决策树

决策树模型

由结点(node)和有向边(directed edge)组成，其中内部结点表示一个特征或属性，叶结点代表一个类

if-then规则

每一个实例都被一条路径或一个规则覆盖，而且只被一条路径或一个规则所覆盖，决策树所构成的路径集合是互斥且完备的

条件概率分布

决策树还表示给定特征条件下类的条件概率分布。这一条件概率分布定义在特征空间的一个划分(partition)上。将特征空间划分为互不相交的单元(cell)或区域(region)，并在每个单元定义一个类的概率分布就构成了一个条件概率分布。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

- 本质：从训练数据集中归纳出一组分类规则，与训练数据集不相矛盾
- 假设空间：由无穷多个条件概率模型组成
- 策略：最小化损失函数
- 特征选择：递归选择最优特征
- 生成：对应特征空间的划分，直到所有训练子集被正确分类
- 剪枝：避免过拟合，具有更好的泛化能力
- 目标：与训练数据矛盾较小的同时具有很好的泛化能力

信息增益

设X是一个取有限个值的离散随机变量，其概率分布为

$$P(X=x_i)=p_i \quad ,i=1,2,...,n$$

则随机变量X的熵定义为

$H(X) = -\sum_{i=1}^{n} p_i \log p_i$,也可记作$H(p)$

熵越大，随机变量的不确定性就越大

$0 \leq H(p) \leq \log n$

条件熵H(YIX)表示在已知随机变量X的条件下随机变量Y的不确定性。随机变量X给定的条件下随机变量Y的条件熵(conditional entropy)H(YIX)，定义为X给定条件下Y的条件概率分布的熵对X的数学期望。

$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i)$

当熵与条件熵中的概率是通过极大似然法估计得到的，则其为经验熵和经验条件熵（$p = \frac{C_K}{D}$ ）

信息增益：得知特征X而使类Y的信息不确定性减少的程度

$g(D, A) = H(D) - H(D|A)$

信息增益比

信息增益值的大小是相对于训练数据集而言的，并没有绝对意义。在分类问题困难时，也就是说在训练数据集的经验熵大的时候，信息增益值会偏大。反之，信息增益值会偏小。使用信息增益比(information gain ratio)可以对这一问题进行校正。这是特征选择的另一准则。

特征A对训练数据集D的信息增益比gR(D,A)定义为其信息增益g(D,A)与训练数据集D的经验熵H(D)之比:

$g_R(D, A) = \frac{g(D,A)}{H(D)}$

决策树生成

ID3算法

输入：训练数据集D，特征集A，阈值ϵ

输出：决策树T

- 判断T是否需要选择特征生成决策树

    - 若D中所有实例属于同一类，则T为单结点树，记录实例C_k 以此为该结点的类标记，并返回T
    - 若D中所有实例无任何特征$(A=\emptyset)$ ，则T为单结点树，记录D中实例个数最多类别C_k ，以此作为该结点的类标记，并返回T
- 否则，计算A中各特征的信息增益，并选择信息增益最大的特征A_g :
    - 若A_g 的信息增益小于ϵ 则T为单结点树，记录D中实例个数最多类别C_k ,以此作为该结点的类标记，返回T
    - 否则，按照A_g 的每个可能取值a_i ,将D分为若干非空子集D_i ，将D_i 中实例个数最多的类别作为标记，构建子结点,以结点和其子结点构成T，并返回T;
- 第i个子结点，以D_i 为训练集，$A-A_g$ 为特征集，递归调用上述步骤得到子树T_i 并返回。

C4.5算法

替换ID3中第二步信息增益为信息增益比

剪枝

一颗优秀的决策树在具有好的拟合与泛化能力同时：

- 深度小
- 叶结点少
- 深度小且叶结点小

～用来解决过拟合问题。

预剪枝

生成过程中，对每个结点划分前进行估计，若当前结点的划分不能提升泛化能力，则停止划分，记当前结点为叶结点。

方法：

- 限定决策树深度
- 设定一个阈值
- 设置某个指标，比较结点划分前后的泛化能力

CART算法

CART 树为二叉树

基尼指数

假设现在由K个类，样本点属于第k个类的概率为p_k 则概率分布的基尼指数为

$Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$

二分类

$Gini(p) = 2p(1-p)$

样本集D的基尼指数

$Gini(D) = 1 - \sum_{k=1}^{K}(\frac{|C_k|}{|D|})^2$

特征条件A下，样本集D的基尼系数为

$Gini(D,A) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2)$

CART分类树算法

输入：训练数据集D，特征集A，阈值ϵ

输出：CART决策树T

- 从根结点出发，进行操作，构建二叉树
- 结点处的训练数据集为D，计算现有特征对该数据集的基尼指数，并选取最优特征
    - 在特征A_g 下对其可能取 的每个值a_g ,根据样本点对$A_g = a_g$ 的测试为 “是“或者”否“，将D分割成D1和D2两部分计算$A_g = a_g$ 时的基尼指数
    - 选择基尼指数最小的那个值作为该特征下的最优切分点
    - 计算每个特征下的最优切分点，并比较在最优切分点下的每个特征的基尼指数，选择基尼指数最小的那个特征，即为最优特征
- 根据最优特征和最优切分点，从现结点生成两个子结点，将训练数据集根据特征分配到两个子结点中去
- 分别对两个子结点递归调用上述步骤，直至满足停止条件