稀牛学院 + 网易云课堂

机器学习工程师微专业
作业及答案

03 树模型初步与进阶

第一章 决策树与分类

讲师：寒小阳

选择题（2 分/题）

1.下列关于信息熵描述错误的是：A

A.分类样本集 D 的信息熵 Ent(D)越小，数据纯度越低

B.信息熵可以用于描述不确定程度

C.分类样本集 D 的信息熵 Ent(D)越小，数据纯度越高

D.决策树可以基于信息熵计算信息增益，用于最优划分属性选择

2.ID3 和 CART 分别是基于什么进行的最优属性选择：C

A.信息增益，信息增益率

B.信息增益率，基尼指数

C.信息增益，基尼指数

D.基尼指数，信息增益率

3.下列描述不正确的是：B

A.在分类数据集中，基尼指数越大，纯度越低

B.分类问题中，CART 算法会选择划分后基尼指数最大的属性作为分裂属性

C.ID3 使用信息增益进行最优属性选择，会倾向于选择取值多的属性

D.让决策树在训练集上生长完全，得到的模型可能会有过拟合的风险

4.关于剪枝，说法错误的是：C

A.剪枝是通过主动去掉一些分支来缓解模型过拟合的风险

B.可以用留出法评估剪枝前后的模型好坏

C.后剪枝的效果一定比预剪枝好

D.预剪枝会提前终止某些分支的生长

判断题（2 分/题）

1.C4.5 是使用信息增益率进行最优属性的选择： 对

2.在分类数据集上，基尼指数和数据纯度成正相关：错

3.后剪枝通常效果优于预剪枝，但是训练时间开销大：对

4.CART 分类决策树是二叉树形态：对

问答题（10 分/题）

请简述三种分类决策树的最优属性选择方法：

答：课程中讲到 ID3、C4.5、CART 3 种方法用于解决分类问题。核心问题是最优划分属性的选择，针对这个问题，3 种方法提出了不同的最优划分属性选择方法。

① ID3 选用信息增益作为衡量标准(公式详见 PPT)，选择信息增益最大的属性作为划分属性，缺点是对数量多的属性可能会有偏袒。

② C4.5 选择信息增益率作为衡量标准(公式详见 PPT)，选择信息增益率最大的属性作为划分属性。

③ CART 选择划分后基尼指数最小的属性作为划分属性。

课程链接：http://course.study.163.com/400000002658002/learning

如有问题，请咨询稀牛学院客服微信