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

信息　　$I = -\log_2 P_k$

信息熵　度量样本集合纯度，熵越小样本纯度越高

$$Ent(D) = -\sum_{k=1}^{|Y|} P_k log_2 P_k$$

P_k代表样本 D 中第 k 类样本所占的比重，$|Y|$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

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

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a^v的样本，记为D^v

ID3 算法

输入 训练集 D=$\{(x_1,y_1),(x_2,y_2),(x_3,y_3),\ldots\ldots(x_n,y_n)\}$

属性集 A=$\{a_1, a_2, \ldots a_n\}$

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 a_*^v in a_* ：

为 node 生成一个分支，令D_v表示 D 中在a_*上的取值为a_*^v的子集

if D_v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($D_v, A/\{a_*\}$)为分支结点