

## 第七次作业

阅读机器学习实战 第6章 决策树部分

### Exercise 7.1

决策树的父节点和子节点的熵的大小关系是什么？

- A、两者相等
- B、子节点的熵更大
- C、父节点的熵更大
- D、不确定

### Exercise 7.2

决策树中属性选择的方法不包括（）

- A、信息值
- B、信息增益
- C、信息增益率
- D、GINI系数

### Exercise 7.3

以下关于决策树特点分析的说法错误的有（）。

- A. 算法自动忽略了对模型没有贡献的属性变量
- B. 推理过程容易理解，计算简单
- C. 算法容易造成过拟合
- D. 算法考虑了数据属性之间的相关性

### Exercise 7.4

以下关于决策树原理介绍错误的有（）。

- A. 决策树算法属于无监督学习
- B. 决策树决策过程从根节点开始
- C. 决策树算法本质上是贪心算法
- D. 决策树生成过程中需要用到分割法

### Exercise 7.5

我们想要在大数据集上训练决策树模型，为了使用较少的时间，可以：（）。

- A. 增加树的深度
- B. 增大学习率
- C. 减少树的深度
- D. 减少树的数量

## Exercise 7.6

以下那种说法是错误的( )。

- A. 中国足球队战胜巴西足球队的信息熵要小于中国乒乓球队战胜巴西乒乓球队的信息熵
- B. 信息增益 = 信息熵 - 条件熵
- C. 一个系统越是有序，信息熵就越低
- D. 一个系统越是混乱，随机变量的不确定性就越大，信息熵就越高

## Exercise 7.7

关于C4.5算法，错误的是( )。

- A. C4.5算法采用基尼系数的大小来度量特征的各个划分点
- B. C4.5算法引入悲观剪枝策略进行后剪枝
- C. C4.5算法可以处理非离散的数据
- D. C4.5 算法最大的特点是克服了 ID3 对特征数目的偏重这一缺点

## Exercise 7.8

关于CART算法，错误的是( )。

- A. CART 分类树采用基尼系数的大小来度量特征的各个划分点
- B. CART算法既可以处理分类问题，也可以处理回归问题
- C. 可以处理样本不平衡问题
- D. CART算法采用信息增益率的大小来度量特征的各个划分点

## Exercise 7.9

考虑表中二元分类问题的训练样本集

实例	$a_1$	$a_2$	$a_3$	目标类
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

整个训练样本集关于类属性的熵是多少？关于这些训练集， $a_1$ ， $a_2$ ， $a_3$ 的信息增益分别是多少？

### Exercise 7.10

利用CART决策树方法，根据职业和年龄来预测月薪

职业	年龄	月薪
程序员	22	20000
程序员	23	26000
程序员	29	30000
教师	23	12000
教师	25	14000

### Exercise 7.11

如果决策树对训练集欠拟合，尝试缩放输入特征是否为一个好主意？