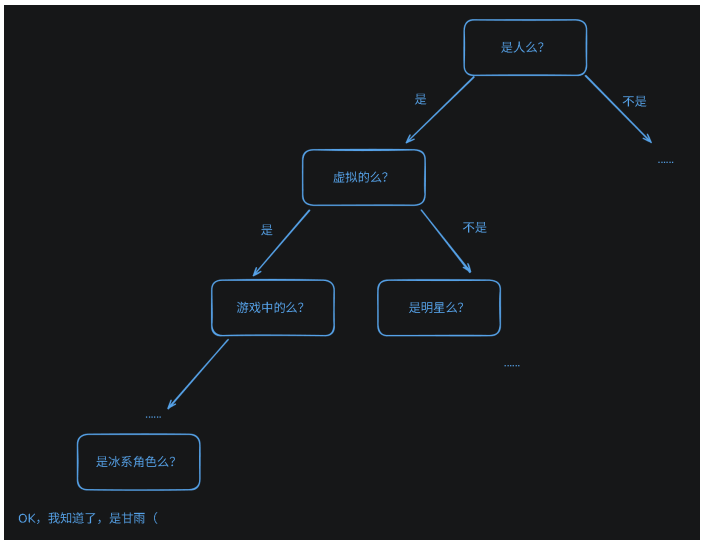
决策树基础

前言

决策树的基本概念其实很简单。想象你在玩20问：小爱同学用一连串“是/否”问题缩小猜测范围，最终锁定答案。决策树做的正是类似的事——把样本空间反复划分（分治策略），直到每个子空间“纯度”足够高。区别在于，树不再靠“随便问”，而是用信息增益（信息论中的信息熵）或基尼系数挑“最佳问题”（无用信息的数量），并且通过剪枝防止问得太多而记错噪声（去掉无用信息）。下面，我们把游戏翻译成数学语言，正式讲一讲这棵“决策树”。

基本流程

在最开始，我们不应该直接上手困难问题，而是从简单的二分类问题入手，也就是二叉决策树。这里以上面的20问为例，每次的答与问都将问题范围进行缩减，并且这个问题只存在“是或者不是”两个答案，最终这个决策的过程就成为了



显然，决策过程的最终结论对应了我们所希望的判定结果，例如“是”或 “不是”；决策过程中提出的每个判定问题都是对某个属性的“测试”，例如：“是人么？”、“虚拟的么？”；每个测试的结果或是导出最终结论，或是导出进一步的判定问题，其考虑范围是在上次决策结果的限定范围之内，例如在虚拟之后判断是否是游戏当中的。

一般的，一棵决策树包含一个根结点、若干个内部结点和若干个叶结点；叶结点对应于决策结果， 其他每个结点则对应于一个属性测试；每个结点包含 的样本集合根据属性测试的结果被划分到子结点中； 根结点包含样本全集。从根结点到每个叶结点的路径对应了一个判定测试序列。决策树学习的目的是为了产生一棵泛化能力强，即处理未见示例能力强的决策树，其基本流程遵循简单且直观的“分而治之” (divide-and-conquer)。如下图所示：

输入： 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
　　属性集 $A = \{a_1, a_2, \ldots, a_d\}$.

过程： 函数 TreeGenerate(D, A)

1: 生成结点 node;
2: **if** D 中样本全属于同一类别 C **then**
3:　将 node 标记为 C 类叶结点; **return**　　递归返回，情形(1).
4: **end if**
5: **if** $A = \varnothing$ **OR** D 中样本在 A 上取值相同 **then**
6:　将 node 标记为叶结点，其类别标记为 D 中样本数最多的类; **return**　　递归返回，情形(2).
7: **end if**
8: 从 A 中选择最优划分属性 a_*;　　我们将在下一节讨论如何获得最优划分属性.
9: **for** a_* 的每一个值 a_*^v **do**
10:　为 node 生成一个分支; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
11:　**if** D_v 为空 **then**
12:　　将分支结点标记为叶结点，其类别标记为 D 中样本最多的类; **return**　　递归返回，情形(3).
13:　**else**
14:　　以 TreeGenerate(D_v, $A \setminus \{a_*\}$)为分支结点　　从 A 中去掉 a_*.
15:　**end if**
16: **end for**

输出： 以 node 为根结点的一棵决策树

图 4.2　决策树学习基本算法

显然，决策树的生成是一个递归过程。在决策树基本算法中，有三种情形会导致递归返回：

1. 当前结点包含的样本全属于同一类别，无需划分
2. 当前属性集为空，或是所有样本在所有属性上取值相同，无法划分
3. 当前结点包含的样本集合为空，不能划分。

在第 (2) 种情形下，我们把当前结点标记为叶结点，并将其类别设定为该结点所含样本最多的类别；在第 (3) 种情形下，同样把当前结点标记为叶结点，但将其类别设定为其父结点所含样本最多的类别。注意这两种情形的处理实质不同：情形 (2) 是在利用当前结点的后验分布，而情形 (3) 则是把父结点的样本分布作为当前结点的先验分布。

举个例子：

假设父结点里有 10 个样本。其中 A 类：7 个，B 类：3 个。现在父结点要按“颜色”划分：

- **红色**：有 6 个样本（A=5，B=1）
- **蓝色**：有 4 个样本（A=2，B=2）
- **绿色**：训练集中根本没人是绿色 → 于是“绿色”分支是空结点（A=0，B=0）。
- 对“红色”子结点：如果还有可用特征，就继续划分；如果没有可用特征，就按 (2) → 标为 **A 类（多数类）**。
- 对“蓝色”子结点：同理，如果不能再划分 → 标为 **A 类（2 vs 2 → 随机或按规则打破平局）**。
- 对“绿色”子结点：没样本 → 按 (3)，回到父结点的多数类 → **A 类**。

所以我们才说：**(2)** 有样本但不能再分 → 叶子类别 = **当前结点多数类**（后验分布）。**(3)** 没样本 → 叶子类别 = **父结点多数类**（先验分布）。

注意：这里举例子的（A=5，B=1）、（A=2，B=2）和（A=0，B=0），这里的个数其实可以计算出信息出现的概率，我们就可以引出上一篇文章所说的——信息论。