决策树作业

说明：请使用该模板作答，格式转成 PDF，文件名为学号_姓名.pdf

学号：2017211146　　班级：2017211303　　姓名：张静雅

1. 请使用最大信息增益算法为课件 73 页的数据构建决策树，写出计算过程并画出决策树。（30 分）

使用最大信息增益算法构建决策树

输入：训练集D 和 特征A

① 训练D的经验熵 $H(D)$

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|}$$

$$= -\left(\frac{1}{17} \times \log_2 \frac{1}{17} + \cdots + \frac{1}{17} \times \log_2 \frac{1}{17}\right)$$

$$= -\log_2 \frac{1}{17}$$

$$= 4.0875$$

② A_1：男

$$H(D|男) = P(男=是)H(D|男=是) + P(男=否)H(D|男=否)$$

$$= \frac{9}{17}\left(-\frac{1}{9} \times \log\frac{1}{9}\right) \times 9 + \frac{8}{17} \times \left(-\frac{1}{8}\log\frac{1}{8}\right) \times 8$$

$$= -\frac{9}{17} \times \log\frac{1}{9} - \frac{8}{17} \times \log\frac{1}{8} = 3.09$$

$H(D|\text{先关}) = P(\text{先关}=\text{是})\,H(D|\text{先关}=\text{是}) + P(\text{先关}=\text{否})\,H(D|\text{先关}=\text{否})$

$= \frac{2}{17}\left(-\frac{1}{2}\log\frac{1}{2}\right)\times 2 + \frac{15}{17}\times\left(-\frac{1}{15}\log\frac{1}{15}\right)\times 15$

$= -\frac{2}{17}\times\log\frac{1}{2} - \frac{5}{17}\times\log\frac{1}{15} = 3.565$

$H(D|\text{运动员}) = -\frac{7}{17}\times\log_2\frac{1}{7} - \frac{10}{17}\times\log_2\frac{1}{10}$

$= 3.11$

$H(D|\text{70后}) = -\frac{4}{17}\times\log_2\frac{1}{4} - \frac{13}{17}\times\log_2\frac{1}{13}$

$= 3.297$

$H(D|\text{80后}) = -\frac{7}{17}\times\log_2\frac{1}{7} - \frac{10}{17}\times\log_2\frac{1}{10}$

$= 3.11$

$H(D|\text{离婚}) = -\frac{3}{17}\times\log_2\frac{1}{3} - \frac{14}{17}\times\log_2\frac{1}{14}$

$= 3.413$

$H(D|\text{姓秀}) = -\frac{3}{17}\times\log_2\frac{1}{3} - \frac{14}{17}\times\log_2\frac{1}{14}$

$= 3.413$

$H(D|\text{篮球}) = -\frac{2}{17}\times\log_2\frac{1}{2} - \frac{15}{17}\times\log_2\frac{1}{15}$

$= 3.565$

$H(D|\text{内地}) = -\frac{11}{17}\times\log_2\frac{1}{11} - \frac{6}{17}\times\log_2\frac{1}{6}$

$= 3.6468$

$H(D|\text{演员}) = -\frac{7}{17}\times\log\frac{1}{7} - \frac{10}{17}\times\log_2\frac{1}{10}$

$= 3.126$

$Gain(D,\text{男}) = 4.0875 - 3.09 = 0.9975$

$Gain(D,\text{先关}) = 4.0875 - 3.565 = 0.5225$

$Gain(D,\text{运动员}) = 4.0875 - 3.11 = 0.9775$

$Gain(D,\text{70后}) = 4.0875 - 3.297 = 0.7905$

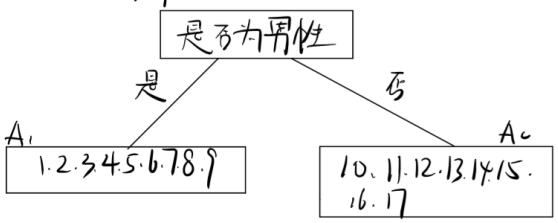
$Gain(D,\text{80后}) = 4.0875 - 3.11 = 0.9775$

$Gain(D,\text{离婚}) = 4.0875 - 3.413 = 0.6745$

$Gain(D,\text{姓秀}) = 4.0875 - 3.413 = 0.6745$

Gain(1. 监评) = 4.085 - 3.565 = 0.522
Gain(2. 内也) = 4.085 - 3.6468 = 0.4407
Gain(1. 演员) = 4.085 - 3.126 = 0.9615

Gain(1. 男)最大



对A1: H(A1)=3.169
再计算 Gain(A1.运动员) = 0.991 ✓ 其中.选秀和离婚可直接确定人物.

Gain(A1. 70后) = 0.991
Gain(A1. 未关) = 0.764
Gain(A1. 80后) = 0.673
Gain(A1. 离婚) = 0.502
Gain(A1. 选秀) = 0.502
Gain(A1. 感评) = 0.764
Gain(A1. 内也) = 0.991
Gain(A1. 演员) = 0.991

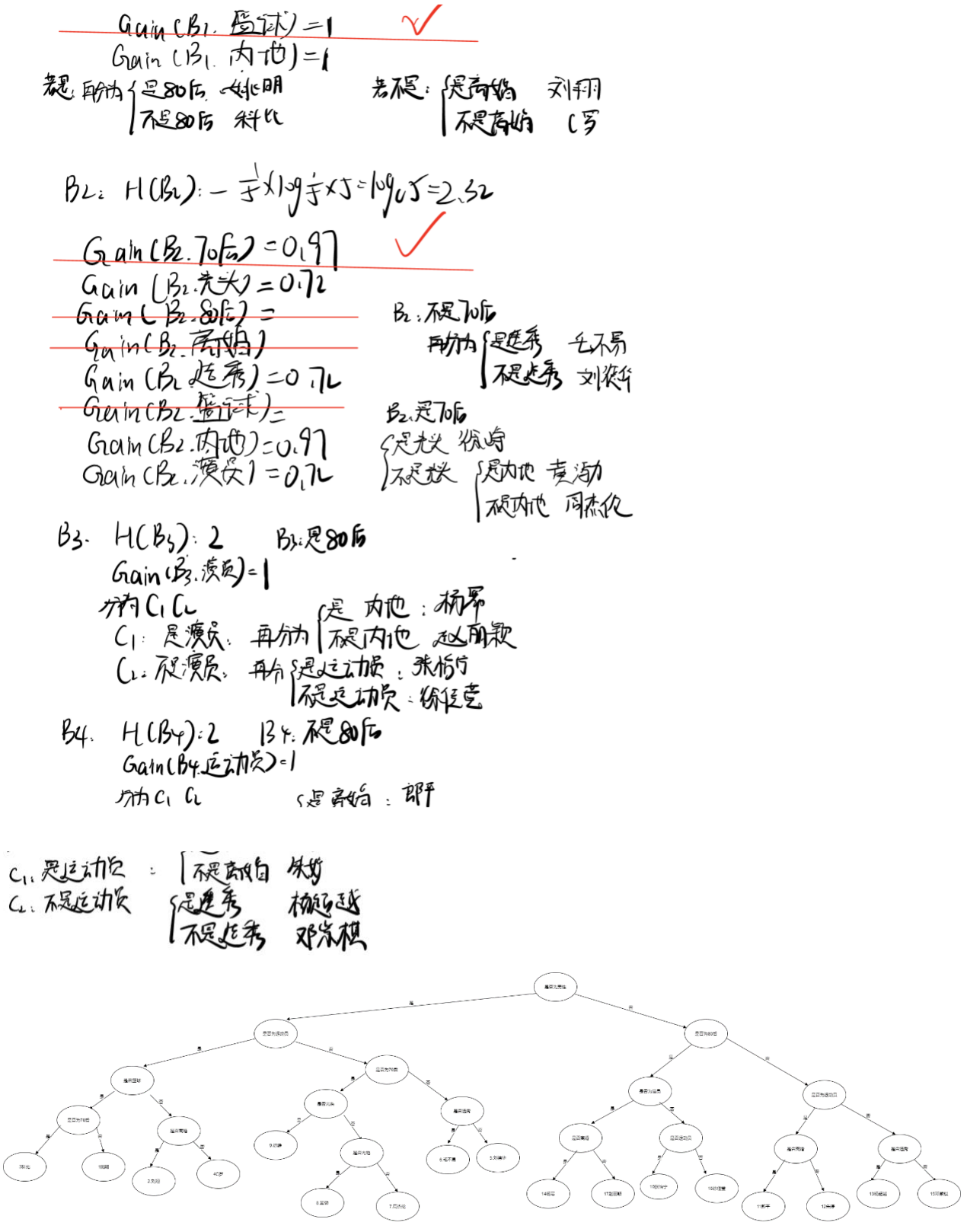
对A2: H=3 Gain(A2. 70后) = Gain(A2. 未关) = Gain(A2. 感评) = 0

Gain(A2. 运动员) = 0.854
Gain(A2. 80后) = 1 ✓

Gain(A2. 离婚) = 0.811
Gain(A2. 选秀) = 0.811
Gain(A2. 内也) = 0.811
Gain(A2. 演员) = 0.854

分成 B1 B2 B3 B4

B1: H(B1) = $-\frac{1}{4}\log_2\frac{1}{4}\times 4 = \log_2 4 = 2$
Gain(B1. 70后) = 0.312
Gain(B1. 未关) = 0.312
Gain(B1. 80后) = 0.312
Gain(B1. 离婚) = 0.312



2. 假定数据库有N个人，第n个人的先验概率γ_n，有K个问题，假定第n个人对第k个问题答案为“是”的概率为α_{nk}，请给出给定第k个问题条件下，数据集的条件熵的计算公式。（20 分）

2.

N个人 第n个人用欠账概率: γ_n 有K个问题
回答是的概率: α_{nk}

欠账概率 $P = \gamma_n$ (第n个人出现的概率)

$$H(K|X) = \sum_n p(n)\, H(K|X=n)$$

$$= -\sum_n p(n) \sum_y p(y|x) \log p(y|x)$$

$$= -\sum_n \gamma_n \sum_y P\, \alpha_{nk} \log \alpha_{nk}$$

3. 请编程实现题目 1，要求代码运行能够直接打印出决策树。代码只能包含一个文件，文件名为学号_姓名.py。编程环境要求如下：
   - Python 3.6.10
   - python standard library
   - numpy == 1.16.2
   - scipy == 1.2.1
   - pandas == 0.24.2
   - networkx == 2.4
   - graphviz==0.13.2
   - matplotlib==3.2.1

   最终生成的结果图：