hw6

Wen Deng

2022/4/19

(1) 根据表 1 数据，利用信息增益比算法（C4.5 算法）生成决策树。请写出详细的计算过程并对生成的决策树作图。(目标分类变量为 **“工作表现”**)

职员	资历	教育程度	有无经验	工作表现
1	3 年以下	硕士	有	优秀
2	5 年以上	硕士	无	普通
3	3 年以下	硕士	有	优秀
4	3 年以下	本科	有	普通
5	3 年以下	硕士	无	优秀
6	5 年以上	硕士	有	优秀
7	3 年至 5 年	本科	无	普通
8	3 年至 5 年	硕士	有	优秀
9	3 年至 5 年	本科	无	普通
10	3 年以下	硕士	有	普通

表 1: 工作考核情况

Step 1

计算目标变量信息熵:

```
info_y = -5/10 * log2(5/10) * 2
info_y
```

```
## [1] 1
```

$info(y) = -\frac{5}{10}log_2\frac{5}{10} - \frac{5}{10}log_2\frac{5}{10} = 1$

计算各个因变量信息增益率并分裂:

```
info_x1 = -5/10 * (3/5*log2(3/5) + 2/5*log2(2/5)) -2/10 * (1/2*log2(1/2) + 1/2*log2(1/2)) -3/10 *
 (1/3*log2(1/3) + 2/3*log2(2/3))
info_x1
```

```
## [1] 0.960964
```

```
info_x2 = -3/10 * (-log2(3/3)) -7/10 * (5/7*log2(5/7) + 2/7*log2(2/7))
info_x2
```

```
## [1] 0.6041844
```

```
info_x3 = -6/10 * (4/6*log2(4/6) + 2/6*log2(2/6)) -4/10 * (1/4*log2(1/4) + 3/4*log2(3/4))
info_x3
```

```
## [1] 0.8754888
```

x_1代表资历，x_2代表教育程度，x_3代表有无经验。

信息熵：

$$info(x_1) = \frac{5}{10}(-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5}) + \frac{2}{10}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) + \frac{3}{10}(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}) = 0.960964$$

$$info(x_2) = \frac{3}{10}(-log_2\frac{3}{3}) + \frac{7}{10}(-\frac{5}{7}log_2\frac{5}{7} - \frac{2}{7}log_2\frac{2}{7}) = 0.6041844$$

$$info(x_3) = \frac{6}{10}(-\frac{4}{6}log_2\frac{4}{6} - \frac{2}{6}log_2\frac{2}{6}) + \frac{4}{10}(-\frac{1}{4}log_2\frac{1}{4} - \frac{3}{4}log_2\frac{3}{4}) = 0.8754888$$

信息增益：

$$gain(x_1) = 1 - 0.960964 = 0.039036$$

$$gain(x_2) = 1 - 0.6041844 = 0.3958156$$

$$gain(x_3) = 1 - 0.8754888 = 0.1245112$$

```
H_x1 = -5/10*log2(5/10) -2/10 * log2(2/10) -3/10 *log2(3/10)
H_x1
```

```
## [1] 1.485475
```

```
H_x2 = -3/10 * log2(3/10) - 7/10 * log2(7/10)
H_x2
```

```
## [1] 0.8812909
```

```
H_x3 = -6/10 * log2(6/10) - 4/10 * log2(4/10)
H_x3
```

```
## [1] 0.9709506
```

分裂信息度量：

$$H(x_1) = -\frac{5}{10}log_2\frac{5}{10} - \frac{2}{10}log_2\frac{2}{10} - \frac{3}{10}log_2\frac{3}{10} = 1.485475$$

$$H(x_2) = -\frac{3}{10}log_2\frac{3}{10} - \frac{7}{10}log_2\frac{7}{10} = 0.8812909$$

$$H(x_3) = -\frac{6}{10}log_2\frac{6}{10} - \frac{4}{10}log_2\frac{4}{10} = 0.9709506$$

信息增益率：

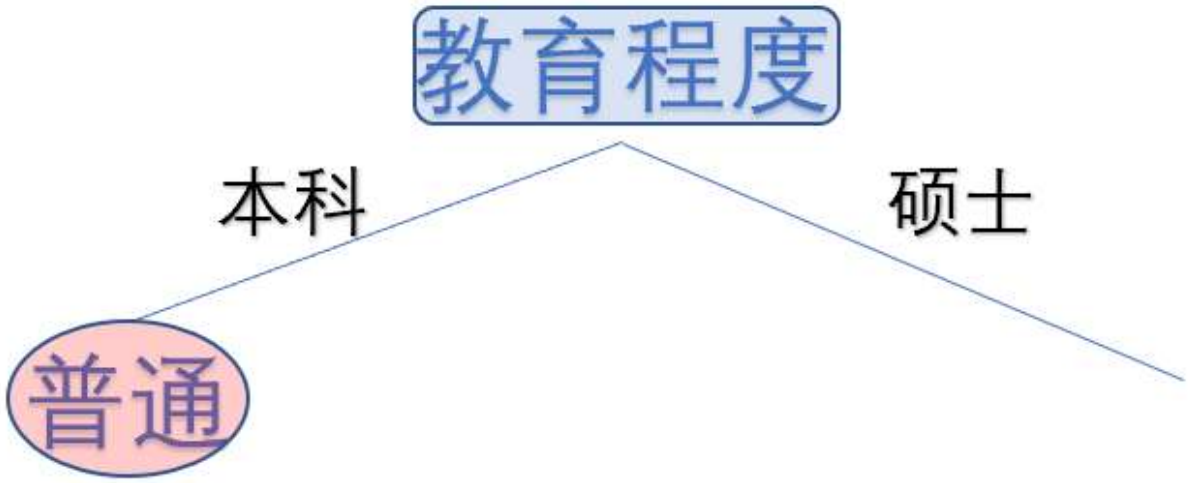
$$IGR(x_1) = \frac{gain(x1)}{H(x1)} = 0.02627846$$

$$IGR(x_2) = \frac{gain(x1)}{H(x1)} = 0.4491316$$

$$IGR(x_3) = \frac{gain(x1)}{H(x1)} = 0.1282364$$

因此选择x_2即教育程度进行分裂，

由于教育程度中的“本科”属性是纯的（即不包含其他类别，因此它直接作为一个叶节点）



接下来对“硕士”属性继续上述步骤：

Step 2

计算目标变量信息熵：

```
info_y = -5/7*log2(5/7) - 2/7*log2(2/7)
info_y
```

```
## [1] 0.8631206
```

$info(y) = -\frac{5}{7}log_2\frac{5}{7} - \frac{2}{7}log_2\frac{2}{7} = 0.8631206$

计算各个因变量信息增益率并分裂：

```
info_x1 = -4/7 * (3/4*log2(3/4) + 1/4*log2(1/4)) -2/7 * (1/2*log2(1/2) + 1/2*log2(1/2)) -1/7 * log
2(1/1)
info_x1
```

```
## [1] 0.7493018
```

```
info_x3 = -5/7 * (4/5*log2(4/5) + 1/5*log2(1/5)) -2/7 * (1/2*log2(1/2) + 1/2*log2(1/2))
info_x3
```

```
## [1] 0.8013772
```

信息熵：

$info(x_1) = \frac{4}{7}(-\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4}) + \frac{2}{7}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) + \frac{1}{7}(-log_21) = 0.7493018$

$info(x_3) = \frac{5}{7}(-\frac{4}{5}log_2\frac{4}{5} - \frac{1}{5}log_2\frac{1}{5}) + \frac{2}{7}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) = 0.8013772$

信息增益：

$gain(x_1) = 0.8631206 - 0.7493018 = 0.1138188$

$gain(x_3) = 0.8631206 - 0.8013772 = 0.0617434$

```
H_x1 = -4/7*log2(4/7) -2/7 * log2(2/7) -1/7 *log2(1/7)
H_x1
```

```
## [1] 1.378783
```

```
H_x3 = -5/7 * log2(5/7) - 2/7 * log2(2/7)
H_x3
```

```
## [1] 0.8631206
```

分裂信息度量：

$$H(x_1) = -\frac{4}{7}log_2\frac{4}{7} - \frac{2}{7}log_2\frac{2}{7} - \frac{1}{7}log_2\frac{1}{7} = 1.378783$$

$$H(x_3) = -\frac{5}{7}log_2\frac{5}{7} - \frac{2}{7}log_2\frac{2}{7} = 0.8631206$$

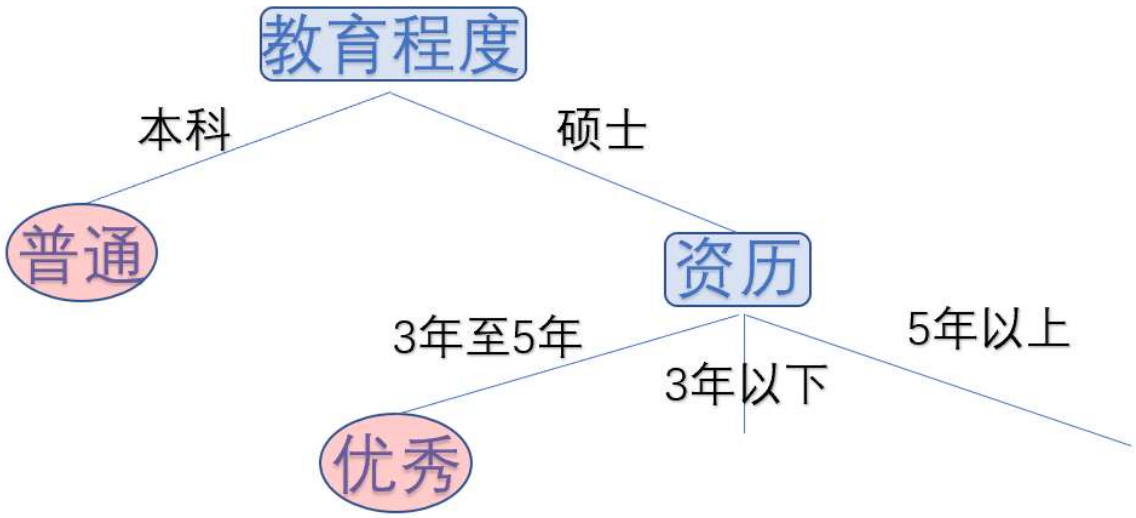
信息增益率：

$$IGR(x_1) = \frac{gain(x1)}{H(x1)} = 0.08255019$$

$$IGR(x_3) = \frac{gain(x1)}{H(x1)} = 0.07153508$$

因此选择x_1即资历进行分裂，

由于教育程度中的“3年至5年”属性是纯的（即不包含其他类别，因此它直接作为一个叶节点）



接下来分别对“3年以下”，“5年以上”属性继续上述步骤：

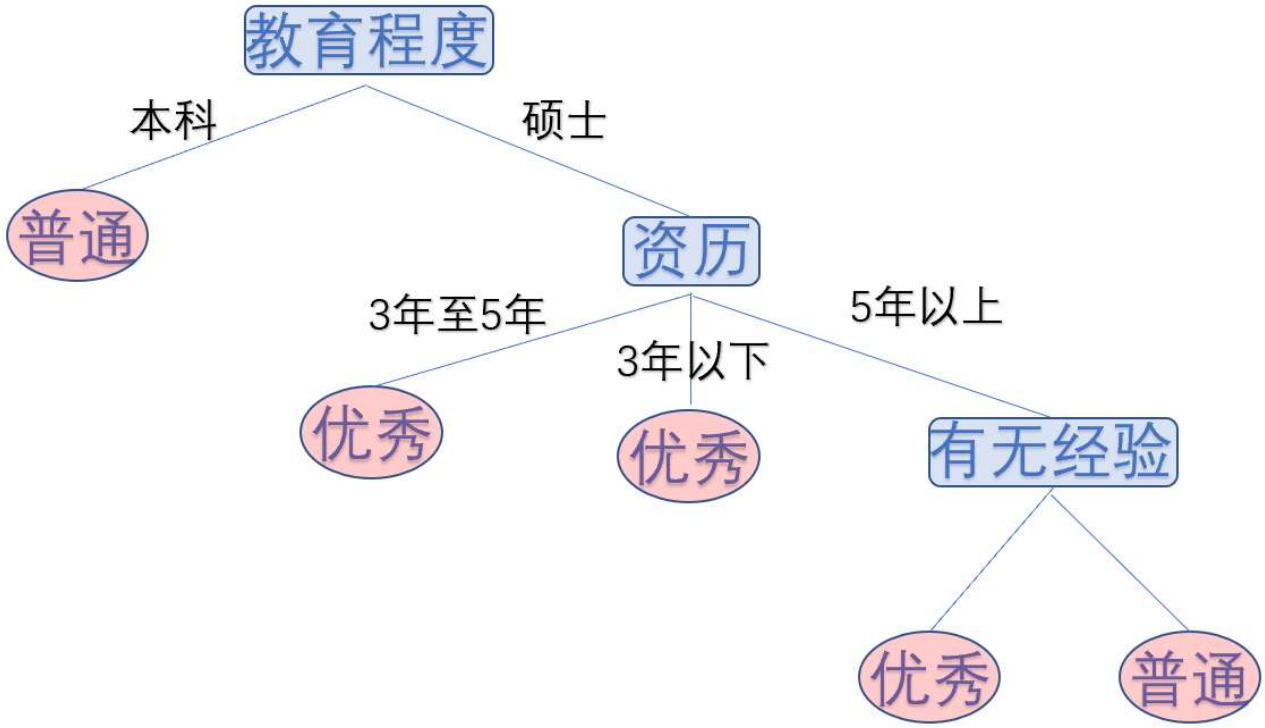
Step 3

由于按照“3年以下”，“5年以上”属性划分数据集后，只剩下了一个特征“有无经验”。

因此只能对它进行划分：

对于“5年以上”的数据集：“有”工作经验的为优秀，“无”工作经验的为普通，两个叶子节点为纯的。

对于“3年以下”的数据集：“无”工作经验的为优秀，且该个叶子节点为纯的。而“有”工作经验的有3个优秀，一个普通，由于无法再进行划分（没有更多的特征值），因此该叶节点为优秀。


教育程度
本科
硕士
普通
资历
3年至5年
3年以下
5年以上
优秀
优秀
有无经验
优秀
普通