西安電子科技大學
XIDIAN UNIVERSITY

智能数据挖掘大作业

题目	决策树
姓名	鲍超俊
学号	15020510059
指导老师	缑水萍

一、决策树

1. 算法简介

机器学习中，决策树是一个预测模型；他代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。决策树仅有单一输出，若欲有复数输出，可以建立独立的决策树以处理不同输出。 数据挖掘中决策树是一种经常要用到的技术，可以用于分析数据，同样也可以用来作预测。从数据产生决策树的机器学习技术叫做决策树学习,通俗说就是决策树。

2. 决策树ID3算法的信息论基础

> 熵

熵度量了事物的不确定性，越不确定的事物，它的熵就越大。具体的，随机变量X的熵的表达式如下：

$$H(X) = -\sum_{i=1}^{n} p_i log_2 p_i$$

其中n代表X的n种不同的离散取值。p_i代表了X取第i个离散值的概率，log_2也可以换成ln.

> 联合熵

随机变量X 和Y的联合熵定义如下：

$$H(X,Y) = -\sum_{i=1}^{n} p(x_i, y_i) log_2 p(x_i, y_i)$$

> 条件熵

有了联合熵，又可以得到条件熵的表达式$H(X|Y)$，条件熵类似于条件概率,它度量了我们的X在知道Y以后剩下的不确定性。表达式如下：

$$H(X|Y) = -\sum_{i=1}^{n} p(x_i, y_i) log p(x_i|y_i) = \sum_{j=1}^{n} p(y_j) H(X|y_j)$$

3. 决策树ID3算法步骤

1) 初始化信息增益的阈值ϵ.

2) 判断样本是否为同一类输出D_i，如果是则返回单节点树T。标记类别为D_i.

3) 判断特征是否为空，如果是则返回单节点树T，标记类别为样本中输出类别D实例数最多的类别。

4) 计算A中的各个特征（一共n个）对输出D的信息增益，选择信息增益最大的特征A_g

5) 如果A_g的信息增益小于阈值ϵ，则返回单节点树T，标记类别为样本中输出类别D实例数最多的类别。

6) 否则，按特征A_g的不同取值A_{gi}将对应的样本输出D分成不同的类别D_i. 每个类别产生一个子节点。对应特 征值为A_{gi}。返回增加了节点的数T。

7)对于所有的子节点，令$D = D_i, A = A - \{A_g\}$递归调用**2) - 6)**步，得到子树T_i并返回。

二、实验报告

1. 实验数据

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rain	Mild	High	Weak	Yes
D_5	Rain	Cool	Normal	Weak	Yes
D_6	Rain	Cool	Normal	Strong	No
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rain	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rain	Mild	High	Strong	No

2. 实验结果

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rain	Mild	High	Weak	Yes
D_5	Rain	Cool	Normal	Weak	Yes
D_6	Rain	Cool	Normal	Strong	Yes
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rain	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rain	Mild	High	Strong	No

3. 程序说明

> DT.py

1）定义决策树节点(Node)

2）定义决策树（decision_tree）

> main.py

1）创建决策树对象：tree = DT.decision_tree()

2）训练决策树：decision_tree.fit()

3）预测：decision_tree.predict()