姓名： 学号：

|  |  |  |
| --- | --- | --- |
| **成绩** | **好好学习** | **通过** |
| 低 | 否 | 否 |
| 低 | 是 | 是 |
| 中 | 否 | 否 |
| 中 | 是 | 是 |
| 高 | 否 | 是 |
| 高 | 是 | 是 |

1. 利用以下数据集构建决策树分类器，根据学生成绩（高、中、低）及是否有好好学习（是或否），预测学生本门课能不能通过。不纯度度量使用基尼指数；利用增益及增益率分别构建决策树（log23≈1.6）。

解：① 构建根结点：

四种可能的划分方式：

划分1：

划分2：

划分3：

划分4：

无论计算增益还是增益率，最佳划分方式都为第四种。

1. 构建内部结点（父结点为上述第四种划分方式的左结点），共有三种划分方式
   1. {成绩=低，成绩=中、高}，
   2. {成绩=低、中，成绩=高}，
   3. {成绩=低，成绩=中，成绩=高}

显然，b、c划分均能达到叶结点纯度最大。因此，考虑增益，最佳决策树为：

 或 

考虑增益率，c划分生成三个叶结点，因此划分信息大，增益率小。最佳决策树为：



1. 假设一个数据集包含正负两类实例，
2. 假设正负类实例数量相等（各50个），分类器将所有实例预测为正类，计算准确率；
3. 假设正负类实例数量相等（各50个），分类器随机将80%的实例预测为正类，计算准确率；
4. 假设正负类实例数量为2:1（正类66个，负类33个），分类器将所有实例预测为正类，计算准确率；
5. 假设正负类实例数量为2:1（正类66个，负类33个），分类器随机将2/3的实例预测为正类，计算准确率。

（提示：先画出混淆矩阵再计算）

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **预测的类** | |
| + | - |
| **实际的类** | + | 50 | 0 |
| - | 50 | 0 |

解：(a)混淆矩阵：

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **预测的类** | |
| + | - |
| **实际的类** | + | 40 | 10 |
| - | 40 | 10 |

(b)混淆矩阵：

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **预测的类** | |
| + | - |
| **实际的类** | + | 66 | 0 |
| - | 33 | 0 |

(c)混淆矩阵：

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **预测的类** | |
| + | - |
| **实际的类** | + | 44 | 22 |
| - | 22 | 11 |

(d)混淆矩阵：