**决策树**

1.决策树的优缺点：

优点：计算复杂度不高，输出结果易于理解，对中间值的缺失不敏感，可以处理不相关的特征数据。

缺点：可能会产生过度匹配的问题。

适用数据类型：数值型和标称型。

2.构造决策树：

1.决策树步骤：

1.收集数据：可以使用任何方法收集数据。

2.准备数据：树构造算法只适用于标称型数据，因此数值型数据必须离散化。

3.分析数据：构造树完成之后，我们应该检查图形是否符合预期。

4.训练算法：构造树的数据结构

5.测试算法：使用经验树计算错误率。

6.使用算法：这个步骤适用于任何监督学习算法，而使用决策树可以更好的理解数据的内在含义。

2.id3算法：

ID3算法是决策树的一种，它是基于奥卡姆剃刀原理的，即用尽量用较少的东西做更多的事。ID3算法，即Iterative Dichotomiser 3，迭代二叉树3代，是Ross Quinlan发明的一种决策树算法，这个算法的基础就是上面提到的奥卡姆剃刀原理，越是小型的决策树越优于大的决策树，尽管如此，也不总是生成最小的树型结构，而是一个启发式算法。

在信息论中，期望信息越小，那么信息增益就越大，从而纯度就越高。ID3算法的核心思想就是以信息增益来度量属性的选择，选择分裂后信息增益最大的属性进行分裂。该算法采用自顶向下的贪婪搜索遍历可能的决策空间。

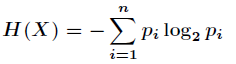
1. 信息熵与信息增益：

在信息增益中，重要性的衡量标准就是看特征能够为分类系统带来多少信息，带来的信息越多，该特征越重要。在认识信息增益之前，先来看看**信息熵**的定义

**熵**这个概念最早起源于物理学，在物理学中是用来度量一个热力学系统的无序程度，而在信息学里面，熵是对不确定性的度量。在**1948**年，**香农**引入了**信息熵**，将其定义为离散随机事件出现的概率，一个系统越 是有序，信息熵就越低，反之一个系统越是混乱，它的信息熵就越高。所以信息熵可以被认为是系统有序化程度的一个度量。

   假如一个随机变量IMG_256的取值为IMG_257，每一种取到的概率分别是IMG_258，那么

IMG_259的熵定义为

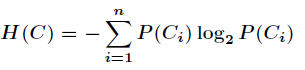


意思是一个变量的变化情况可能越多，那么它携带的信息量就越大。

 对于**分类系统**来说，类别IMG_261是变量，它的取值是IMG_262，而每一个类别出现的概率分别是

IMG_263

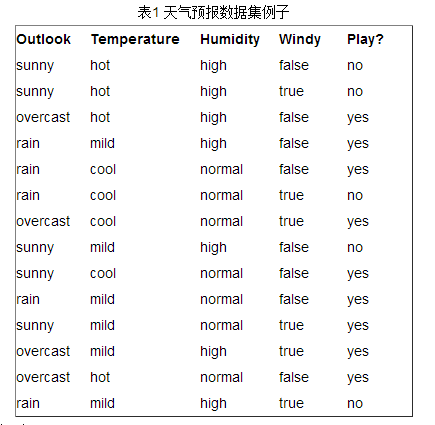
 而这里的IMG_264就是类别的总数，此时分类系统的熵就可以表示为



以上就是信息熵的定义，接下来介绍**信息增益**。

  信息增益是针对一个一个特征而言的，就是看一个特征IMG_266，系统有它和没有它时的信息量各是多少，两者的差值就是这个特征给系统带来的信息量，即**信息增益**。

接下来以**天气预报**的例子来说明。下面是描述天气数据表，学习目标是**play**或者**not play**。

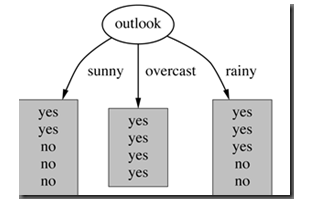


   可以看出，一共**14**个样例，包括**9**个正例和**5**个负例。那么当前信息的熵计算如下

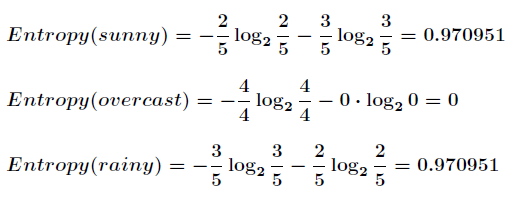
IMG_257

   在决策树分类问题中，信息增益就是决策树在进行属性选择划分前和划分后信息的差值。假设利用

   属性**Outlook**来分类，那么如下图



      划分后，数据被分为三部分了，那么各个分支的信息熵计算如下



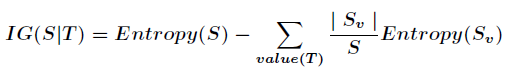
       那么划分后的信息熵为

IMG_260

IMG_261代表在特征属性IMG_262的条件下样本的**条件熵**。那么最终得到特征属性IMG_263带来的信息增益为

IMG_264

**信息增益的计算公式**如下



   其中IMG_266为全部样本集合，IMG_267是属性IMG_268所有取值的集合，IMG_269是IMG_270的其中一个属性值，IMG_271是IMG_272中属性IMG_273的

   值为IMG_274的样例集合，IMG_275为IMG_276中所含样例数。

   在决策树的每一个非叶子结点划分之前，先计算每一个属性所带来的信息增益，选择最大信息增益的属性来划分，因为信息增益越大，区分样本的能力就越强，越具有代表性，很显然这是一种自顶向下的贪心策略。以上就是**ID3算法**的核心思想。

<以上为参考前人的成果，现在我们就以Python来实现信息熵，信息增益，以及决策树的构造，代码在相应的代码文件>