# Machine Learning---决策树

## 引言

决策树(Decision Trees)，一个非常经典的监督式学习案例。

## 决策树基本介绍

### 1.决策树

在建立决策树时候，计算机使用已分类的数据，建立一个具有预测功能的树形结构。

建立决策树有许多种算法,这里我只介绍ID3算法。

### 2.核心算法

在建立决策树的时候，首先要选出哪一个属性作为最重要的属性，成为决策树的根节点。这里将用到信息理论中的一个概念：熵（entropy）。具体计算使用下面的公式。

公式（1）

对于这个公式我稍作解释。其中p(I)便是在集合S中属于类型I样本的比例。

如果S中所有的样本都属于I的话，P(I)便是1，所以E(S) = -1 \* = 0，这也表示集合S是纯净的。

如果S中所有的个体都随机分布，这样的话

因为每一个个体都是随机分布，所以每一个个体属于类型I的概率都是一样的。这样。所以

公式(2)

公式2便是信息增量，用于决定那个属性作为决策树的根节点。其中Gain越大的属性便是根节点。

## 决策树例子

|  |  |  |  |
| --- | --- | --- | --- |
| **天气（weather）** | **身体状况（Health）** | **工作量（Wordload）** | **行为（Behavior）** |
| 晴天 | 良好 | 轻松 | 打球 |
| 晴天 | 糟糕 | 繁重 | 不打球 |
| 阴天 | 良好 | 轻松 | 打球 |
| 阴天 | 糟糕 | 轻松 | 不打球 |
| 雨天 | 良好 | 轻松 | 不打球 |
| 雨天 | 良好 | 繁重 | 不打球 |
| 雨天 | 糟糕 | 轻松 | 不打球 |
| 雨天 | 糟糕 | 繁重 | 不打球 |

表2.1

上表中包含八个样本，通过此表我们如何建立决策树？首先向计算出决策树的根节点，这样就要计算天气、身体状况和工作量的信息增益（Information Gain）。取其中最大信息增益的条件。

首先计算

其中晴天的熵计算，其中天气为晴天的有两个样本，在这两个样本中，有两个行为结果。

同样方式计算阴天和雨天的熵

所以得到

同样的方法可以计算出另外两个(身体状况和工作量)的信息增益

三个条件计算完毕，得出天气和身体两个条件的信息增益是一样的，都比工作量大，我们这里就取天气作为根节点（当然也可以取身体）。我们已经找到了决策树的根节点，但是任务还没有完成。我们要开始建立三棵子树。

雨

晴

阴

**？**

**？**

**？**

**天 气**

晴天条件下

|  |  |  |
| --- | --- | --- |
| **身体状况（Health）** | **工作量（Wordload）** | **行为（Behavior）** |
| 良好 | 轻松 | 打球 |
| 糟糕 | 繁重 | 不打球 |

表2.2

阴天条件

|  |  |  |
| --- | --- | --- |
| **身体状况（Health）** | **工作量（Wordload）** | **行为（Behavior）** |
| 良好 | 轻松 | 打球 |
| 糟糕 | 轻松 | 不打球 |

表2.3

雨天条件

|  |  |  |
| --- | --- | --- |
| **身体状况（Health）** | **工作量（Wordload）** | **行为（Behavior）** |
| 良好 | 轻松 | 不打球 |
| 良好 | 繁重 | 不打球 |
| 糟糕 | 轻松 | 不打球 |
| 糟糕 | 繁重 | 不打球 |

表2.4

按照一开始介绍的方法，分别为上面三张表找决策树的根节点（信息增益最大的条件）。就这样不断递归建立。

## 总结

建立决策树的方法有很多种，这里我使用的是ID3算法，另外还有C4.5和C5.0算法。这里就暂时不介绍了。

如果上面有计算错误之处，敬请指出。

如果有兴趣的可以留言，一起交流一下算法学习的心得。

声明：本文章是笔者整理资料所得原创文章，如转载需注明出处，谢谢。