决策树

机器学习笔记 create by siwanghu v1.0

决策树是一种常见的机器学习分类方法

ID3算法

输入 训练集D={(,),(),(),……(,)}

属性集A={}

treeGenerate(D,A):

生成结点

if D中样本全属于同一类别C then 将node标记为C类叶结点 return

if A 是空集 OR D中样本在A是取值相同 then 将node标记为叶结点，类别为D中样本数最多的类 return

从A中求出所有属性的信息增益，选择信息增益最大的属性作为划分属性

for in ：

为node生成一个分支，令表示D中在上的取值为的子集

if 是空集 then 将分支结点标记为叶结点，其类别为D中样本最多的类 return

else 以 treeGenerate()为分支结点

信息

信息熵 度量样本集合纯度，熵越小样本纯度越高

代表样本D中第k类样本所占的比重，代表样本D中类别总数

信息增益 决策树中使用信息增益最大的属性进行划分，代表用这个属性划分，样本所获得的纯度提升越大

假设离散属性a有v种取值，若使用a划分，则会产生v个分支结点。第v个分支结点包

了样本D中属性a中所有取值为的样本，记为