**实验五 决策树算法ID3**

**【实验目的】**

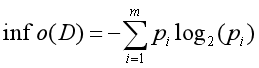
编程实现决策树算法ID3；理解算法原理。

【**实验内容**】

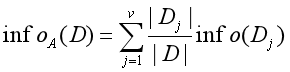
利用taindata.txt的数据（75\*5，第5列为标签）进行训练，构造决策树；利用构造好的决策树对testdata.txt的数据进行分类，并输出分类准确率。

【**实验原理**】

ID3算法的核心思想就是以信息增益度量属性选择，选择分裂后信息增益最大的属性进行分裂。下面先定义几个要用到的概念。设D为用类别对训练元组进行的划分，则D的[熵](http://en.wikipedia.org/wiki/Entropy)（entropy）表示为：



其中pi表示第i个类别在整个训练元组中出现的概率，可以用属于此类别元素的数量除以训练元组元素总数量作为估计。熵的实际意义表示是D中元组的类标号所需要的平均信息量。现在我们假设将训练元组D按属性A进行划分，则A对D划分的期望信息为：



而信息增益即为两者的差值：

*gain(A)=info(D)-infoA(D)*

ID3算法就是在每次需要分裂时，计算每个属性的增益率，然后选择增益率最大的属性进行分裂。

对于特征属性为连续值，可以如此使用ID3算法：先将D中元素按照特征属性排序，则每两个相邻元素的中间点可以看作潜在分裂点，从第一个潜在分裂点开始，分裂D并计算两个集合的期望信息，具有最小期望信息的点称为这个属性的最佳分裂点，其信息期望作为此属性的信息期望。

【**实验要求**】

（1）提交源代码及可执行文件。

（2）提交实验报告，内容包括：对代码的简单说明、运行结果的截图及说明等。

（3）需画出决策树，指明每个分支所对应的特征/属性，以及分裂值。

注：如用到了剪枝、限定深度等技巧（加分项），请加以说明。