机器人足球第四次作业



题目：学习一个吃饭的时候是否等待的决策树。

需求：按照ID3算法推出最终的决策树。

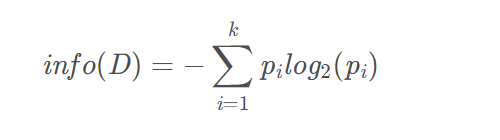
可程序实现；也可以手工推出（要有手动推出过程）。

**本次作业我采取手工推出的方式来完成此次作业。**

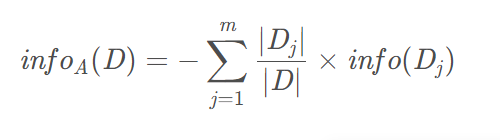
**思路描述：**

ID3算法使用的数据特征函数（标准）为信息增益

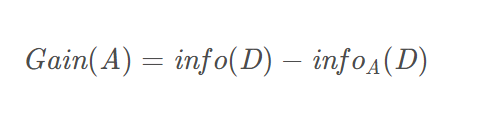
1. 假设在数据D中有k个类别，其中第i个类别数据在总数据中的占有率为p，则熵的计算公式为：



1. 当我们使用某一特征A对数据分类之后，其不确定度就会减小，即熵会减小，假设特征Ａ有ｍ个类别，其计算公式为：



1. 那么分类前后熵减小的差值就是信息增益。



1. 一一计算所有变量的信息增益，选择信息增益最大的那个变量作为此分类节点。产生第一次决策树之后，再对每个叶节点再次利用同样的过程，最终得出完整的决策树。

**实现过程：**



从上图可以看出，属于是的结果有6个，否的结果也有6个，可计算出训练集的熵为：

下面对各个属性计算对应的信息增益：

|  |  |  |
| --- | --- | --- |
| **是否有其他选择** | **是/熵** | **否／增益** |
| **是** | 2 | 4 |
| **否** | 4 | 2 |
| **熵** | 0.9182 | 0.0818 |

|  |  |  |
| --- | --- | --- |
| **饿否** | **是/熵** | **否／增益** |
| **是** | 5 | 2 |
| **否** | 1 | 4 |
| **熵** | 0.8042 | 0.1958 |

|  |  |  |
| --- | --- | --- |
| **价格** | **是/熵** | **否／增益** |
| **低** | 3 | 4 |
| **中** | 2 | 0 |
| **高** | 1 | 2 |
| **熵** | 0.8042 | 0.1958 |

|  |  |  |
| --- | --- | --- |
| **餐馆类型** | **是/熵** | **否／增益** |
| **中式** | 2 | 2 |
| **意大利式** | 1 | 1 |
| **法式** | 1 | 1 |
| **快餐** | 2 | 2 |
| **熵** | 1.0 | 0.0 |

|  |  |  |
| --- | --- | --- |
| **餐馆顾客人数** | **是/熵** | **否／增益** |
| **无人** | 0 | 2 |
| **有人** | 4 | 0 |
| **客满** | 2 | 4 |
| **熵** | 0.4591 | 0.5409 |

|  |  |  |
| --- | --- | --- |
| **等待时间** | **是/熵** | **否／增益** |
| **0-10** | 4 | 2 |
| **10-30** | 1 | 1 |
| **30-60** | 1 | 1 |
| **＞60** | 0 | 2 |
| **熵** | 0.7924 | 0.2076 |

从上面表格数据中可以得出餐馆**顾客人数**的信息增益最大，所以选择顾客人数作为根节点的测试属性。

餐馆顾客人数为无人时均为否。

餐馆顾客人数为有人时均为是。

以上二者的熵均为0，将其作为叶节点。

故现在只要考虑餐馆人数为客满的状况：

餐馆顾客人数为客满时的数据为：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **是否有其他选择** | **饿否** | **价格** | **餐馆类型** | **等待时间** | **是否等待** |
| **是** | 是 | 低 | 中餐 | 30-60 | 否 |
| **是** | 是 | 低 | 中餐 | 10-30 | 是 |
| **是** | 否 | 高 | 法式 | ＞６０ | 否 |
| **否** | 否 | 低 | 快餐 | ＞６０ | 否 |
| **是** | 是 | 高 | 意大利式 | 10-30 | 否 |
| **否** | 是 | 低 | 快餐 | 30-60 | 是 |

由上方表格可以看出有2个是，4个否，可以计算出上方表格中数据的熵为：

下面对各个属性计算对应的信息增益：

|  |  |  |
| --- | --- | --- |
| **是否有其他选择** | **是/熵** | **否／增益** |
| **是** | 1 | 3 |
| **否** | 1 | 1 |
| **熵** | 0.8742 | 0.044 |

|  |  |  |
| --- | --- | --- |
| **饿否** | **是/熵** | **否／增益** |
| **是** | 2 | 2 |
| **否** | 0 | 2 |
| **熵** | 0.6667 | 0.2515 |

|  |  |  |
| --- | --- | --- |
| **价格** | **是/熵** | **否／增益** |
| **低** | 2 | 2 |
| **中** | 0 | 0 |
| **高** | 0 | 2 |
| **熵** | 0.6667 | 0.2515 |

|  |  |  |
| --- | --- | --- |
| **餐馆类型** | **是/熵** | **否／增益** |
| **中式** | 1 | 1 |
| **意大利式** | 0 | 1 |
| **法式** | 0 | 1 |
| **快餐** | 1 | 1 |
| **熵** | 0.6667 | 0.2515 |

|  |  |  |
| --- | --- | --- |
| **等待时间** | **是/熵** | **否／增益** |
| **0-10** | 0 | 0 |
| **10-30** | 1 | 1 |
| **30-60** | 1 | 1 |
| **＞60** | 0 | 2 |
| **熵** | 0.6667 | 0.2515 |

根据以上数据可以得出饿否、价格、餐馆类型与等待时间的增益相同且最高均为0.2515，此处任选一个作为节点，我选饿否。

饿否为否时均为否，熵为0，可作为叶节点。

故只需分析饿否为是的情况。

饿否为是时的数据为：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **是否有其他选择** | **价格** | **餐馆类型** | **等待时间** | **是否等待** |
| **是** | 低 | 中餐 | 30-60 | 否 |
| **是** | 低 | 中餐 | 10-30 | 是 |
| **是** | 高 | 意大利式 | 10-30 | 否 |
| **否** | 低 | 快餐 | 30-60 | 是 |

由上方表格可以看出有2个是，2个否，可以计算出上方表格中数据的熵为：

下面对各个属性计算对应的信息增益：

|  |  |  |
| --- | --- | --- |
| **是否有其他选择** | **是/熵** | **否／增益** |
| **是** | 1 | 2 |
| **否** | 1 | 0 |
| **熵** | 0.6887 | 0.3113 |

|  |  |  |
| --- | --- | --- |
| **价格** | **是/熵** | **否／增益** |
| **低** | 2 | 1 |
| **中** | 0 | 0 |
| **高** | 0 | 1 |
| **熵** | 0.6887 | 0.3113 |

|  |  |  |
| --- | --- | --- |
| **餐馆类型** | **是/熵** | **否／增益** |
| **中式** | 1 | 1 |
| **意大利式** | 0 | 1 |
| **法式** | 0 | 0 |
| **快餐** | 1 | 0 |
| **熵** | 0.5 | 0.5 |

|  |  |  |
| --- | --- | --- |
| **等待时间** | **是/熵** | **否／增益** |
| **0-10** | 0 | 0 |
| **10-30** | 1 | 1 |
| **30-60** | 1 | 1 |
| **＞60** | 0 | 0 |
| **熵** | 1 | 0 |

根据上述数据可以得出餐馆**类型**的信息增益最大，所以选择顾客人数作为根节点的测试属性。

餐馆类型为快餐时，均为是。

餐馆类型为意大利式时，均为否。

以上二者的熵均为0，将其作为叶节点。

故现在只需要分析餐馆类型为中餐的情况。

餐馆类型为中餐时数据如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **是否有其他选择** | **价格** | **等待时间** | **是否等待** |
| **是** | 低 | 30-60 | 否 |
| **是** | 低 | 10-30 | 是 |

由上方表格可以看出有2个是，2个否，可以计算出上方表格中数据的熵为：

下面对各个属性计算对应的信息增益：

|  |  |  |
| --- | --- | --- |
| **是否有其他选择** | **是/熵** | **否／增益** |
| **是** | 1 | 1 |
| **否** | 0 | 0 |
| **熵** | 1 | 0 |

|  |  |  |
| --- | --- | --- |
| **价格** | **是/熵** | **否／增益** |
| **低** | 1 | 1 |
| **中** | 0 | 0 |
| **高** | 0 | 0 |
| **熵** | 1 | 0 |

|  |  |  |
| --- | --- | --- |
| **等待时间** | **是/熵** | **否／增益** |
| **0-10** | 0 | 0 |
| **10-30** | 1 | 0 |
| **30-60** | 0 | 1 |
| **＞60** | 0 | 0 |
| **熵** | 0 | 1 |

根据上述数据可以得出等待时间的信息增益最大，所以选择顾客人数作为根节点的测试属性。

等待时间为10-30时，为是。

等待时间为30-60时，为否。

以上二者的熵均为0，将其作为叶节点。

**至此无法形成新的节点，从而生成了一棵“决策树”。**

**决策树如下图所示：**

