**实验二 预测隐形眼镜类型的决策树模型**

1. 实验目的和任务

（1）了解决策树算法原理。

（2）掌握ID3决策树生成算法、决策树剪枝。

（3）编程实现ID3算法。

（4）了解随机森林算法原理，利用sklearn实现随机森林算法。

（5）掌握sklearn的调参方法。

1. 实验预习与思考

预习ID3决策树生成算法的相关内容，编写实现代码。

1. 实验原理

决策树算法是一种逼近离散函数值的方法。它是一种典型的[分类方法](http://baike.baidu.com/view/8348989.htm)，首先对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策对新数据进行分析。本质上决策树是通过一系列规则对数据进行分类的过程。决策树方法的基本思想是：利用训练集数据自动地构造决策树，然后根据这个决策树对任意实例进行判定。其中决策树（Decision Tree）是一种简单但是广泛使用的分类器。通过训练数据构建决策树，可以高效的对未知的数据进行分类。决策数有两大优点：1）决策树模型可以读性好，具有描述性，有助于人工分析；2）效率高，决策树只需要一次构建，反复使用，每一次预测的最大计算次数不超过决策树的深度。

决策树算法构造决策树来发现数据中蕴涵的分类规则．如何构造精度高、规模小的决策树是决策树算法的核心内容。决策树构造可以分两步进行。第一步，决策树的生成：由训练样本集生成决策树的过程。一般情况下，训练样本数据集是根据实际需要有历史的、有一定综合程度的，用于数据分析处理的数据集。第二步，决策树的剪技：决策树的剪枝是对上一阶段生成的决策树进行检验、校正和修下的过程，主要是用新的样本数扼集（称为测试数据集）中的数据校验决策树生成过程中产生的初步规则，将那些影响预衡准确性的分枝剪除、

决策树方法最早产生于上世纪60年代，到70年代末。由J Ross Quinlan提出了[ID3算法](http://baike.baidu.com/view/3098338.htm)，此算法的目的在于减少树的深度。但是忽略了叶子数目的研究。C4.5算法在ID3算法的基础上进行了改进，对于预测变量的缺值处理、剪枝技术、派生规则等方面作了较大改进，既适合于分类问题，又适合于回归问题。

本节将就ID3算法展开分析和实现。

ID3算法最早是由罗斯昆（J. Ross Quinlan）于1975年在悉尼大学提出的一种分类预测算法，算法的核心是“信息熵”。ID3算法通过计算每个属性的信息增益，认为信息增益高的是好属性，每次划分选取信息增益最高的属性为划分标准，重复这个过程，直至生成一个能完美分类训练样例的决策树。

在ID3算法中，决策节点属性的选择运用了信息论中的熵概念作为启发式函数。

在这种属性选择方法中，选择具有最大信息增益（information gain）的属性作为当前划分节点。通过这种方式选择的节点属性可以保证决策树具有最小的分枝数量，使得到的决策树冗余最小。

1. 实验内容

（1）收集数据，提供文本文件。

（2）准备数据，解析tab键分隔的数据行。

（3）分析数据，快速检查数据，确保正确地解析数据内容。

（4）训练算法，实现ID3决策树算法。

（5）测试算法，编写测试函数验证决策树可以正确分类给定的数据实例。

（6）使用算法，存储树的数据结构，以便下次使用时，无需重新构造树。

1. 实验报告要求

1、认真记录实验数据。

2、需要详细算法设计的源代码。

1. 实验仪器设备

（1）PC电脑

（2）pyhton

七. 思考题

分析ID3算法的优略和局限性，如何修改能够使模型更加健壮。提示：分析ID3算法和C4.5的不同。