第七次作业参考答案

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

解析：父节点的熵实际上是所有子节点熵的加权平均数

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

解析：后面三种是常用的分类问题的属性选择方法。

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

解析：决策树是单独检查每一个特征的，忽略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

解析：决策树算法属于监督学习

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

解析：树的深度决定了模型的复杂程度。

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

解析：发生概率较少的事件具有更大的信息熵

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

解析：C4.5算法采用信息增益率度量各个划分点

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

解释：CART采用基尼系数来度量各个特征的划分点，分为分类树和回归树

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？

解析:

总熵为:

$$
\begin{aligned}
Entropy(S) &= -\sum_{i=1}^{m}\frac{s_i}{s}log_2\frac{s_i}{s} \\
&= -\frac{4}{9}*log_2\frac{4}{9}-\frac{5}{9}*log_2\frac{5}{9} \\
&= 0.9911
\end{aligned}
$$

a1的信息增益为

$$
\begin{aligned}
Gain(a1,\ S) &= Entropy(S)-Entropy(a1,\ S) \\
&= 0.9911-\frac{4}{9}Entropy(3,1)-\frac{5}{9}Entropy(1,4) \\
&= 0.9911-\frac{4}{9}(-\frac{1}{4}log_2\frac{1}{4}-\frac{3}{4}log_2\frac{3}{4})-\frac{5}{9}(-\frac{1}{5}log_2\frac{1}{5}-\frac{4}{5}log_2\frac{4}{5}) \\
&= 0.9911-0.7616 \\
&= 0.2295
\end{aligned}
$$

a2的信息增益为

$$
\begin{aligned}
Gain(a2,\ S) &= Entropy(S)-Entropy(a2,\ S) \\
&= 0.9911-\frac{4}{9}Entropy(2,3)-\frac{5}{9}Entropy(2,2) \\
&= 0.9911-\frac{4}{9}(-\frac{2}{5}log_2\frac{2}{5}-\frac{3}{5}log_2\frac{3}{5})-\frac{5}{9}(-\frac{1}{2}log_2\frac{1}{2}-\frac{1}{2}log_2\frac{1}{2}) \\
&= 0.9911-0.9839 \\
&= 0.0072
\end{aligned}
$$

a3的信息增益为,此处考虑以2作为划分

$$
\begin{aligned}
Gain(a3,\ S) &= Entropy(S)-Entropy(a3,\ S) \\
&= 0.9911-\frac{1}{9}Entropy(1,0)-\frac{8}{9}Entropy(3,5) \\
&= 0.9911-\frac{8}{9}(-\frac{3}{8}log_2\frac{3}{8}-\frac{5}{8}log_2\frac{5}{8}) \\
&= 0.9911-0.8484 \\
&= 0.1427
\end{aligned}
$$

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

解析:

CART全称叫Classification and Regression Tree。首先要强调的是CART假设决策树是二叉树。作为分类决策树时，待预测样本落至某一叶子节点，则输出该叶子节点中所有样本所属**类别最多的那一类**（即叶子节点中的样本可能不是属于同一个类别，则多数为主）；作为回归决策树时，待预测样本落至某一叶子节点，则**输出该叶子节点中所有样本的均值。**

程序员月薪均值为：(20000+26000+30000)/3=25333.33

教师月薪均值为：(12000+14000)/2=13000

- 以职业为切分点

  平方误差：$(20000-25333.33)^2$ +$(26000-25333.33)^2$ +$(30000-25333.33)^2$ +$(12000-13000)^2$ +$(14000-13000)^2$ =51666666+2000000 = 53666666

- 以年龄为切分点
  - 以22为切分点
    - 第一类：年龄≤22，月薪均值：20000
    - 第二类：年龄>22,月薪均值：(26000+30000+12000+14000)/4=20500
    - 平方误差：$(20000-20000)^2$ +$(26000-20500)^2$+$(30000-20500)^2$+$(12000-20500)^2$+$(14000-20500)^2$= 235000000
  - 以23为切分点
    - 第一类：年龄≤23，月薪均值：(20000+12000+26000)/3=19333.33
    - 第二类：年龄>23,月薪均值：(30000+14000)/2=22000
    - 平方误差：$(20000-19333.33)^2$+$(26000-19333.33)^2$+$(30000-22000)^2$+$(12000-22000)^2$+$(14000-22000)^2$=272888938
  - 以25为切分点
    - 第一类：年龄≤25，月薪均值：(20000+12000+26000+14000)/4=18500
    - 第二类：年龄>25,月薪均值：30000
    - 平方误差：$(20000-18500)^2$+$(26000-18500)^2$+$(30000-18500)^2$+$(12000-18500)^2$+$(30000-30000)^2$=233000000

  故以职业为切分点预测月薪，如果职业为程序员，则月薪预测值为25333.33；如果职业为教师，则月薪预测值为13000。

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？

解析：决策树的优点之一就是它们不关心训练数据是缩放还是集中，所以如果决策树不适合训练集，缩放输入特征不过是浪费时间罢了。