决策树

决策树是一个树形结构（可以是二叉树或非二叉树）

内部节点表示一个属性上的判断，每个分支表示一个判断结果的输出，叶子节点代表分类结果

有监督的多分类模型

决策树的生成步骤：

1）节点的分裂

2）阈值的确定

目的是寻找一个特征，用该特征划分后，数据集的熵最小，依次选择特征，直至完全分类。

☆ ID3.5

entropy

信息论中有熵的概念，熵越大越混乱

熵的变化可看做是信息增益

ID3.5的核心思想是以信息增益做为属性选择的标准，选择分裂后信息增益最大的属性分裂

设数据集为D，其熵为

$$Entropy(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

其中p_i表示第i个类别在D中出现的概率.

将训练数据集D按照A属性进行划分

则A对D划分的条件熵为

$$Entropy(D|A) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Entropy(D_j) \quad (2)$$

其中v表示属性A的特征值个数，D_j表示按照第j个特征值划分的数据集

因此属性A的信息增益为

$$gain(A) = Entropy(D) - Entropy(D|A) \quad (3)$$

☆ C4.5

C4.5算法实现了分裂信息和信息增益率

属性A对应的分裂信息为

$$Split(A) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2\left(\frac{|D_j|}{|D|}\right) \quad (4)$$

信息增益率为 $gain\_rate = \frac{gain(A)}{Split(A)}$ (5)

C4.5的核心思想是选择信息增益率最大的属性进行分裂

☆实现时：某个属性的特征值对应的记录数

决策树的构造是递归的，所以需要停止条件

一种可行的办法是当前节点中的记录数低于固定阈值时停止，选择概率最大的类别作为当前叶节点的分类。