决策树的一个重要的任务就是为了数据中所蕴含的知识信息，因此决策树可以使用不熟悉的数据集合，并从中提取一系列的规则，在这些机器根据数据集创建规则时， 就是机器学习的过程。

在构造决策树的时候，我们需要解决的第一个问题就是，当前数据集上哪一个特征在划分数据分类时起决定性作用，为了找到决定性的特征，划分出最好的结果，我们必须评估每一个特征。

信息增益：

组织杂乱无章数据的一种方式就是使用信息论度量信息，我们可以在划分数据之前或之后使用信息论量化度量信息的内容。在划分数据集之前之后信息发生的变化称为信息增益，知道如何计算信息增益，我们就可以计算每一个特征值划分数据集获得的信息增益，获得信息增益最高的特征就是最好的选择。

信息的定义：

如果待分类的事务可能划分在多个类别中，则类别Xi的信息定义为：

http://img.blog.csdn.net/20140616200705796?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvcGVuZzgyNTIyMzIwOA==/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

其中p(xi)为训练样本中i类的数量除以所有训练样本的总量

熵定义为信息的期望值

http://img.blog.csdn.net/20140616200713937?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvcGVuZzgyNTIyMzIwOA==/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

决策树会非常好的匹配实验数据，然而这些匹配选项可能太多了，我们将这种问题称为过度匹配。为了减少过度匹配的问题，我们可以裁剪决策树，去掉一些不必要的叶子节点。如果叶子节点只能增加少许的信息，这可以删除该节点，将他并入到其他的叶子节点中。本章使用的算法为ID3，他是一个好的算法但并不完美，ID3算法无法直接处理数值型数据，尽管我们可以通过量化的方法将数值型数据转换为标称型数据，但是如果村子太多的特征划分，ID3算法仍然会面临其他问题。

优点：

计算复杂读不高，输出结果易于理解，对中间值的缺失不敏感，可以处理不相关特征数据

缺点：

可能会产生过度匹配问题

适应数据类型：

数值型和标称形