决策树习题

浙江大学

赵洲

题目

- **1. 以下关于决策树的描述，哪一项是正确的?**

A. 决策树只能处理离散型特征

B. 剪枝可以防止决策树过拟合

C. 决策树模型不需要任何参数调整

D. 决策树无法处理多分类问题

解答

- A. 决策树只能处理离散型特征。
- **错误**

- 决策树不仅可以处理离散型特征，也可以处理连续型特征。在处理连续型特征时，决策树会选择一个阈值，然后根据该阈值将数据分为两部分。这是决策树处理连续数据的常见方式。
- **结论**：决策树不仅可以处理离散型特征，也能处理连续型特征，因此这个选项是错误的。

解答

- **B. 剪枝可以防止决策树过拟合**

- **正确**：
  - 剪枝是决策树训练过程中的一个重要步骤，用于防止模型过拟合。过拟合通常发生在决策树过于复杂时（例如树的深度过大，或包含过多的分支），剪枝通过去除一些不必要的分支来简化树的结构，从而提高模型的泛化能力。
  - 剪枝有两种常见的方式：
    - **预剪枝（Pre-pruning）**：在树的生长过程中，提前停止分裂，以限制树的复杂度。
    - **后剪枝（Post-pruning）**：在决策树完全生长后，去除一些不必要的叶子节点或分支。
  - **结论**：剪枝确实有助于防止过拟合，因此此选项是正确的。

解答

- **C. 决策树模型不需要任何参数调整**
- **错误**：
  - 虽然决策树是一个易于理解和实现的算法，但它并非无需调参。决策树的性能受多种超参数的影响，包括：
    - **树的最大深度**（max_depth）：控制树的复杂度，过深的树容易过拟合，过浅的树可能欠拟合。
    - **每个叶子节点的最小样本数**（min_samples_leaf）：控制树的分裂是否足够精细。
    - **分裂节点的最小样本数**（min_samples_split）：决定内部节点是否继续分裂。
    - **最大特征数**（max_features）：控制在每次分裂时考虑的特征数量。
  - 因此，决策树模型需要调整和选择合适的超参数来优化性能。
  - **结论**：决策树需要进行参数调整，因此此选项是错误的。

解答

- **D. 决策树无法处理多分类问题**

- **错误**：
  - 决策树不仅能够处理二分类问题，也能够处理多分类问题。在多分类任务中，决策树通过选择最优的特征和切分点来进行分裂，直到树达到停止条件（例如，达到最大深度或满足最小样本数要求）。在每个叶子节点，决策树会选择一个类别作为最终的预测结果。
  - **结论**：决策树可以处理多分类问题，因此此选项是错误的。

题目

- **2. 在构建决策树时，信息增益最大的特征通常被优先选择用于分裂。信息增益基于以下哪种度量?**

A. Gini不纯度

B. 熵

C. 方差

D. 均值绝对误差

题目

- **2. 在构建决策树时，信息增益最大的特征通常被优先选择用于分裂。信息增益基于以下哪种度量?**

A. Gini不纯度
B. 熵
C. 方差
D. 均值绝对误差

解答

■ A. Gini不纯度

- **错误**:
  - Gini不纯度是决策树中另一个常用的度量，特别是在**CART（Classification and Regression Trees）**算法中。它用来衡量一个节点的不纯度，值越低表示节点越纯净。虽然Gini不纯度也用于选择特征进行分裂，但它与信息增益无关。
  - Gini不纯度和信息增益是两种不同的度量方法。

解答

- **B. 熵**

- **正确**：
  - **信息增益**是基于**熵**的度量。熵是一个度量不确定性的量，信息增益则是通过某一特征进行分裂后，数据集熵的减少量。选择信息增益最大的特征进行分裂，目的是最大限度地减少数据的混乱程度或不确定性。
  - **结论**：信息增益基于熵，因此此选项是正确的。

解答

- **C. 方差**

- **错误**:
  - 方差主要用于回归任务中度量数据点的分散程度，而不是用于分类任务中的决策树分裂。在回归树（例如CART回归树）中，可以使用方差减少来衡量分裂的好坏，但方差不是信息增益的基础。

解答

- **D. 均值绝对误差**

- **错误**：
  - 均值绝对误差（MAE）是回归问题中评估模型预测误差的一种度量，它与决策树的特征选择和分裂标准无关。MAE不用于信息增益的计算。

题目

- **3. 以下哪种算法通常用于生成决策树?**

A. K-均值聚类

B. Apriori算法

C. ID3算法

D. 支持向量机

解答

- **A. K-均值聚类**

- **错误**：
  - **K-均值聚类**是一种常用的聚类算法，用于将数据分为若干个簇。它并不涉及生成决策树。K-均值算法通过最小化簇内数据点的方差来进行聚类，通常用于无监督学习任务。
  - 因此，K-均值算法不用于生成决策树。

解答

- **B. Apriori算法**

- **错误**：
  - **Apriori算法**是一种经典的关联规则学习算法，通常用于发现频繁项集和关联规则。它与决策树的生成无关，主要应用于市场篮分析、推荐系统等任务。
  - Apriori算法与决策树没有直接关系，因此不适用于生成决策树。

解答

- **C. ID3算法**

- **正确**：
  - **ID3算法**（Iterative Dichotomiser 3）是生成决策树的经典算法之一，常用于分类任务。它通过计算信息增益来选择最佳的特征进行分裂。ID3算法是基于**熵**和**信息增益**的，它通过递归地选择信息增益最大的特征，逐步构建决策树。
  - 因此，ID3算法通常用于生成决策树。

解答

■ 支持向量机

- **错误**：
  - **支持向量机（SVM）**是一种常用的分类和回归算法，主要通过寻找一个最佳的超平面来将数据分成不同的类别。它并不是生成决策树的算法，而是一种用于分类的算法。
  - 支持向量机与决策树的生成没有直接关系。

题目

- 判断1.
- 决策树模型容易受到训练数据中噪声和异常值的影响。

解答

- 正确

决策树的敏感性： 决策树通过逐步分裂数据集来构建模型，依赖于数据中的特征和类别分布。当训练数据中存在噪声或异常值时，决策树可能会根据这些异常情况进行错误的分裂。

存在问题： 过拟合、树结构复杂度增加、不稳定性等。

解决方法： 剪枝、数据预处理、集成方法。

题目

- 判断2.
- 在决策树中，叶节点表示决策结果或类别标签。

解答

- 正确

- **叶节点的功能：**

- **决策结果：** 叶节点提供了基于特征分裂的最终决策结果。例如，在二分类问题中，一个叶节点可能代表“是”类别，另一个叶节点代表“否”类别。
- **类别标签：** 叶节点通常包含类别标签的概率分布或多数类标签，用于对新样本进行分类预测。

题目

- 判断3.
- 所有决策树无法处理缺失值，必须要在预处理阶段填补所有缺失值。

解答

- 错误

- **决策树处理缺失值的能力**：虽然在实际应用中，缺失值处理通常在数据预处理阶段进行（如填补缺失值或删除缺失样本），但有一些决策树算法具备处理缺失值的机制，能够在构建过程中直接应对缺失数据。

- 算法支持：
- **CART（Classification and Regression Trees）、C4.5**

题目

- 1 给定以下简单的数据集，包含两个特征 X 和 Y，以及类别标签 Z，计算特征 X 和特征 Y 的信息增益，判断哪一个特征应当首先用于分裂。

X	Y	Z
0	0	A
0	1	A
1	0	B
1	1	B

解答

解答：

首先，计算数据集的熵 $H(Z)$。

数据集中有2个类别 A 和 B，各占2个样本。

$$H(Z) = -\sum_{i} P(i)\log_2 P(i) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1 \text{ bit}$$

计算基于特征 X 的信息增益：

特征 X 有两个取值：0和1，每个取值有2个样本。

- 当 $X = 0$，类别 Z 都是 A：

$$H(Z|X=0) = -\left(\frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

- 当 $X = 1$，类别 Z 都是 B：

$$H(Z|X=1) = -\left(\frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

整体条件熵：

$$H(Z|X) = \frac{2}{4}\times 0 + \frac{2}{4}\times 0 = 0$$

信息增益：

$$IG(X) = H(Z) - H(Z|X) = 1 - 0 = 1 \text{ bit}$$

解答

计算基于特征 Y 的信息增益：

特征 Y 有两个取值：0和1，各有2个样本。

- 当 $Y=0$，类别 Z 包含 A 和 B 各1个：

$$H(Z|Y=0) = -\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)=1$$

- 当 $Y=1$，类别 Z 包含 A 和 B 各1个：

$$H(Z|Y=1) = -\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)=1$$

解答

整体条件熵:

$$H(Z|Y) = \frac{2}{4} \times 1 + \frac{2}{4} \times 1 = 1$$

信息增益:

$$IG(Y) = H(Z) - H(Z|Y) = 1 - 1 = 0 \text{ bit}$$

结论:

由于 $IG(X) = 1$ 大于 $IG(Y) = 0$，应首先使用特征 X 进行分裂。

答案: 应首先使用特征 X 进行分裂。

题目

- 2.考虑以下数据集，其中包含一个连续型特征 X 和类别标签 Y，使用基尼不纯度（Gini Impurity）作为分裂标准，计算最佳分裂点。

X	Y
2	0
4	0
6	1
8	1

解答

解答：

首先，对连续特征 X 进行排序并找到可能的分裂点。可能的分裂点位于相邻样本的中间：

- 分裂点1：$(2+4)/2=3$
- 分裂点2：$(4+6)/2=5$
- 分裂点3：$(6+8)/2=7$

计算每个分裂点的基尼不纯度。

基尼不纯度公式：

$$Gini = 1 - \sum_{i}(p_i)^2$$

解答

分裂点1：3

- 左子集 $X \leq 3$: {(2,0)}

$$Gini_{left} = 1 - (1)^2 = 0$$

- 右子集 $X > 3$: {(4,0), (6,1), (8,1)}

  类别分布：0 → 1，1 → 2

$$p_0 = \frac{1}{3}, p_1 = \frac{2}{3}$$

$$Gini_{right} = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) = 1 - \left(\frac{1}{9} + \frac{4}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9} \approx 0.444$$

整体基尼：

$$Gini_{split1} = \frac{1}{4} \times 0 + \frac{3}{4} \times 0.444 = 0.333$$

解答

分裂点2：5

- 左子集 $X \leq 5$: {(2,0), (4,0)}

$$Gini_{left} = 1 - \left(\left(\frac{2}{2}\right)^2\right) = 0$$

- 右子集 $X > 5$: {(6,1), (8,1)}

$$Gini_{right} = 1 - \left(\left(\frac{2}{2}\right)^2\right) = 0$$

整体基尼:

$$Gini_{split2} = \frac{2}{4} \times 0 + \frac{2}{4} \times 0 = 0$$

解答

分裂点3：7

- 左子集 $X \leq 7$: {(2,0), (4,0), (6,1)}

  类别分布：0 → 2，1 → 1

$$p_0 = \frac{2}{3}, p_1 = \frac{1}{3}$$

$$Gini_{left} = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) = 1 - \left(\frac{4}{9} + \frac{1}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9} \approx 0.444$$

- 左子集 $X > 7$: {(8,1)}

$$Gini_{right} = 1 - (1)^2 = 0$$

整体基尼:

$$Gini_{split3} = \frac{3}{4} \times 0.444 + \frac{1}{4} \times 0 = 0.333$$

解答

比较各分裂点的基尼不纯度：

- 分裂点1：0.333
- 分裂点2：0
- 分裂点3：0.333

最佳分裂点：5

答案： 最佳分裂点为 5。

题目

- **3.考虑一个决策树的节点，有以下数据分布，计算该节点的熵，并计算如果使用某特征将数据分为两个子集后，左子集有类别分布** A:4, B:12, C:4，**右子集有类别分布** A:6, B:18, C:16。**求该特征的条件熵和信息增益。**

类别	样本数
A	10
B	30
C	20

解答

解答：

1. 计算原始节点的熵 $H(Z)$：

总样本数：$10 + 30 + 20 = 60$

类别概率：

$$P(A) = \frac{10}{60} = \frac{1}{6}, \quad P(B) = \frac{30}{60} = \frac{1}{2}, \quad P(C) = \frac{20}{60} = \frac{1}{3}$$

熵公式：

$$H(Z) = -\sum_i P(i) \log_2 P(i)$$

计算：

$$H(Z) = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{3}\log_2\frac{1}{3}\right)$$

$$= -\left(\frac{1}{6}\times(-2.585) + \frac{1}{2}\times(-1) + \frac{1}{3}\times(-1.585)\right)$$

$$= 0.4308 + 0.5 + 0.5283 \approx 1.459$$

解答

2. 计算条件熵 $H(Z|X)$:

左子集：A:4, B:12, C:4

总样本数左子集：$4+12+4=20$

类别概率：

$$P(A)=\frac{4}{20}=0.2,\quad P(B)=\frac{12}{20}=0.6,\quad P(C)=\frac{4}{20}=0.2$$

熵：

$$H(Z|\text{左})=-(0.2\log_2 0.2+0.6\log_2 0.6+0.2\log_2 0.2)$$

$$=-(0.2\times(-2.322)+0.6\times(-0.737)+0.2\times(-2.322))$$

$$=0.4644+0.4422+0.4644\approx 1.371$$

解答

右子集：A:6, B:18, C:16

总样本数右子集：$6+18+16=40$

类别概率：

$$P(A)=\frac{6}{40}=0.15,\quad P(B)=\frac{18}{40}=0.45,\quad P(C)=\frac{16}{40}=0.4$$

熵：

$$H(Z|\text{右})=-(0.15\log_2 0.15+0.45\log_2 0.45+0.4\log_2 0.4)$$
$$=-(0.15\times(-2.737)+0.45\times(-1.152)+0.4\times(-1.322))$$
$$=0.4106+0.5184+0.5288\approx 1.4578$$

解答

整体条件熵：

$$H(Z|X) = \frac{20}{60} \times 1.371 + \frac{40}{60} \times 1.4578 = \frac{1}{3} \times 1.371 + \frac{2}{3} \times 1.4578 \approx 0.457 + 0.972 = 1.429$$

3. 计算信息增益 $IG(X)$**：**

$$IG(X) = H(Z) - H(Z|X) = 1.459 - 1.429 = 0.030$$

解答

答案：

- **原始节点的熵： 约 1.459 比特**
- **条件熵 $H(Z|X)$： 约 1.429 比特**
- **信息增益： 约 0.030 比特**

题目

- 4.计算各特征的信息增益，找出最佳的根节点特征。

年龄	收入	学生	信用评级	是否购买电脑
青年	高	否	公平	否
青年	高	否	优秀	否
青年	中等	否	公平	是
青年	低	是	公平	是
中年	低	是	公平	是
老年	低	是	公平	是
老年	中等	否	优秀	否
老年	高	否	优秀	否
老年	中等	是	优秀	是

解答

Step 1：计算目标变量的熵（Entropy）

目标变量是"是否购买电脑"。目标变量有两个可能值：是（1），否（0）。先计算其熵。

- 否的样本数 = 4
- 是的样本数 = 5
- 总样本数 = 9

解答

目标变量的熵公式为：

$$H(D) = -p(是)\log_2 p(是) - p(否)\log_2 p(否)$$

其中：

$$p(是) = \frac{5}{9}, \quad p(否) = \frac{4}{9}$$

$$H(D) = -\frac{5}{9}\log_2\frac{5}{9} - \frac{4}{9}\log_2\frac{4}{9}$$

计算结果：

$$H(D) = -\frac{5}{9}\times(-0.847) - \frac{4}{9}\times(-1.322) \approx 0.998$$

所以，目标变量的熵 $H(D) \approx 0.998$。

解答

Step 2：计算每个特征的信息增益

我们将分别计算"年龄"、"收入"、"学生"和"信用评级"这四个特征的信息增益。

特征"年龄"的信息增益：

- 年龄的取值有三类：青年、中年、老年。
- 先计算每一类的条件熵，再计算整体的条件熵。

- 青年类样本：

年龄	收入	学生	信用评级	是否购买电脑
青年	高	否	公平	否
青年	高	否	优秀	否
青年	中等	否	公平	是
青年	低	是	公平	是

在"青年"类别中，是否购买电脑的分布是：否（2个），是（2个）。

所以，青年类别的熵为：

$$H(\text{青年}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

解答

- **中年类样本：**

年龄	收入	学生	信用评级	是否购买电脑
中年	低	是	公平	是

在"中年"类别中，只有1个"是"样本，熵为0：

$$H(\text{中年}) = 0$$

解答

- **老年类样本：**

年龄	收入	学生	信用评级	是否购买电脑
老年	低	是	公平	是
老年	中等	否	优秀	否
老年	高	否	优秀	否
老年	中等	是	优秀	是

在"老年"类别中，是否购买电脑的分布是：否（2个），是（2个）。

所以，老年类别的熵为：

$$H(\text{老年}) = 1$$

解答

- 总的条件熵：

现在我们可以计算"年龄"特征的条件熵。假设"年龄"取"青年"、"中年"、"老年"时的样本比例分别是：4/9, 1/9, 4/9。

所以条件熵为：

$$H(\text{年龄}|D) = \frac{4}{9} \times 1 + \frac{1}{9} \times 0 + \frac{4}{9} \times 1 = \frac{8}{9} \approx 0.889$$

信息增益：

信息增益的计算公式为：

$$IG(\text{年龄}) = H(D) - H(\text{年龄}|D)$$

$$IG(\text{年龄}) = 0.998 - 0.889 = 0.109$$

解答

Step 3：计算其它特征的信息增益

同理，可以计算“收入”、“学生”和“信用评级”这三个特征的信息增益。计算过程较为繁琐，这里省略中间的详细步骤，最终结果如下：

- 信息增益（收入）：0.0
- 信息增益（学生）：0.2
- 信息增益（信用评级）：0.0

Step 4：结论

根据计算结果，“学生”特征的信息增益最大（0.2），因此在决策树的根节点上选择“学生”特征。

End