姓名和学号：李胜志 2210180232

章节名称：CART算法

知识目标：掌握基尼指数的计算方法和使用CART算法构建决策树。

能力目标：能够使用CART算法构建决策树。

素质目标：关注国家资讯，增强民族意识，培养爱国主义精神；理解决策树算法的基本原理，培养勇为人先的精神。

知识重点：根据基尼指数选择最优特征属性作为决策树的根节点。

知识难点：基尼指数的计算。

CART算法：

CART算法(Classification and Regression Tree)分类与回归树算法，该算法通过递归地将当前样本集划分为两个子样本集，使得生成的每个非叶子节点都有两个分支，从而构建出结构简洁的二叉树模型。CART算法既可以用于分类任务，也可以用于回归任务。CART算法的基本思路是：使用基尼指数（gini index）作为度量数据集纯度的指标，其值越大，样本的不确定性越大，数据集样本纯度越低，因此，选择基尼指数最小的属性作为决策树的根节点。

基尼指数的计算公式：

假设训练集数据D根据特征A是否取某一可能值被划分为D1和D2两部分，则在特征属性A的条件下，训练集D的基尼指数为：

其中，的计算公式为：

其中，表示类别i在D1数据集中出现的概率。

CART算法的步骤

1. 计算训练数据集中每个特征属性中每个划分（属性值）对该数据集的基尼指数。
2. 在所有特征属性及其对应划分点（属性值）中，选择基尼指数最小的特征属性与对应的划分点作为最优特征属性与最优划分点。
3. 选择最优特征属性作为决策树（或子树）的根节点，选择最优划分点作为决策树的两个分支。
4. 更新数据集。根据决策树（或子树）的根节点特征属性的取值将训练数据集分配到两个分支中。
5. 重复以上步骤，直至节点的样本个数小于预定阈值或者样本集的基尼指数小于预定阈值（样本基本属于同一类）。
6. 生成CART决策树。

【例6-3】银行在办理贷款业务时，需要对客户是否会拖欠贷款进行评估，现有某银行的客户数据集（表 1），使用CART算法构造决策树，判定新用户（有房，单身年收入90k）是否会被欠贷款。

表 1银行客户数据集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 是否有房 | 婚姻状况 | 年收入/（k） | 是否拖欠贷款 |
| 1 | 是 | 单身 | 125 | 否 |
| 2 | 否 | 已婚 | 100 | 否 |
| 3 | 否 | 单身 | 70 | 否 |
| 4 | 是 | 已婚 | 120 | 是 |
| 5 | 否 | 离异 | 95 | 否 |
| 6 | 否 | 已婚 | 60 | 否 |
| 7 | 是 | 离异 | 220 | 是 |
| 8 | 否 | 单身 | 85 | 否 |
| 9 | 否 | 已婚 | 75 | 否 |
| 10 | 否 | 单身 | 90 | 是 |

CART算法构造决策树并对新样本进行预测的步骤如下。

1. “是否有房”特征属性只有两个取值，可以根据其取值“是”或“否”将数据集划分为两部分，计算每个部分的基本尼值，然后计算其基尼指数。
2. 计算取值为“是”的基尼值。属性值为“是”的3个样本中，有0个拖欠贷款（占0/3），有3个没有拖欠贷款（占3/3），则基尼值为

gini(是)=1-(0/3)^2-(3/3)^2=0

1. 计算取值为“否”的基尼值。属性值为“否”的7个样本中，有3个拖欠贷款（占3/7），有4个没有拖欠贷款（占4/7），则基尼值为
2. 计算已知“是否有房”条件下，类别属性的基尼指数。
3. “婚姻状况”特征属性有3个取值，根据每个值可将数据集进行划分，计算每个划分的基尼指数。
4. 计算“婚姻状况”为“单身”条件下，类别属性的基尼指数。属性值为“单身”的样本有4个（占总样本的4/10），其中有2个拖欠贷款（占2/4），有2个没有拖欠贷款（占2/4）；属性值为“已婚”或“离异”的样本有6个（占总样本的6/10），其中有一个拖欠贷款（占1/6），有5个没有拖欠贷款（占5/6），则其基尼指数为
5. 使用同样的方法计算“婚姻状况”为“已婚”条件下，类别属性的基尼指数。

3

1. 使用同样的方法计算“婚姻状况”为“离异”条件下，类别属性的基尼指数。

4

1. “年收入”特征属性的取值是连续性数据，这些连续性数据可将数据集进行不同的划分，计算每个划分的基尼指数。
2. 对“年收入”的属性值进行排序，然后分别计算两个相邻值的中值。本例有10个属性值，因此有9个中值，分别是65、72.5、80、87.5、92.5、97.5、110、122.5和172.5。
3. 计算“年收入”划分点为65时，类别属性的基尼指数。属性值小于或等于65的样本有1个（占总样本的1/10），其中有0个拖欠贷款（占0/1），有1个没有拖欠贷款（占1/1）；属性值大于65的样本有9个（占总样本的9/10），其中有3个拖欠贷款（占3/9）
4. 使用同样的方法计算“年收入”为其他划分点时，类别属性的基尼指数，计算结果如表6-8所示。
5. 在这些特征属性与对应的划分点中，选择基尼指数最小的特征属性与对应的划分点作为最优特征属性与最优划分点。比较发现，“已婚”作为“婚姻状况”的划分点与“97.5”作为“年收入”的划分点的基尼指数相同，选择其中一个作为最优特征属性和最优划分点即可。本例中选择“婚姻状况”作为最优特征属性，其对应属性值“已婚”作为最优划分点，得到的决策树根节点如图 1所示。

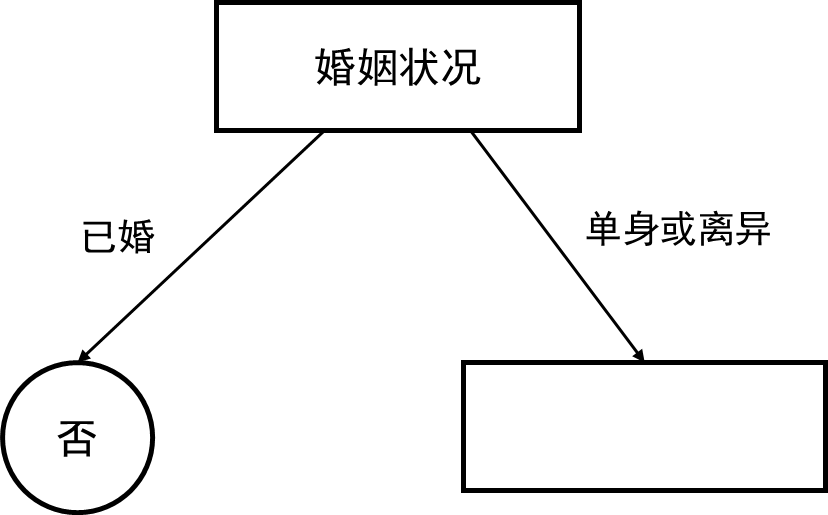


图 1决策树根节点

1. 确定右分支的根节点。

表 2右分支数据集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 是否有房 | 婚姻状况 | 年收入/（k） | 是否拖欠贷款 |
| 1 | 是 | 单身 | 125 | 否 |
| 3 | 否 | 单身 | 70 | 否 |
| 5 | 否 | 离异 | 95 | 否 |
| 7 | 是 | 离异 | 220 | 是 |
| 8 | 否 | 单身 | 85 | 否 |
| 10 | 否 | 单身 | 90 | 是 |

根据表 2的样本数据，重复步骤(1)(2)(3)(4)，确定该分支的根节点为“是否有房”。

1. 继续进行同样的运算，最后得到的决策树如图 2所示。

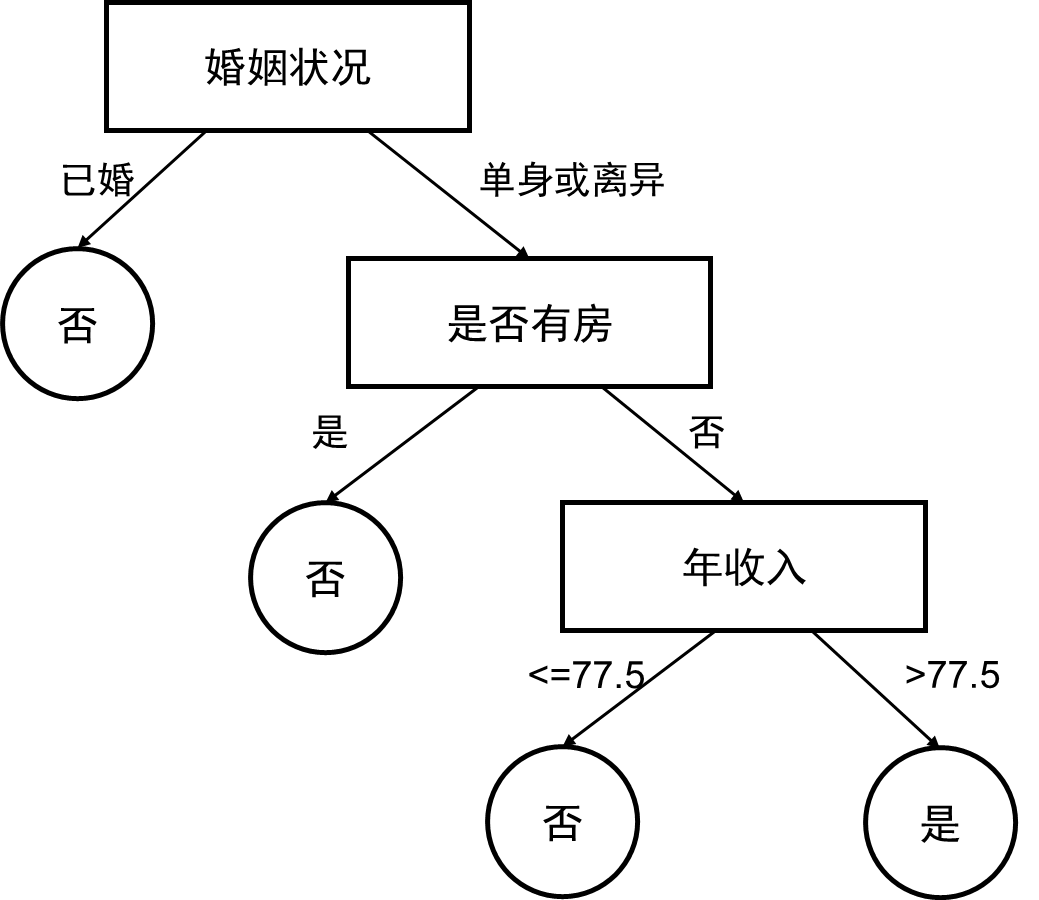


图 2完整决策树

1. 根据决策树写出决策规则。从决策树的根节点到叶节点的每条路径对应一条合取规则，故图6.7中决策树对应的决策规则为：
2. 使用决策规则对新样本进行预测。对于新用户（有房，单身，年收入90k），根据决策规则，应划分为“没有拖欠贷款”。