

机器学习

Homework4:决策树

线性回归

题1（88分）：根据下面给出的西瓜数据集，利用决策树学习基本算法实现西瓜的自动分类。使用信息增益选择最优化分属性。

画出决策树（画决策树不需要写程序，可手工画图）。

文件data4a.xlsx, data4b.xlsx都是这个训练数据集， data4b.xlsx将文字用数字替换了。

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

决策树学习基本算法

Algorithm 1 决策树学习基本算法

输入:

- 训练集 $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$;
- 属性集 $A = \{a_1, \dots, a_d\}$.

过程: 函数 TreeGenerate(D, A)

- 1: 生成结点 node;
- 2: if D 中样本全属于同一类别 C then
- 3: 将 node 标记为 C 类叶结点; return
- 4: end if
- 5: if $A = \emptyset$ OR D 中样本在 A 上取值相同 then
- 6: 将 node 标记叶结点, 其类别标记为 D 中样本数最多的类; return
- 7: end if
- 8: 从 A 中选择最优划分属性 a_* ;
- 9: for a_* 的每一个值 a_*^v do
- 10: 为 node 生成每一个分枝; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
- 11: if D_v 为空 then
- 12: 将分枝结点标记为叶结点, 其类别标记为 D 中样本最多的类; return
- 13: else
- 14: 以 TreeGenerate($D_v, A - \{a_*\}$) 为分枝结点
- 15: end if
- 16: end for

输出: 以 node 为根结点的一棵决策树

线性回归

题2（12分）：对于上题给出的西瓜数据集，分别使用预剪枝和后剪枝技术生成决策树（仍旧使用信息增益选择最优化分属性）。

画出决策树（画图不需要写程序）。

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

实验报告内容

1、实验内容

实验要解决的问题、采用的模型或算法等

2、实验设置和实验结果

3、其它（其它你觉得需要写在实验报告中的内容）

4、实验过程中遇到的问题

5、实验心得体会。

注意事项：

1、实验报告请使用老师提供的实验模板，源代码作为单独的文件。

2、实验报告命名：完整学号_姓名_ML_project4.doc

3、建立个人文件夹放实验报告和源代码(源代码需加注释说明)，文件夹名“完整学号_姓名_ML_project4”。

4、严禁抄袭，一旦发现，按不及格处理。

报告提交时间：11月11日周三下午3点前。