# 8.3.1 决策树的理论

2023 年 9 月 18 日

## 决策树的理论

决策树是监督学习中最常用、最实用的方法之一。它既可用于解决回归任务，也可用于解决分类任务，而其分类任务被更多地用于实际应用，如广告推荐、商户分类等。

## 1. 一个例子

我们以一个生活中常见的相亲例子，来开启决策树的介绍和讨论：

一个眼光略高对自己婚姻伴侣略有要求的女青年**刘小姐**，熟练得在相亲网站上筛选着相亲对象。

这个相亲网站有超过 10 万个真实注册用户，她固然可以直接设置一定的筛选规则，例如：“年龄在 25～35 之间”，“收入在中档以上”，“相貌较好”等条件，让网站在数据库里检索。

但是你已经意识到了，这样的规则会过滤掉不符合其中一些条件，但是在其他条件上却很好的男性。有没有什么方法来解决这个问题呢？

同时，当有新的男士资料被添加到服务器，刘小姐也不是每时每刻都在线，错过了，如何是好？

我们希望用机器学习的方法，让机器习得刘小姐**对于另一半的偏好**。这样，机器就可以 24 小时全天候帮她把关，并及时通过短信和邮件通知刘小姐。

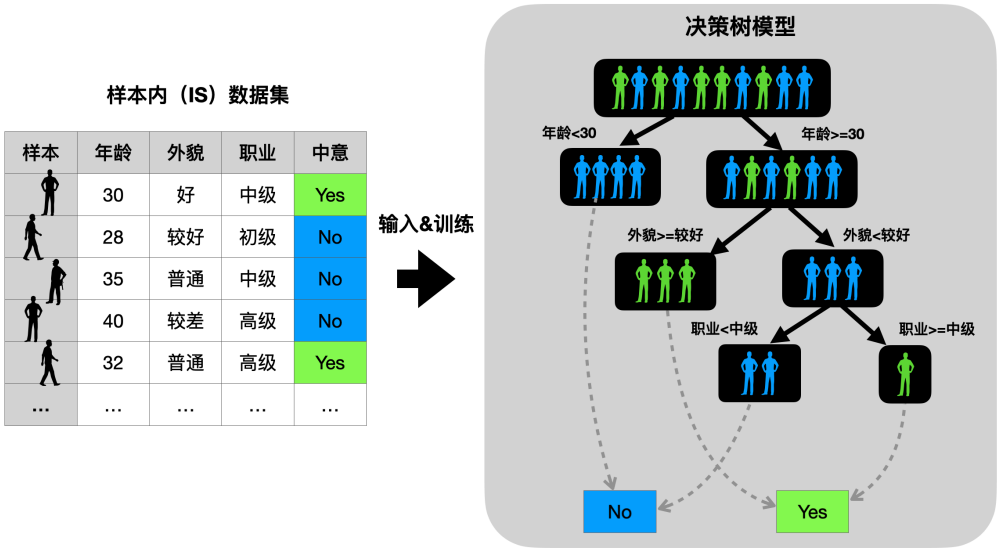
## 2. 模型的训练和预测过程

首先，准备数据样本：

1）根据过去的浏览记录，将刘小姐点“□”的男士样本打上标签，“Yes”，点“□”或者拉黑的男士样本打上“No”；
2）根据随机对照实验理论，将样本分为样本内（IS）和样本外（OOS），比例可设为 8:2，前者用来训练模型，后者用来验证模型的分类能力；

## 2.1 样本内（IS）训练过程

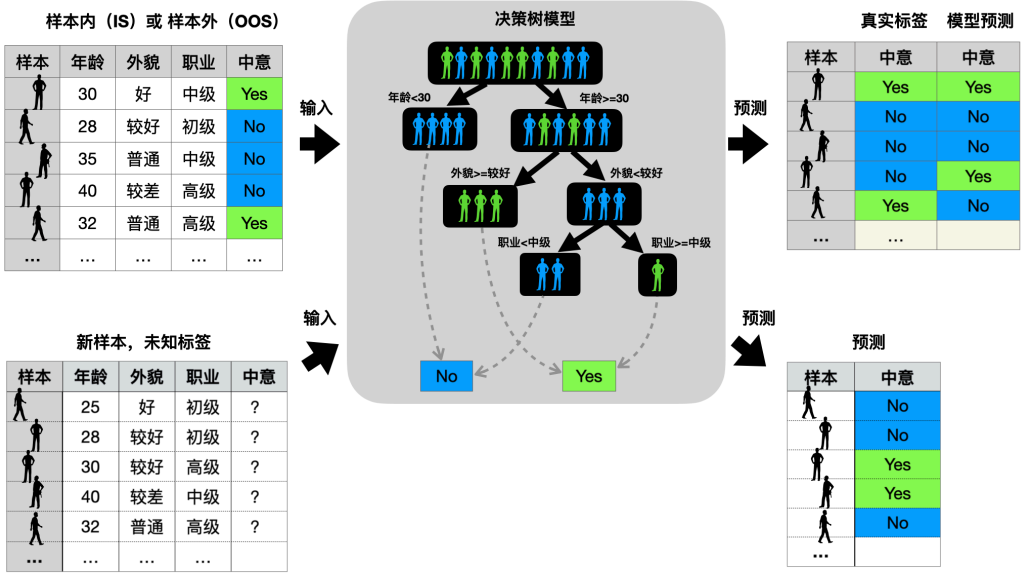
训练过程如下：1. 从上至下的方向，创建决策树的根节点，将样本内的数据集放入根节点中；2. 依次选择各个属性，对属性的数值进行划分（大于等于属性的某个值和小于这个属性值），并计算划分后的纯度；3. 选择划分后纯度最高或者标签一致的属性，作为当前分叉的标准，标签一致的为叶节点，另一个为中间节点；4. 回到步骤 3，并重复执行 1-3，直到所有中间节点都划分为叶节点。




模型训练完，也就是决策树生成完毕后，就可以将样本内的数据输入到模型中。然后，观察模型预测的结果和真实的标签，两者之间的差异，这代表了模型的学习能力。

## 2.2 样本外（OOS）的评估以及未知样本的预测

在模型学习到**刘小姐**的偏好后，就可以在**刘小姐**不在线的时候，也可以帮助其筛选样本了。当输入新的未知标签的数据到模型，模型会根据已经训练后的决策树的路径，很多就可以对样本进行分类，属于“Yes”还是“No”。




## 3. 划分的选择

关于决策树从最上面的**根节点**，到下面的**中间节点**以及末端的**叶节点**，如何划分，换句话说在每一次划分选择哪个属性进行划分，决定了决策树的形状。

如何选择最优划分属性。一般而言，随着划分过程不断进行，我们希望决策树的分支结点所包含的样本尽可能属于同一类别，即结点“纯度”（purity）越来越高。

节点的 Gini 属性用于测量它的纯度：如果一个节点包含的所有训练样例全都是同一类别的，我们就说这个节点是纯的（Gini=0）。Gini 计算如下：

$$G = \sum_{i=1}^{C}(p(i) * (1 - p(i)))$$

## 4. 模型的性能评估

### 4.1 精度与错误率

如何评估一个模型的好坏，一个自然而然的想法就是：模型给出的预测值与真实值进行对比。

错误率：分类错误的样本数占样本总数的比例
精度：分类正确样本数占样本总数的比例

精度计算如下：

$$accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i)$$

$$= 1 -$$

### 4.2 查全率与查准率（局部考量结果）

在二分类任务（类别为两类）中，假如我们定义'positive' 和 'negative' 为分类的预测结果，而 'true' and 'false' 指的是该预测是否符合，可以得到以下表格。上表为英文，下表为中文翻译。记住！多分类任务（类别为 3 类及以上）没有查全率和查准率。

| | | Actual class | |
|---|---|---|---|
| | | positive | negative |
| Predicted class | positive | tp (ture positive) | fp (false positive) |
| | negative | fn (false negative) | tn (true negative) |

| | | 真实类别 | |
|---|---|---|---|
| | | 正 | 负 |
| 预测类别 | 正 | 真正例 | 假正例 |
| | 负 | 假反例 | 真负例 |

查准率（准确率，precision）

$$precision = \frac{tp}{tp + fp}$$

查全率（也叫召回率，灵敏度，recall）

$$recall = \frac{tp}{tp + fn}$$

综合查准率与查全率

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

从上图可以知道：

**查准率**：是基于「预测数据」，考察「真正例」的占比。
**查全率**：是基于「真实数据」，考察「真正例」的占比。

来看看两个例子：

例子 1：如：在病情诊断时，我们希望查准率越高越好，减少病情误判。这样就需要约束条件比较严苛，落在约束条件下的样本数量较小，查全率自然就小了。

例子 2：如：在逃犯搜捕过程中，我们希望不放过一个漏网之鱼，所以就希望查全率越高越好。这样就需要约束条件比较宽松，落在约束条件下的样本数量较大，查准率自然就小了。