C4.5

(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”）

(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

In [ ]:
```
from numpy import log2
```

In [ ]:
```
e_D = - (5/10 * log2(5/10) + 5/10 * log2(5/10))
e_资历 = - (
    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))
)
```

In [ ]:
```
def cal_entropy(*l):
    # 计算 l 所定义的分布的熵
    s = sum(l)
    res = 0
    for i in l:
        res -= i/s * log2(i/s)
    return res

def cal_conditional_entropy(*lists):
    # 计算经过某一类别划分后的条件熵
    ss = [sum(l) for l in lists]
    res = 0
    for l in lists:
        res += cal_entropy(*l) * sum(l) / sum(ss)
    return res

def cal_gain_ratio(entropy, *lists):
    # 计算信息增益率
    conditional_entropy = cal_conditional_entropy(*lists)
    lens = [sum(l) for l in lists]
    return (entropy - conditional_entropy) / cal_entropy(*lens)
```

分别计算三个变量的信息增益率

In [ ]:

```
# 工作表现 5/5 个 优秀/普通
e_D = cal_entropy(5,5)
# 3 年以下、5 年以上、3 年至 5 年 的员工 优秀/普通 的数量
gr_资历 = cal_gain_ratio(e_D, [3,2], [1,1], [1,2])
# 硕士、本科
gr_教育程度 = cal_gain_ratio(e_D, [5,2], [3])
# 有、经验
gr_有无经验 = cal_gain_ratio(e_D, [4,2], [1,3])
gr_资历, gr_教育程度, gr_有无经验 = [round(i, 2) for i in [gr_资历, gr_教育程度, gr_
print(gr_资历, gr_教育程度, gr_有无经验)
```

```
0.03 0.45 0.13
```

选择教育程度划分; 决策: 教育经验本科 -> 普通.

然后需要对「硕士」子表进行划分

In [ ]:

```
e_D_硕士 = cal_entropy(5,2)
# 3 年以下、5 年以上、3 年至 5 年
gr_资历_硕士 = cal_gain_ratio(e_D_硕士, [3,1], [1,1], [1])
gr_有无经验_硕士 = cal_gain_ratio(e_D_硕士, [4,1], [1,1])
gr_资历_硕士, gr_有无经验_硕士 = [round(i, 2) for i in [gr_资历_硕士, gr_有无经验_硕
print(gr_资历_硕士, gr_有无经验_硕士)
```

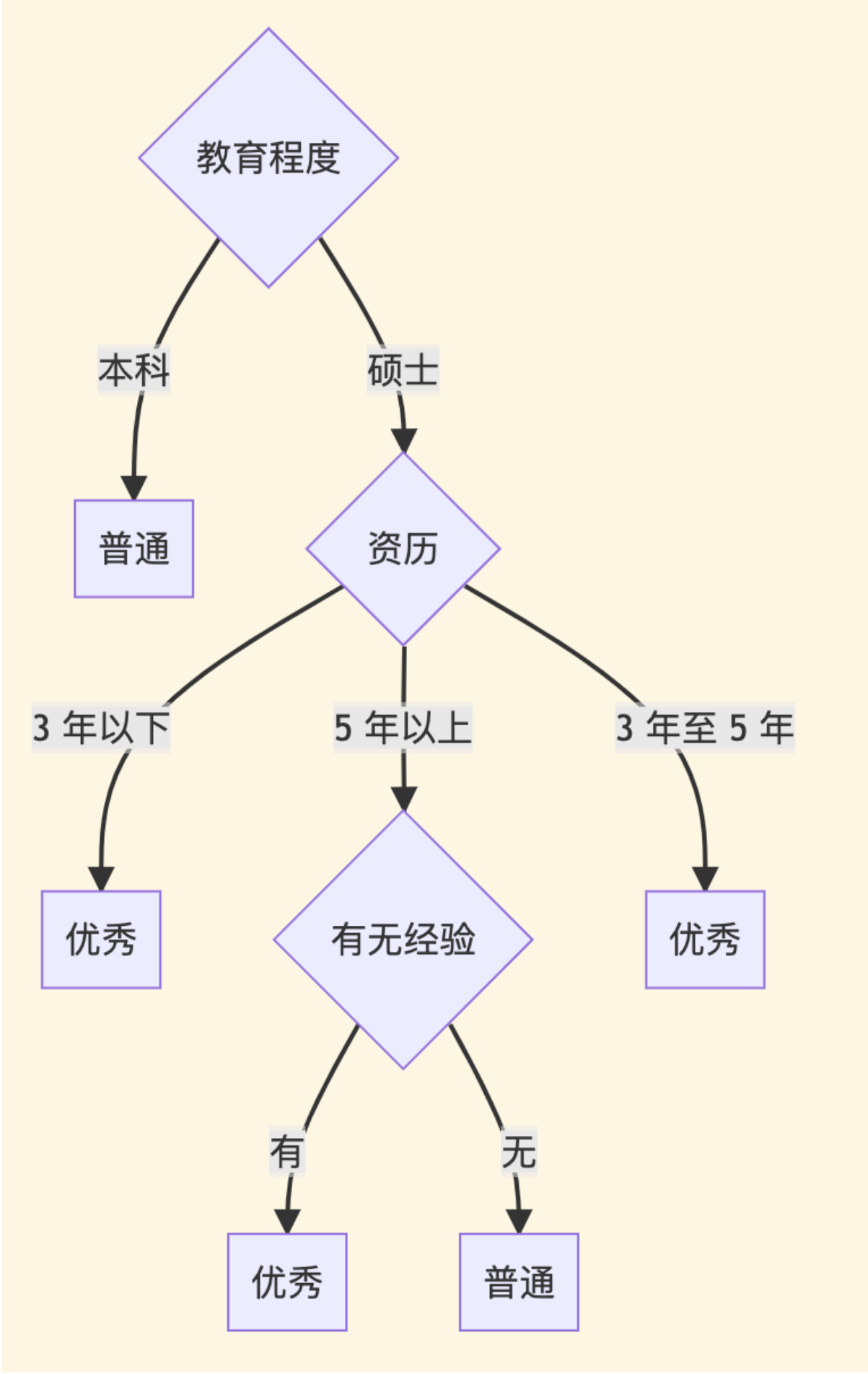
```
0.08 0.07
```

根据信息增益率, 选择资历进行划分.

决策: 资历 3 年至 5 年 -> 优秀. 对于「5 年以上」的两个员工, 再根据有无经验进行划分; 而「3 年以下」节点所包括的四个员工中, 一个经验为「无」的为优秀, 三个经验为「有」的多数为优秀, 因此统一划分为优秀.

综上, 决策树为

```
mermaid
graph TD
A{教育程度}
    A --> | 本科 | B[普通]
    A --> | 硕士 | C{资历}
    C --> | 3 年以下 | D[优秀]
    C --> | 5 年以上 | E{有无经验}
    C --> | 3 年至 5 年| F[优秀]
    E --> | 有 | G[优秀]
    E --> | 无 | H[普通]
```


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