

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 计算 D 的熵 $H(D)$

$$\begin{aligned} H(D) &= - \sum_{k=1}^K \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \\ &= - \left(\frac{1}{17} \times \log_2 \frac{1}{17} + \dots + \frac{1}{17} \times \log_2 \frac{1}{17} \right) \\ &= - \log_2 \frac{1}{17} \\ &= 4.0875 \end{aligned}$$

② A_1 : 男

$$\begin{aligned} H(D|A_1) &= P(\text{男}=是) H(D|\text{男}=是) + P(\text{男}=否) H(D|\text{男}=否) \\ &= \frac{9}{17} \left(-\frac{1}{9} \times \log_2 \frac{1}{9} \right) \times 9 + \frac{8}{17} \left(-\frac{1}{8} \times \log_2 \frac{1}{8} \right) \times 8 \\ &= -\frac{9}{17} \times \log_2 \frac{1}{9} - \frac{8}{17} \times \log_2 \frac{1}{8} = 3.09 \end{aligned}$$

$$\begin{aligned}
 H(D|\text{开头}) &= P(\text{开头}=\text{是}) H(D|\text{开头}=\text{是}) + P(\text{开头}=\text{否}) H(D|\text{开头}=\text{否}) \\
 &= \frac{2}{17} (-\frac{1}{2} \log_2 \frac{1}{2}) \times 2 + \frac{15}{17} \times (-\frac{1}{15} \log_2 \frac{1}{15}) \times 15 \\
 &= -\frac{2}{17} \times \log_2 \frac{1}{2} - \frac{5}{17} \times \log_2 \frac{1}{15} = 3.565
 \end{aligned}$$

$$\begin{aligned}
 H(D|\text{运动员}) &= -\frac{7}{17} \times \log_2 \frac{1}{7} - \frac{10}{17} \times \log_2 \frac{1}{10} \\
 &= 3.11
 \end{aligned}$$

$$\begin{aligned}
 H(D|\text{70后}) &= -\frac{4}{17} \times \log_2 \frac{1}{4} - \frac{13}{17} \times \log_2 \frac{1}{13} \\
 &= 3.297
 \end{aligned}$$

$$\begin{aligned}
 H(D|\text{80后}) &= -\frac{7}{17} \times \log_2 \frac{1}{7} - \frac{10}{17} \times \log_2 \frac{1}{10} \\
 &= 3.11
 \end{aligned}$$

$$\begin{aligned}
 H(D|\text{离婚}) &= -\frac{3}{17} \times \log_2 \frac{1}{3} - \frac{14}{17} \times \log_2 \frac{1}{14} \\
 &= 3.413
 \end{aligned}$$

$$\begin{aligned}
 H(D|\text{选秀}) &= -\frac{7}{17} \times \log_2 \frac{1}{7} - \frac{10}{17} \times \log_2 \frac{1}{10} \\
 &= 3.413
 \end{aligned}$$

$$\begin{aligned}
 H(D|\text{篮球}) &= -\frac{2}{17} \times \log_2 \frac{1}{2} - \frac{15}{17} \times \log_2 \frac{1}{15} \\
 &= 3.565
 \end{aligned}$$

$$\begin{aligned}
 H(D|\text{内地}) &= -\frac{11}{17} \times \log_2 \frac{1}{11} - \frac{6}{17} \times \log_2 \frac{1}{6} \\
 &= 3.6468
 \end{aligned}$$

$$\begin{aligned}
 H(D|\text{演员}) &= -\frac{7}{17} \times \log_2 \frac{1}{7} - \frac{10}{17} \times \log_2 \frac{1}{10} \\
 &= 3.126
 \end{aligned}$$

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

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

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

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

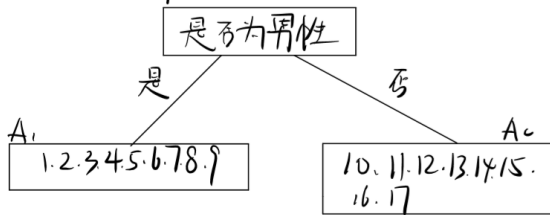
$$\text{Gain}(D, \text{80后}) = 4.0875 - 3.11 = 0.9775$$

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

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

$$\begin{aligned} \text{Gain}(C) \text{ 篮球} &= 4.0875 - 3.565 = 0.5225 \\ \text{Gain}(C) \text{ 内地} &= 4.0875 - 3.6468 = 0.4407 \\ \text{Gain}(C) \text{ 演员} &= 4.0875 - 3.126 = 0.9615 \end{aligned}$$

Gain(D) 另最大



针对 A1: $H(A1) = 3.169$
再计算 $\text{Gain}(A1 \text{ 运动员}) = 0.991$ ✓ 其中运动员和商
始可直接确定人物

$$\text{Gain}(A1 \text{ 70后}) = 0.991$$

$$\text{Gain}(A1 \text{ 丈夫}) = 0.764$$

$$\text{Gain}(A1 \text{ 80后}) = 0.653$$

$$\text{Gain}(A1 \text{ 离婚}) = 0.502$$

$$\text{Gain}(A1 \text{ 选秀}) = 0.502$$

$$\text{Gain}(A1 \text{ 篮球}) = 0.764$$

$$\text{Gain}(A1 \text{ 内地}) = 0.991$$

$$\text{Gain}(A1 \text{ 演员}) = 0.991$$

针对 A2: $H(A2) = 3$
 $\text{Gain}(A2 \text{ 70后}) = \text{Gain}(A2 \text{ 丈夫}) = \text{Gain}(A2 \text{ 篮球}) = 0$

$$\text{Gain}(A2 \text{ 运动员}) = 0.954$$

$$\text{Gain}(A2 \text{ 80后}) = 1 \quad \checkmark$$

$$\text{Gain}(A2 \text{ 离婚}) = 0.811$$

$$\text{Gain}(A2 \text{ 选秀}) = 0.811$$

$$\text{Gain}(A2 \text{ 内地}) = 0.811$$

$$\text{Gain}(A2 \text{ 演员}) = 0.954$$

分成 B1 B2 B3 B4

$$B1: H(B1) = -\frac{1}{4} \times \log_2 \frac{1}{4} \times 4 = 1.0924 = 2$$

$$\text{Gain}(B1 \text{ 70后}) = 0.312$$

$$\text{Gain}(B1 \text{ 丈夫}) = 0.312$$

$$\text{Gain}(B1 \text{ 80后}) = 0.312$$

$$\text{Gain}(B1 \text{ 离婚}) = 0.312$$



親再分爲 $\begin{cases} \text{是80斤} & \text{姓明} \\ \text{不是80斤} & \text{姓比} \end{cases}$

老根: 是青姑 刘翔
不是青姑 C罗

$$B_{L2}: H(B_{L2}): -\frac{1}{5} \times \log \frac{1}{5} \times 5 = \log 5 = 2.32$$



B₂: 不是正合

再力为退避秀 乞不易

不是这样 刘荣华

長 70 分

材料 複合

不是找 { 是內地 黃渤
不是內地 周杰倫

B3. $H(B_3): 2$ B3: 是 80 后

$$\text{Gain}(B_3, \text{读页}) = 1$$

分为 C_1, C_2

C_1 是漢人，再分為：

- 是內地：楊鼎
- 不是內地：趙爾毅

C1: 假演员: 再分 { 是运动员: 张怡宁
不是运动员: 徐亚克

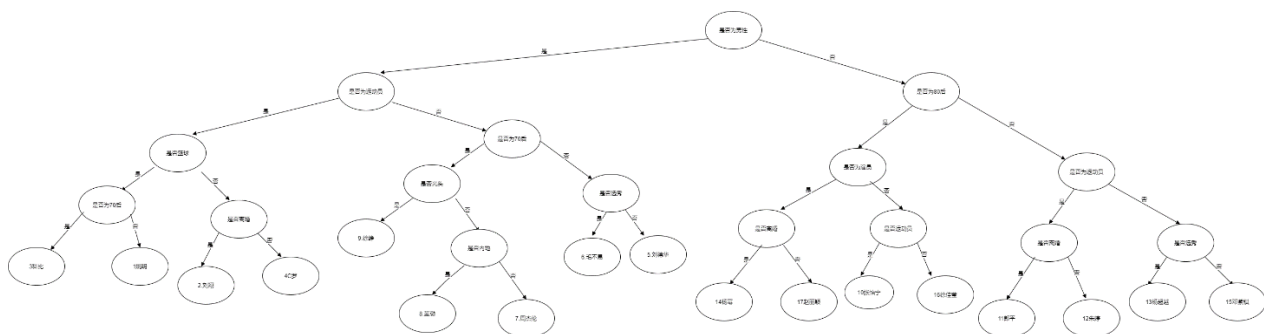
B4. $H(B_4) = 2$ B4: 不是80分

$$\text{Gain}(\text{By. 运动员}) = 1$$

方为 C_1, C_2

是 齋 始：郎平

C₁: 是运动学 : 不是动力学 朱芳
C₂: 不是运动学 : 是力学 杨超越
不是力学 邓紫棋



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个人回答问题的概率: Y_n 有K个问题
回答是的概率: α_{nk}

失配概率 $p = Y_n$ (第n个人出现的概率).

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

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

$$= - \sum_n Y_n \sum_y p \alpha_{nk} \log \alpha_{nk}$$

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:

