决策树作业

学号：2017211594　　班级：　2017211301　　姓名：袁子麒

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）

答：我们将数据集记作 D ，将属性集合记为 $A=\{A_1,\cdots,A_{10}\}$

其中 A_1-A_{10} 依次表示数据集中从左到右的属性，（即 A_1 表示 是否为男性，A_5 表示是否为 80 后， A_{10} 表示 是否为演员）

首先，进行根结点划分属性选择：

我们计算各个属性在全体数据集上的条件熵：利用公式（1）

$$H(D|A_i)=P(A_i=\text{是})\cdot H(D|A_i=\text{是})+P(A_i=\text{否})\cdot H(D|A_i=\text{否}) \quad (1)$$

依次求得 $H(D|\textit{男})\approx 3.09$，$H(D|\textit{运动员})\approx 3.11$，$H(D|70\ \textit{后})\approx 3.30$，

$H(D|\textit{光头})\approx 3.56$，$H(D|80\ \textit{后})\approx 3.11$，$H(D|\textit{离婚})\approx 3.42$，$H(D|\textit{选秀})\approx 3.42$，

$H(D|\textit{篮球})\approx 3.56$，$H(D|\textit{内地})\approx 3.15$，$H(D|\textit{演员})\approx 3.11$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性。故根结点的划分属性为 **男**

然后进入第二层节点划分属性选择

我们记数据集中 $(\textit{男}=\text{是})$ 的子集为 D_1，记数据集中 $(\textit{男}=\text{否})$ 的子集为 D_2。

仍然利用公式（1），分别计算除 男 属性以外的属性在D_1上的条件熵：

依次求得 $H(D_1|\textit{运动员})\approx 2.18$，$H(D_1|70\ \textit{后})\approx 2.18$，$H(D_1|\textit{光头})\approx 2.41$，

$H(D_1|80\ \textit{后})\approx 2.25$，$H(D_1|\textit{离婚})\approx 2.67$，$H(D_1|\textit{选秀})\approx 2.67$，$H(D_1|\textit{篮球})\approx$

2.41，$H(D_1|\textit{内地})\approx 2.18$，$H(D_1|\textit{演员})\approx 2.18$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **运动员**

仍然利用公式（1），分别计算除 男 属性以外的属性在D_2上的条件熵：

依次求得 $H(D_2|\textit{运动员})\approx 2.05$，$H(D_2|70\ \textit{后})\approx 3$，$H(D_2|\textit{光头})\approx 3$，

$H(D_2|80\ \textit{后})\approx 2$，$H(D_2|\textit{离婚})\approx 2.19$，$H(D_2|\textit{选秀})\approx 2.19$，$H(D_2|\textit{篮球})\approx 3$，

$H(D_2|\textit{内地})\approx 2.19$，$H(D_2|\textit{演员})\approx 2.05$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同

的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **80 后**

然后进入第三层节点划分属性选择

我们记数据集中 $男=是\&运动员=是$ 的子集为 D_3，记数据集中 $男=是\&运动员=否$ 的子集为 D_4。

仍然利用公式（1），分别计算除 男 运动员 属性以外的属性在D_3上的条件熵：

依次求得 $H(D_3|70\ 后)\approx 1.19$，$H(D_3|光头)\approx 1.19$，$H(D_3|80\ 后)\approx 1.19$，$H(D_3|离婚)\approx 1.19$，$H(D_3|选秀)\approx 2$，$H(D_3|篮球)\approx 1$，$H(D_3|内地)\approx 1$，$H(D_3|演员)\approx 2$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **篮球**

仍然利用公式（1），分别计算除 男 运动员 属性以外的属性在D_4上的条件熵：

依次求得 $H(D_4|70\ 后)\approx 1.35$，$H(D_4|光头)\approx 1.6$，$H(D_4|80\ 后)\approx 2.32$，$H(D_4|离婚)\approx 2.32, H(D_4|选秀)\approx 1.6, H(D_4|篮球)\approx 2.32, H(D_4|内地)\approx 1.35$，$H(D_4|演员)\approx 1.6$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **70 后**

我们记数据集中 $男=否\&80\ 后=是$ 的子集为 D_5，记数据集中 $男=否\&80\ 后=否$ 的子集为 D_6。

仍然利用公式（1），分别计算除 男 80 后 属性以外的属性在D_5上的条件熵：

依次求得 $H(D_5|运动员)\approx 1.19$，$H(D_5|70\ 后)\approx 2$，$H(D_5|光头)\approx 2$，$H(D_5|离婚)\approx 1.19, H(D_5|选秀)\approx 1.19, H(D_5|篮球)\approx 2, H(D_5|内地)\approx 1.19$，$H(D_5|演员)\approx 1$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **演员**

仍然利用公式（1），分别计算除 男 80 后 属性以外的属性在D_6上的条件熵：

依次求得$H(D_6|运动员)\approx 1, H(D_6|70\ 后)\approx 2, H(D_6|光头)\approx 2, H(D_6|离婚)\approx 1.19, H(D_6|选秀)\approx 1.19, H(D_6|篮球)\approx 2, H(D_6|内地)\approx 1.19, H(D_6|演员)\approx$

1.19

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **运动员**

然后进入第四层节点划分属性选择

我们记数据集中 $男=是\,\&\,运动员=是\,\&篮球=是$ 的子集为 D_7，记数据集中 $男=是\,\&\,运动员=是\,\&篮球=否$ 的子集为 D_8。

仍然利用公式(1)，分别计算除 男 运动员 篮球 属性以外的属性在D_7上的条件熵：

依次求得$H(D_7|70\,后)\approx 0$，$H(D_7|光头)\approx 0$，$H(D_7|80\,后)\approx 0$，$H(D_7|离婚)\approx 1$，$H(D_7|选秀)\approx 1$，$H(D_7|内地)\approx 0$，$H(D_7|演员)\approx 1$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **70 后**

在 70 后节点下，根据属性 男=是 运动员=是 篮球=是 70 后=是 判断人物为科比， 根据属性 男=是 运动员=是 篮球=是 70 后=否，判断人物为姚明，结束

仍然利用公式(1)，分别计算除 男 运动员 篮球 属性以外的属性在D_8上的条件熵：

依次求得$H(D_8|70\,后)\approx 1$，$H(D_8|光头)\approx 1$，$H(D_8|80\,后)\approx 1$，$H(D_8|离婚)\approx 0$，$H(D_8|选秀)\approx 1$，$H(D_8|内地)\approx 0$，$H(D_8|演员)\approx 1$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **离婚**

在 离婚 节点下，根据属性 男=是 运动员=是 篮球=是 离婚=是 判断人物为刘翔， 根据属性 男=是 运动员=是 篮球=是 离婚=否，判断人物为 C 罗，结束

我们记数据集中 $男=是\,\&\,运动员=否\,\&\,70\,后=是$ 的子集为 D_9，记数据集中 $男=是\,\&\,运动员=否\,\&\,70\,后=否$ 的子集为 D_{10}。

仍然利用公式(1)，分别计算除 男 运动员 70 后 属性以外的属性在D_9上的条件熵：

依次求得 $H(D_9|光头)\approx 0.66$，$H(D_9|80\,后)\approx 1.58$，$H(D_9|离婚)\approx 1.58$，$H(D_9|选秀)\approx 1.58$，$H(D_9|篮球)\approx 1.58$，$H(D_9|内地)\approx 0.66$，$H(D_9|演员)\approx 1.58$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **光头**

在 光头 节点下，根据属性 男=是 运动员=否 70 后=是 光头=是 判断人物为徐峥

仍然利用公式（1），分别计算除 男 运动员 70 后 属性以外的属性在D_{10}上的条件熵：

依次求得$H(D_{10}|光头) \approx 1$，$H(D_{10}|80\ 后) \approx 1$，$H(D_{10}|离婚) \approx 1$，$H(D_{10}|选秀) \approx 0$，$H(D_{10}|篮球) \approx 1$，$H(D_{10}|内地) \approx 0$，$H(D_{10}|演员) \approx 0$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **选秀**

在 选秀 节点下，根据属性 男=是 运动员=否 70 后=否 选秀=是 判断人物为毛不易， 根据属性 男=是 运动员=否 70 后=否 选秀=否，判断人物为刘德华，结束

我们记数据集中 $男=否\&80\ 后=是\&演员=是$ 的子集为D_{11}，记数据集中 $男=否\&80\ 后=是\&演员=否$ 的子集为D_{12}。

仍然利用公式（1），分别计算除 男 80 后 演员 属性以外的属性在D_{11}上的条件熵：

依次求得 $H(D_{11}|运动员) \approx 1$ ， $H(D_{11}|70\ 后) \approx 1$ ， $H(D_{11}|光头) \approx 1$ ，$H(D_{11}|离婚) \approx 0$，$H(D_{11}|选秀) \approx 1$，$H(D_{11}|篮球) \approx 1$， $H(D_{11}|内地) \approx 1$，

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **离婚**

在 离婚 节点下，根据属性 男=否 80 后=是 演员=是 离婚=是 判断人物为杨幂， 根据属性 男=否 80 后=是 演员=是 离婚=否，判断人物为赵丽颖，结束

仍然利用公式（1），分别计算除 男 80 后 演员 属性以外的属性在D_{12}上的条件熵：

依次求得 $H(D_{12}|运动员) \approx 0$ ， $H(D_{12}|70\ 后) \approx 1$ ， $H(D_{12}|光头) \approx 1$ ，$H(D_{12}|离婚) \approx 1$，$H(D_{12}|选秀) \approx 0$，$H(D_{12}|篮球) \approx 1$， $H(D_{12}|内地) \approx 0$，

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **运动员**

在 运动员 节点下，根据属性 男=否 80 后=是 演员=否 运动员=是 判断人物为张怡宁， 根据属性 男=否 80 后=是 演员=否 运动员=否，判断人物为徐佳莹，

结束

我们记数据集中 $男=否\&80后=否\&运动员=是$ 的子集为 D_{13}，记数据集中 $男=否\&80后=否\&运动员=否$ 的子集为 D_{14}。

仍然利用公式（1），分别计算除 男 80后 运动员 属性以外的属性在D_{13}上的条件熵：

依次求得$H(D_{13}|70后)\approx 1, H(D_{13}|光头)\approx 1, H(D_{13}|离婚)\approx 0, H(D_{13}|选秀)\approx 1, H(D_{13}|篮球)\approx 1, H(D_{13}|内地)\approx 1, H(D_{13}|演员)\approx 1$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **离婚**

在 离婚 节点下，根据属性 男=否 80后=是 运动员=是 离婚=是 判断人物为郎平， 根据属性 男=否 80后=是 运动员=是 离婚=否，判断人物为朱婷，结束

仍然利用公式（1），分别计算除 男 80后 运动员 属性以外的属性在D_{14}上的条件熵：

依次求得$H(D_{14}|70后)\approx 1, H(D_{14}|光头)\approx 1, H(D_{14}|离婚)\approx 1, H(D_{14}|选秀)\approx 0, H(D_{14}|篮球)\approx 1, H(D_{14}|内地)\approx 0, H(D_{14}|演员)\approx 0$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **选秀**

在 选秀 节点下，根据属性 男=否 80后=是 运动员=是 选秀=是 判断人物为杨超越， 根据属性 男=否 80后=是 运动员=否 选秀=否，判断人物为邓紫棋，结束

然后进入第五层节点划分属性选择：（男=是 运动员=否 70后=是 光头=否）

我们记数据集中 $男=是\&运动员=否\&70后=是\ 光头=是$ 的子集为 D_{15}

利用公式（1），分别计算除 男 运动员 70后 光头 属性以外的属性在D_{15}上的条件熵：

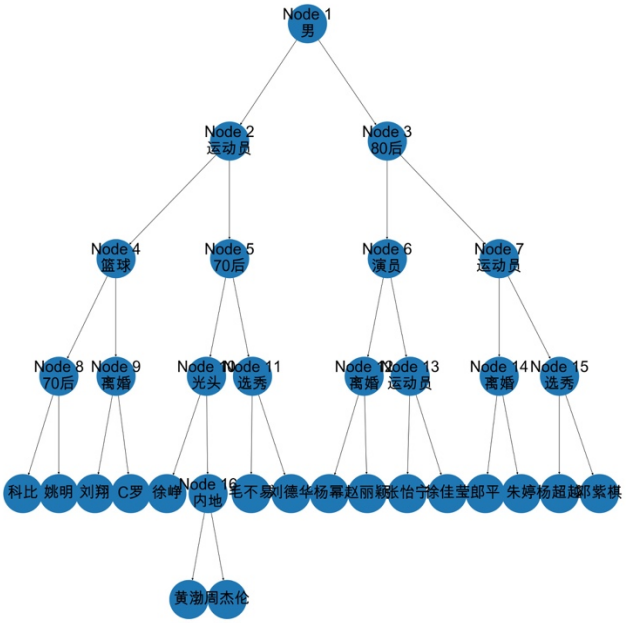
依次求得$H(D_{11}|80后)\approx 1, H(D_{11}|离婚)\approx 1, H(D_{11}|选秀)\approx 1, H(D_{11}|篮球)\approx 1, H(D_{11}|内地)\approx 0, H(D_{11}|演员)\approx 1$

我们选择信息增益最大的属性即条件熵最小的属性作为划分属性(如果存在相同的最小值，选择属性下标最小的属性)。故根结点的划分属性为 **内地**

在 内地 节点下，根据属性 男=是 运动员=否 70后=是 光头=否 内地=是 判断

人物为黄渤， 根据属性 男=是 运动员=否 70 后=是 光头=否 内地=否，判断人物为周杰伦，结束

决策树绘制如下：



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）

解题思路：首先这是一个将数据库中每一个体分成一类的问题。首先分析不考虑先验概率（对应与$\forall n \in \{1,2,\cdots,N\}, \gamma_n = \frac{1}{N}$）并假设每个人对每个问题答案为“是”的概率取值属于$\{0,1\}$的情况。

对于答案只有“是”或“否”的问题k，我们只要确定对于问题k有多少个成员答案为“是”，多少个成员答案为“否”即可确定给定第k个问题条件下数据集的条件熵。我们不妨设对于问题k，回答“是”的人数为m ，回答“否”的人数为n $(n = N - m)$ 。则该问题在数据集上的条件熵为

$$H(D|k) = \frac{n}{m+n}\log_2 n + \frac{m}{m+n}\log_2 m。$$

基于此公式，在有先验概率（bias）和回答概率属于$[0,1]$的问题中，我们只需要给出回答为“是”和“否”的人数的广义定义即可。

首先根据先验概率重新分配个体i所占权重。基于以下两点：

1. 保持总人数不变 $\sum_{i=1}^{N} s_i = N$
2. 保持每个人权重$s_i \propto \gamma_i$

故此有$s_i = N \times \gamma_i$。

然后，根据第n个人对第k个问题答案为“是”的概率为α_{nk}，将第n个人的权重s_n分配到“是”，“否”两类。使用“贡献法”，第n个人对“是”类的贡献为$s_n \times a_{nk}$，对“否”类贡献为$s_n \times (1 - a_{nk})$.

根据此重新计算“是”“否”两类的虚拟总人数$m', n'(m' + n' = N)$

$$m' = \sum_{i=1}^{N} s_i \times a_{ik} = N \sum_{i=1}^{N} a_{ik} \times \gamma_i$$

$$n' = \sum_{i=1}^{N} s_i \times (1 - a_{ik}) = N \sum_{i=1}^{N} \gamma_i \times (1 - a_{ik})$$

将m', n'带入公式$H(D|k) = \frac{n'}{m'+n'} \log_2 n' + \frac{m'}{m'+n'} \log_2 m'$即可。

以下两道题目，二选一即可。（50 分）

3. 请选择一个你认为有意义或有趣味的领域，收集一个可以用作 20 问读心游戏的数据集。角色数不小于 100 个，问题数不小于 20 个。请写出你选择该领域的理由和数据集的收集方法。（数据集列出角色和问题，角色对问题的答案仅选择 20 个角色和 10 个问题即可）

4. 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
   - python standard library
   - numpy == 1.16.2
   - scipy == 1.2.1
   - pandas == 0.24.2