## 决策树学习分类

**15191011 段易通**

1. **背景介绍**

决策树（Decision Tree）是一种通过树状结构判断来进行分类的机器学习方法，有些类似于搜索树，它通过每次对样本不同特征的判别来获得一个最终的分类结果。在这个过程中，每到达一个内部节点，都需要对样本的某种特性做出一个判断，并根据判断结果前往相应的下一个节点，最终到达叶节点，即得到分类结果。

1. **算法原理**
2. **信息熵**

在构建决策树的过程中，在每一个节点，我们都需要对样本做出一个最优的决策，为此算法中引入了信息熵的概念，类比于物理中热学熵的概念，它用来表述一组已知样本集的“混乱程度”，熵越大，样本集越“混乱”，越 “不纯粹”，即不同类样本的比例较为平均。于是有定义：

1. **信息增益**

对于决策树的每一次决策，我们当然希望它能将样本集尽可能地分开，也就是说使得分类后的信息熵尽可能减小，为此算法中提出了信息增益的概念，信息增益越大，则决策的分类效果越好，那么我们就可以通过比较不同决策的信息增益大小来决定采用哪一种决策效果最佳。信息增益的定义为：

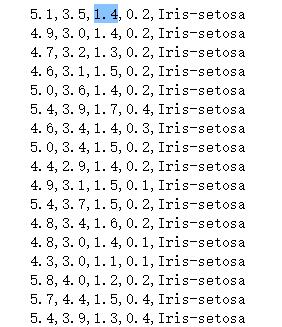
其中代表分类后的某一样本子集的样本数

1. **算法设计**

为实现决策树分类，我们采用了MATLAB语言进行程序的编写，首先通过读给定的Iris数据文件，如图一所示。获取已知的样本集合后，再利用上述的判断方法进行决策，实现对已知集合的分类，然后就可以根据分类时给定的最优的判断条件，对于新的未知样本进行分类，如图二所示。

**图二 决策树分类**

**图一 数据集**



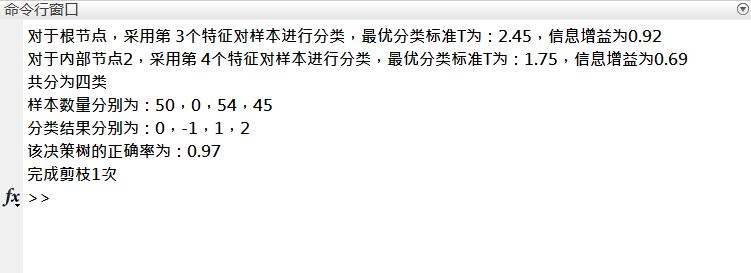
连续的

1. **连续属性**

对于所给的Iris数据集（图一），可以看出每一个特征都为连续的，属性不再有限，也就是说我们无法按照固定的离散属性对其进行分类，因此我们考虑将连续的属性离散化，通过选定一个最好的分类标准，将大量的连续属性分为两个部分，而评判最优标准方法依然是信息增益。则有

我们所要选择的就是对于信息增益来说最大的划分点t，对于Iris数据集，分类过程及结果如图三所示。

**图三 决策树分类**



其中第2、4类没有数据样本，故没有分类结果，标记为-1。

1. **决策树的规模**

通过观察图三中的结果，我们发现，其实当进行了第一次决策之后，该数据集就已经被完美分离开了，因此对于该数据集，高度为1的决策树已经足够（进行一次决策），故应该对决策树进行剪枝，在这里我们采用了预剪枝的方法，从根节点开始，自上而下地进行判断，通过对剪枝后的样本信息增益进行比较，来决定是否剪枝，其原理如图四所示：

**图四 决策树剪枝**

对于剪枝我们需要小心，如果剪枝过多，就会使得决策树的规模过小，这样就会有欠拟合的风险；而若在一些情况下不进行剪枝，又可能会出现过拟合的现象。

1. **决策树的“软边界”**

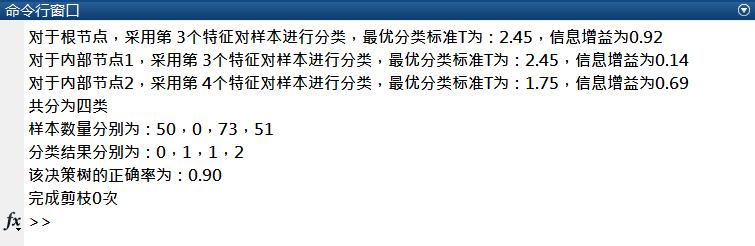
对上述讨论的决策均是基于“硬边界”，也就是说我们严格地按照小于和大于的方法来对样本进行分类，这样做则可能会造成一些边缘样本的分类错误，而为了实现软边界的划分，如图五所示，我们考虑将划分标准不再定位某一个确定的值，而是改为一个区间，这样我们只有样本的属性大于或小于整个区间时，才会被归为分一类，否则就将该样本同时归于两类，等待之后的进一步决策。只不过这样可能会造成一些样本最后被同时分为两类，如图六所示。

**图五 决策树剪枝**

**大于或者处于区间**

**小于或者处于区间**

**图六 软边界分类结果**



1. **总结思考**
2. **决策树与朴素贝叶斯**

归于两类决策树模型和朴素贝叶斯模型是目前应用最为广泛的两种分类模型。

决策树模型通过构造一颗树通过每个内部节点的决策来解决分类问题，它利用训练数据集来构造一棵决策树之后，就可以对为未知的新样本进行分类。决策树模型有着很多的优点，它很符合人类的思维方式，而且根据决策树可以很容易地构造出规则；另一方面，决策树可以用来对具有许多属性的数据集进行分类。决策树模型也有一些缺点，比如剪枝导致过拟合或欠拟合问题的出现，以及忽略数据集中属性之间的相关性等。

和决策树模型相比，朴素贝叶斯模型来自于数学，这种模型虽然抽象一些，但是算法简单而稳定。在一些诸如对缺失参数的样本进行分类的情况下，它往往能够表现出更加优良的效果。

1. **代码链接**

GitHub链接：https://github.com/dythuhu/decisiontree.git