益率

$$Gain\_ratio(D,a)=\frac{Gain(D,a)}{IV(a)}$$

其中

$$IV(a)=-\sum_{\nu=1}^{V}\frac{|D^{\nu}|}{|D|}\log_2\frac{|D^{\nu}|}{|D|}$$

划分属性：增益率对取值数目较少的属性有所偏好，使用启发式方法，先从候选划分属性中找到信息增益率高于平均水平的属性，再从中选择增益率最高的。

CART决策树

Classification And Regression Tree，即分类回归树算法，简称**CART算法**。CART算法是一种二分递归分割技术，把当前样本划分为两个子样本，使得生成的每个非叶子结点都有两个分支，因此CART算法生成的决策树是结构简洁的二叉树。由于CART算法构成的是一个二叉树，它在每一步的决策时只能是"是"或者"否"，即使一个feature有多个取值，也是把数据分为两部分。在CART算法中主要分为两个步骤

- 将样本递归划分进行建树过程
- 用验证数据进行剪枝

递归建立二叉树

设$x_1,\ldots,x_n$代表单个样本的n个属性，y表示所属类别。CART算法通过递归的方式将n维的空间划分为不重叠的矩形。划分步骤大致如下

1. 选一个自变量x_i，再选取x_i的一个值v_i，v_i把n维空间划分为两部分，一部分的所有点都满足$x_i\le v_i$，另一部分的所有点都满足$x_i>v_i$，对非连续变量来说属性值的取值只有两个，即等于该值或不等于该值。
2. 递归处理，将上面得到的两部分按步骤（1）重新选取一个属性继续划分，直到把整个n维空间都划分完。

划分标准

对于一个变量属性来说，它的划分点是一对连续变量属性值的中点。假设m个样本的集合一个属性有m个取值，那么则有$m-1$个分裂点，每个分裂点为相邻两个连续值的均值。每个属性的划分按照能减少的杂质的量来进行排序，而杂质的减少量定义为划分前的杂质减去划分后的每个节点的杂质量划分所占比率之和。而杂质度量方法常用**Gini指标**，假设一个样本共有C类，那么一个节点A的Gini不纯度可定义为

$$Gini(A)=1-\sum_{i=1}^{C}p_i^2$$

如果当前节点的所有样本都不属于同一类或者只剩下一个样本，那么此节点为非叶子节点，所以会尝试样本的每个属性以及每个属性对应的分裂点，尝试找到Gini不纯度最小的一个划分，该属性划分的子树即为最优分支。

缺失值处理

在样本属性值缺失的情况下，我们主要解决俩个问题

- 如何划分属性选择
- 给定划分属性，如何对样本进行划分

对于问题1，我们根据没有缺失值的样本集进行划分属性选择。

对于问题2，如果样本x在属性a上取值未知，则将x同时划入所有子节点，根据各个子节点的样本数调整权重。