《计算机应用数学第一次作业》

截止时间：2024.04.17

计算题（80 分）

1、 （20 分）求 E(X)， Var(X)

(1) X服从[a,b]均匀分布：

(2) $X=x_1+x_2+\cdots x_n$，$x_i\in\{0,1\}$， 相互独立。

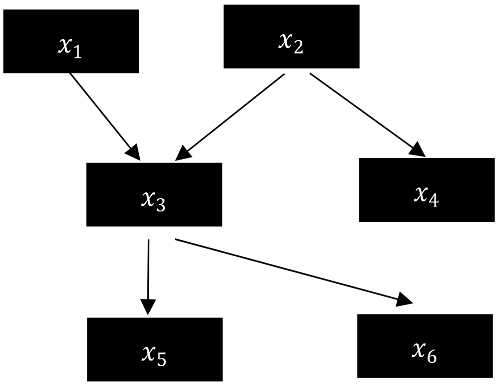
2. （15 分）布隆过滤器(Bloom Filter) 是一种空间效率高、查询效率快的数据结构，用于判断一个元素是否属于一个集合。它实质上是一个长度为 m 的 01 数组和 k 个不同的哈希函数构成。在添加元素时，将元素经过 k 个哈希函数得到哈希值，并将数组上这些哈希值位置标记为 1。在查询元素时，将元素经过同样的 k 个哈希函数得到哈希值，若数组上这些哈希值位置都为 1，则说明元素可能在集合中，否则一定不在集合中。布隆过滤器可能会出现误判，但不会出现漏判。假设每个哈希函数都是随机函数，请计算在插入 n 个元素后，布隆过滤器出现误判的概率，即一个不在集合中的元素被判定为在集合中。

3、（15 分）我们有两种硬币：一种是公平的硬币，即抛一次正反的概率均为 1/2；另一种是产生正面朝上的概率为 2/3 的硬币。从两枚硬币中挑出一枚，将这枚硬币掷 n 次。抛多少次足以让我们有 99% 的把握选择了哪种硬币？请写出计算过程。

4、（10 分）考虑转移概率矩阵，六个状态分别为{0,1,2,3,4,5}，判断给定马尔可夫链是否存在稳定状态？

$$\boldsymbol{P}=\begin{pmatrix} 1/4 & 1/2 & 0 & 0 & 1/4 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 1/4 & 0 & 1/2 & 0 & 1/4 \\ 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/3 & 0 & 0 & 0 & 1/3 & 1/3 \\ 1/5 & 1/5 & 2/5 & 1/5 & 0 & 0 \end{pmatrix}$$

5、（20 分）$x_1, \cdots, x_6$属性满足以下网络图关系，给出对应的因子图以及联合分布概率公式以及x_3的边缘概率公式（即$P(x_1; x_2; x_3; x_4; x_5; x_6)$ ，$P(x_3)$），对比直接计算$P(x_3)$和使用 Belief Propagation 算法（或称作 Sum-Product Message Passing）的计算代价之间的差异（即比较乘法加法次数）。



编程题 （20 分）

说明：建议使用开源工具包，例如 scikit-learn 中有朴素贝叶斯函数实现
朴素贝叶斯分类器（Naive Bayes Classifier）
数据集：Bayesian_Dataset_train.csv, Bayesian_Dataset_test.csv。
数据描述：列名分别为“年纪、工作性质、家庭收入、学位、工作类型、婚姻状况、族裔、性别、工作地点”，最后一列是标签，即收入是否大于 50k 每年。
任务描述：使用朴素贝叶斯（Naïve Bayesian）预测一个人的收入是否高于 50K 每年。
要求输出：
采用不同评估方式，至少包含两种（如交叉验证和留一法等），给出结果统计，包括 Accuracy、precision、recall、F1 score、ROC；
Optional：探索不同参数对结果的影响。

$$= \frac{(b-a)^2}{12}$$

(2) 二项分布【离散型】：

$X \sim B(n,p)$ (n重伯努利试验)

$E(x_i) = 0 \times q + 1 \times p = p$

$E(x_i^2) = 0^2 \times q + 1^2 \times p = p$

$Var(x_i) = E(x_i^2) - [E(x_i)]^2 = p - p^2 = p(1-p)$

由于$x_1, x_2, \dots, x_i$相互独立：

$E(X) = \sum_{i=1}^{n} E(x_i) = np$

$Var(X) = \sum_{i=1}^{n} Var(x_i) = np(1-p)$

由于题目没有说明$x_1, x_2, \dots, x_i$的概率，可以写成以下形式：

$$E(X) = \sum_{i=1}^{n} E(x_i) = \sum_{i=1}^{n} p_i$$

$$Var(X) = \sum_{i=1}^{n} Var(x_i) = \sum_{i=1}^{n} p_i(1-p_i)$$

有些同学默认$p = 0.5$

插入n个元素后，不在集合中的元素被判定为在集合中的概率，即k个哈希函数指向的bit均被置为1的概率

参数：

- 数组长度 m
- 哈希函数的个数 k
- 插入元素的个数 n

对于一个插入元素，经过一个哈希函数后，在某一节点被置为1的概率为：$\frac{1}{m}$（哈希函数相当于在m个bit中挑选一个）

对于一个插入元素，经过一个哈希函数后，在某一节点没有被置为1的概率为$1-\frac{1}{m}$

对于一个插入元素，经过k个哈希函数后，在某一节点没有被置为1的概率为$(1-\frac{1}{m})^k$（即每一次哈希都没有选中这个节点）

对于n个插入元素，经过k个哈希函数后，在某一节点没有被置为1的概率为$(1-\frac{1}{m})^{nk}$（每个元素、每一次哈希都没有选中这个节点）

对于n个插入元素，经过k个哈希函数后，在某一节点被置为1的概率为$1-(1-\frac{1}{m})^{nk}$

出现误判的概率：新插入的元素经过k个哈希函数后，这k个位置都被置为1，概率为$(1-(1-\frac{1}{m})^{nk})^k$（选择k次，每次都选中已经置为1的节点）

$$
\begin{aligned}
P_{error} &= [1-(1-\frac{1}{m})^{kn}]^k \\
&= [1-(1-\frac{1}{m})^{-m(-\frac{kn}{m})}]^k \\
&\approx (1-e^{-\frac{kn}{m}})^k
\end{aligned}
$$

（当m足够大时，根据极限：$\lim_{m\to\infty}(1-\frac{1}{m})^m=\frac{1}{e}$）

不可约、非周期且正常返的马尔可夫链，有唯一平稳分布存在，并且转移概率的极限分布是马尔可夫链的平稳分布

- 不可约：每个状态都可达
- 非周期：在当前状态转移到自己的可能性（任一状态返回到自身的间隔步数没有固定的周期）
- 正常返：每个状态出发后总能够经过有限步回到自己

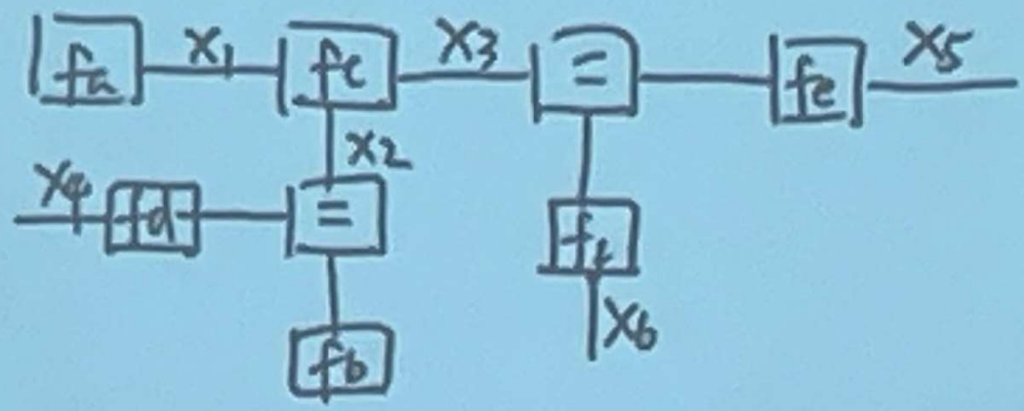
因此存在平稳分布，可以求出稳定状态：

$$\pi P = \pi$$

$$\sum_i \pi_i = 1$$

解得：$\pi = [\frac{110}{1123}, \frac{423}{2246}, \frac{208}{1123}, \frac{278}{1123}, \frac{291}{2246}, \frac{170}{1123}] = [0.098, 0.188, 0.185, 0.248, 0.130, 0.151]$.

编程求解/方程式计算均可



- 联合分布概率公式：

$$P(x_1,x_2,x_3,x_4,x_5,x_6)=P(x_5|x_3)*P(x_6|x_3)*P(x_3|x_1,x_2)*P(x_4|x_2)*P(x_1)*P(x_2)$$

- x_3边缘概率公式：

$$P(x_3)=\tilde{f}_3(x_3)=\sum_{x_1}\sum_{x_2}\sum_{x_4}\sum_{x_5}\sum_{x_6}P(x_5|x_3)*P(x_6|x_3)*P(x_3|x_1,x_2)*P(x_4|x_2)*P(x_1)*P(x_2)$$

直接计算：假设每个x_i有n种取值，则直接计算共有n^5次取值，每次取值需要1次加法和5次乘法，因此乘法和加法都是$O(n^5)$

使用BP算法：

$$\tilde{f}_3(x_3)=m_{\rightarrow}x_3\times m_{\leftarrow}x_3$$
$$m_{\rightarrow}x_1=P(x_1)$$
$$m_{\rightarrow}x_2=\sum_{x_4}P(x_2)P(x_4|x_2)$$
$$m_{\rightarrow}x_3=\sum_{x_1}\sum_{x_2}m_{\rightarrow}x_1P(x_3|x_1,x_2)m_{\rightarrow}x_2=\sum_{x_1}\sum_{x_2}P(x_1)P(x_3|x_1,x_2)P(x_2)(\sum_{x_4}P(x_4|x_2))$$
$$m_{\leftarrow}x_5=P(x_5|x_3)$$
$$m_{\leftarrow}x_6=P(x_6|x_3)$$
$$m_{\leftarrow}x_3=\sum_{x_5}\sum_{x_6}P(x_5|x_3)P(x_6|x_3)$$
$$\tilde{f}_3(x_3)=(\sum_{x_1}\sum_{x_2}P(x_1)P(x_3|x_1,x_2)P(x_2)(\sum_{x_4}P(x_4|x_2)))\sum_{x_5}\sum_{x_6}P(x_5|x_3)P(x_6|x_3)$$

因此加法次数是$O(n^3)$，[illegible]算法可以提高效率。