



# 软工18级多元统计作业1

总分 180

## 判断题

得分：34 总分：34

1-1 因子分析把变量分成公共因子和独立因子两部分因素。 (1分)

- T  F

评测结果 答案正确 (1分)

1-2 选取主成分还可根据特征值的变化来确定。求解主成分的过程实际就是对矩阵结构进行分析的过程。 (T) (1分)

- T  F

评测结果 答案正确 (1分)

1-3 对于度量单位不同的指标或是取值范围彼此差异非常大的指标，可直接从协方差矩阵出发进行主成分分析。 (1分)

- T  F

评测结果 答案正确 (1分)

1-4 若A是退化矩阵，则A-1一定存在。 (1分)

- T  F

评测结果 答案正确 (1分)

1-5 若A为p阶对称矩阵，则存在正交矩阵T和对角矩阵Λ=diag(λ\_1, λ\_2, …, λ\_p)，使得A=TΛT'。 (1分)

- T  F

评测结果 答案正确 (1分)

1-6 若向量x和y的内积为0，则说明向量x和y垂直。 (1分)

- T  F

评测结果 答案正确 (1分)

1-7 若A是一个正交矩阵，则A的行列式为1。 (1分)

- T  F

评测结果 答案正确 (1分)

1-8 若A和B均为p阶方阵，则|AB|=|A||B|。 (1分)

- T  F

评测结果 答案正确 (1分)

1-9 为了使得因子分析的结果更易于解释，进行正交因子旋转，旋转后，新的公共因子仍然彼此独立。 (1分)

- T  F

评测结果 答案正确 (1分)

1-10 因子得分是为了考察每一个样品性质之间的关系。 (1分)

- T  F

1-11 因子分析中,载荷矩阵中的每一个元素代表的是变量 $X_i$ 与公共因子 $F_j$ 之间的关系。 (1分)

 T  F

评测结果 答案正确 (1分)

1-12 因子分析中,利用主成分法求解的载荷系数与主成分分析中的主成分线性方程的系数一样。 (1分)

 T  F

评测结果 答案正确 (1分)

1-13 因子分析中,公因子 $F_j$ 的方差贡献表示的是公共因子 $F_j$ 对于原始数据 $X$ 中的每一分量 $X_i$  ( $i=1, 2, \dots, p$ ) 所提供的方差的总和。 (1分)

 T  F

评测结果 答案正确 (1分)

1-14 利用主成分法得到的因子载荷是唯一的。 (1分)

 T  F

评测结果 答案正确 (1分)

1-15 聚类分析仅能进行样本聚类。 (1分)

 T  F

评测结果 答案正确 (1分)

1-16 聚类分析属于有指导的学习分类方法。 (1分)

 T  F

评测结果 答案正确 (1分)

1-17 进行样品聚类分析时,“靠近”往往由某种距离来刻画。 (1分)

 T  F

评测结果 答案正确 (1分)

1-18 进行多元数据的指标聚类时,可根据相关系数或某种关联性度量来聚类。 (1分)

 T  F

评测结果 答案正确 (1分)

1-19 类平均法进行系统聚类的思想是来源于方差分析,如果类分得正确,同类样品的离差平方和应当较小,类与类之间的离差平方和应当较大。 (1分)

 T  F

评测结果 答案正确 (1分)

1-20 聚类分析时可通过碎石图确定最终的分类数。 (1分)

 T  F

评测结果 答案正确 (1分)



1-21 针对同一多元数据,不同的方法聚类的结果相同。 (1分)

 T  F

**评测结果** 答案正确 (1分)

1-22 系统聚类法中，对于那些先前已被“错误”分类的样品不再提供重新分类的机会。 (1分)

T  F

**评测结果** 答案正确 (1分)

1-23 K-均值法只能用于对样品的聚类，而不能用于对变量的聚类。 (1分)

T  F

**评测结果** 答案正确 (1分)

1-24 有序样品的聚类的实质上是需要找出一些分点，将数据划分成几个分段，每个分段看作一类，称这种分类也可称为分割。 (1分)

T  F

**评测结果** 答案正确 (1分)

1-25 k均值法的类个数需事先指定。 (1分)

T  F

**评测结果** 答案正确 (1分)

1-26 主成分分析实质上是线性变换，无假设检验。 (1分)

T  F

**评测结果** 答案正确 (1分)

1-27 一般地说，从同一原始变量的协方差矩阵出发求得的主成分与从原始变量的相关矩阵出发求得的主成分相同。 (1分)

T  F

**评测结果** 答案正确 (1分)

1-28 选取主成分还可根据特征值的变化来确定。求解主成分的过程实际就是对矩阵结构进行分析的过程。 (T) (1分) (1分)

T  F

**评测结果** 答案正确 (1分)

1-29 对于度量单位不同的指标或是取值范围彼此差异非常大的指标，可直接从协方差矩阵出发进行主成分分析。 (1分)

T  F

**评测结果** 答案正确 (1分)

1-30 主成分分析时，对数据进行标准化的过程可能抹杀原始变量离散程度差异。 (1分)

T  F

**评测结果** 答案正确 (1分)

1-31 主成分分析要求数据来自于正态总体。 (1分)

T  F

**评测结果** 答案正确 (1分)

1-32 对来自多元正态总体的数据，主成分分析的主成分就是按数据离散程度最大的方向进行坐标轴旋转。 (1分)

T  F

1-33 主成分分析只是要达到目的的一个中间结果（或步骤），没有实际意义。

(1分)

- T       F

评测结果 答案正确 (1分)

1-34 在主成分分析的过程中，得到的各个主成分的方差依次递减。

(1分)

- T       F

评测结果 答案正确 (1分)

**单选题**

得分：16 总分：21

2-1 设A是3阶方阵，且 $|A|=-2$ ，则 $|A^{-1}|=()$ 。

(1分)

- A.  $1/2$   
 B.  $-2$   
 C.  $2$   
 D.  $-1/2$

评测结果 答案正确 (1分)

2-2 对于任意n阶方阵A,B，总有()。

(1分)

- A.  $AB=BA$   
 B.  $|AB|=|BA|$   
 C.  $(AB)'=A'B'$   
 D.  $(AB)^2=A^2 B^2$

评测结果 答案正确 (1分)

2-3 设 $A=\text{diag}(a_{11}, a_{22}, \dots, a_{pp})$ 且 $a_{ii} \neq 0 (i=1, 2, \dots, p)$ ，以下说法错误的是()。

(1分)

- A.  $a_{11}, a_{22}, \dots, a_{pp}$ 为矩阵A的特征值  
 B. 矩阵A有小于p个特征向量  
 C.  $(1, 0, 0, \dots, 0)'$ 是矩阵A的特征向量之一  
 D. 矩阵A的特征向量构成的特征矩阵是正交矩阵

评测结果 答案正确 (1分)

2-4 设 $\alpha_1, \alpha_2, \dots, \alpha_m$ 均为n维向量，那么，下面关于向量的说法正确的是()。

(1分)

- A. 若 $k_1\alpha_1+k_2\alpha_2+\dots+k_m\alpha_m=0$ ，则 $\alpha_1, \alpha_2, \dots, \alpha_m$ 线性相关  
 B. 若对任意一组不全为零的数 $k_1, k_2, \dots, k_m$ ，都有 $k_1\alpha_1+k_2\alpha_2+\dots+k_m\alpha_m \neq 0$ ，则 $\alpha_1, \alpha_2, \dots, \alpha_m$ 线性无关  
 C. 若 $\alpha_1, \alpha_2, \dots, \alpha_m$ 线性相关，则对任意一组不全为零的数 $k_1, k_2, \dots, k_m$ ，都有 $k_1\alpha_1+k_2\alpha_2+\dots+k_m\alpha_m=0$   
 D. 若 $0 \cdot \alpha_1 + 0 \cdot \alpha_2 + \dots + 0 \cdot \alpha_m = 0$ ，则 $\alpha_1, \alpha_2, \dots, \alpha_m$ 线性无关

评测结果 答案正确 (1分)

2-5 设 $a=(2, -4, 1)', b=(3, 5, -1)'$ ， $ab'$ 的非零特征值为()。

(1分)

- A.  $-10$   
 B.  $-15$



C. -12

D. -8

**评测结果 答案正确 (1分)**

2-6 如果对某公司在一个城市中的各个营业点按彼此之间的路程远近来进行聚类，则最适合采用的距离是 (1分)

A. 马氏距离

B. 绝对值距离

C. 欧氏距离

D. 各变量标准化之后的欧氏距离

**评测结果 答案正确 (1分)**

2-7 不适合对变量聚类的方法有 (1分)

A. 最长距离法

B. 类平均法

C. K均值法

D. 最短距离法

**评测结果 答案正确 (1分)**

2-8 聚类分析的目的是 (1分)

A. 降维

B. 抽样

C. 检验

D. 描述

**评测结果 答案正确 (1分)**

2-9 对样本分类还不清楚的时候，适合首先进行哪种分析 (1分)

A. 判别分析

B. 因子分析

C. 聚类分析

D. 主成分分析

**评测结果 答案正确 (1分)**

2-10 适合对大样本数据进行聚类分析的方法是 (1分)

A. 系统聚类

B. K-均值聚类

C. 条件聚类

D. 有序聚类

**评测结果 答案正确 (1分)**



2-11 根据聚类的对象不同，聚类可以分为 (1分)

- A. 样本聚类和变量聚类
- B. 有序聚类和K均值聚类
- C. 系统聚类和动态聚类
- D. 模糊聚类和系统聚类

评测结果 答案正确 (1分)

2-12 根据聚类的对象不同，聚类可以分为 (1分)

- A. 样本聚类和变量聚类
- B. 有序聚类和K均值聚类
- C. 系统聚类和动态聚类
- D. 模糊聚类和系统聚类

评测结果 答案正确 (1分)

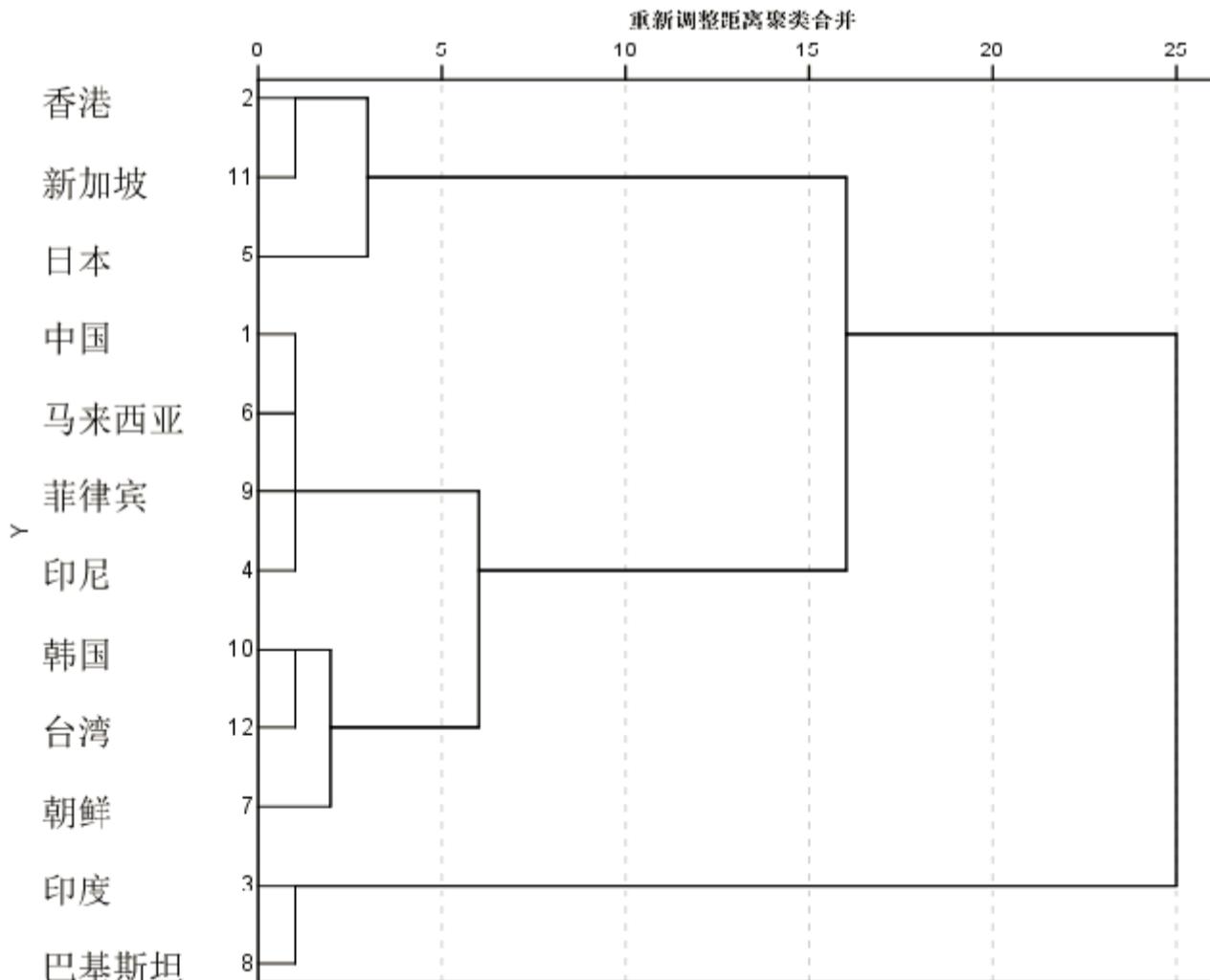
2-13 聚类分析中，通常使用（ ）来衡量两个对象之间的相异度 (1分)

- A. 距离
- B. 位置
- C. 大小
- D. 比较

评测结果 答案正确 (1分)

2-14 一项研究中对中国、香港等12个国家或地区的多项经济指标进行了聚类分析，得到的树状图如下， (1分)

使用质心联接的树状图



说法不正确的是

- A. 分为三类时，日本、香港、新加坡在同一个类内
- B. 分为两类时，日本、香港、新加坡在同一个类内
- C. 分为四类时，印度和巴基斯坦构成一个类



- D. 分为四类时，日本、中国、马来西亚构成一个类

评测结果 答案正确 (1分)

2-15 主成分分析中， (1分)

- A. 有必要考虑变量的量纲，避免出现“大数吃小数”  
 B. 需要考虑样品间的相关性  
 C. 得到的各个主成分实际上是影响各原始变量的潜在因素  
 D. 有多少个原始变量，就需要选取多少个主成分

评测结果 答案正确 (1分)

2-16 关于第一个主成分，下列描述正确的是 (1分)

- A. 与原始变量的相关性最大  
 B. 其系数的绝对值显著地比其他主成分系数的绝对值大  
 C. 与第二个主成分之间的相关性最大  
 D. 其累积方差贡献率最大

评测结果 答案正确 (1分)

2-17 关于第一个主成分，下列描述正确的是 (1分)

- A. 与原始变量的相关性最大  
 B. 其系数的绝对值显著地比其他主成分系数的绝对值大  
 C. 与第二个主成分之间的相关性最大  
 D. 其累积方差贡献率最大

评测结果 答案正确 (1分)

2-18 主成分分析中各主成分之间是 (1分)

- A. 相互独立  
 B. 彼此相关  
 C. 存在线性相关  
 D. 互不相关

评测结果 答案正确 (1分)

2-19 在利用主成分分析进行综合评价时，要对样本观测进行一些变换，最常用的是 (1分)

- A. 同向化变换  
 B. 标准化变换  
 C. 对数变换  
 D. 极差规格化变换

评测结果 答案正确 (1分)

2-20 为了更充分有效地代表原始变量的信息，不同的主成分应携带不同的信息。以第一、第二主成分Y1\Y2为例， (1分)

- A.  $\text{cov}(Y_1, Y_2) \neq 0$



- B. Y<sub>1</sub>和Y<sub>2</sub>高度相关
- C. cov(Y<sub>1</sub>,Y<sub>2</sub>)=0
- D. cov(Y<sub>1</sub>,Y<sub>2</sub>)=1

评测结果 答案正确 (1分)

2-21 主成分分析的主要任务有 (1分)

- A. 计算因子个数
- B. 确定主成分个数
- C. 计算平均值
- D. 预测变量分类

评测结果 答案正确 (1分)

### 填空题

得分: 24 总分: 30

4-1 设随机向量X=(x<sub>1</sub>, x<sub>2</sub>, x<sub>3</sub>, x<sub>4</sub>)'的相关阵R为

$$\begin{pmatrix} 1 & 0.2 & -0.5 & 0.4 \\ 0.2 & 1 & 0.3 & 0.8 \\ -0.5 & 0.3 & 1 & 0.6 \\ 0.4 & 0.8 & 0.6 & 1 \end{pmatrix}$$

则x<sub>1</sub>和x<sub>3</sub>的相关系数为 -0.5 (1分)。

评测结果 答案正确 (1分)

测试点得分	序号	结果	得分
	0	答案正确	1

4-2 聚类和分类的区别是, 分类 (1分) 分析是一种有监督学习方法, 而聚类 (1分) 分析是一种无监督学习方法。

评测结果 答案正确 (2分)

测试点得分	序号	结果	得分
	0	答案正确	1
	1	答案正确	1

4-3 因子分析中, 将每个原始变量分解为两个部分, 一个部分由所有变量共同具有的少数几个公共因子 (1分) 组成的, 另一个部分是每个变量独自具有的因素, 即特殊因子 (1分)。

评测结果 答案正确 (2分)

测试点得分	序号	结果	得分
	0	答案正确	1
	1	答案正确	1

4-4 设a=(2,-4,1)', b=(4,1,-4)', 则a和b的夹角为 90 (1分)。

评测结果 答案正确 (1分)

测试点得分	序号	结果	得分
	0	答案正确	1

4-5 若A为4阶非退化矩阵, 若2为矩阵A的一个特征值, 对应的特征向量为 (1, 0, 3, 4), 则A逆矩阵的一个特征值为 0.5 (1分), 对应特征向量为 (1, 0, 3, 4) (1分))。

评测结果 答案正确 (2分)

测试点得分	序号	结果	得分

序号	结果	得分
0	答案正确	1
1	答案正确	1

4-6 设矩阵  $A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 2 & 3 & \lambda + 1 \end{pmatrix}$  的秩为2, 则  $\lambda = \boxed{1}$  (1分)。

评测结果 答案正确 (1分)

测试点得分	序号	结果	得分
	0	答案正确	1

4-7 设  $X = (x_1, x_2, x_3)$  为标准化后的随机变量, 将其协方差矩阵通过因子分析分解为:

$$\begin{aligned}\Sigma &= \begin{pmatrix} 1 & -\frac{1}{3} & \frac{2}{3} \\ -\frac{1}{3} & 1 & 0 \\ \frac{2}{3} & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0.934 & 0 & 0 \\ -0.417 & 0.894 & 0 \\ 0.835 & 0.447 & 1 \end{pmatrix} \begin{pmatrix} 0.934 & -0.417 & 0.835 \\ 0 & 0.894 & 0.447 \\ 0 & 0 & 1 \end{pmatrix} \\ &\quad + \begin{pmatrix} 0.128 & 0 & 0 \\ 0 & 0.027 & 0 \\ 0 & 0 & 0.103 \end{pmatrix}\end{aligned}$$

$x_1$  的共同度  $h_1^2 = \boxed{0.872}$  (1分),  $x_1$  的剩余方差  $\sigma_1^2 = \boxed{0.128}$  (1分), 公因子  $F_1$  与  $x_1$  的协方差 =  $\boxed{0.934}$  (1分)。 (结果保留小数点后三位)

评测结果 答案正确 (1分)

测试点得分	序号	结果	得分
	0	答案正确	1
	1	答案正确	1
	2	答案正确	1

4-8 聚类分析 (1分) 是将分类对象分成若干类, 相似的归为同一类, 不相似的归为不同的类。

评测结果 答案正确 (1分)

测试点得分	序号	结果	得分
	0	答案正确	1

4-9 进行系统聚类分析时, 计算初始6个样本 ( $X_1 \dots X_6$ ) 的距离矩阵为:

$$D_0 = \left[ \begin{array}{cccccc} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \\ X_1 & 0 & & & & & \\ X_2 & 3 & 0 & & & & \\ X_3 & 5 & 6 & 0 & & & \\ X_4 & 6 & 8 & 3 & 0 & & \\ X_5 & 9 & 8 & 3 & 4 & 0 & \\ X_6 & 6 & 5 & 2 & 1 & 0 & \end{array} \right],$$

若类之间连接应用最大距离方法, 最先聚类的是  $X_5$  和  $X_6$  |  $X_5, X_6 | X_5, X_6$  (1分) 和。

评测结果 答案正确 (1分)

测试点得分	序号	结果	得分
	0	答案正确	1

4-10 进行系统聚类分析时, 计算初始6个样本 ( $X_1 \dots X_6$ ) 的距离矩阵为:

$$D_0 = \begin{bmatrix} \dots & X1 & X2 & X3 & X4 & X5 & X6 \\ X1 & 0 & \dots & \dots & \dots & \dots & \dots \\ X2 & 3 & 0 & \dots & \dots & \dots & \dots \\ X3 & 5 & 6 & 0 & \dots & \dots & \dots \\ X4 & 6 & 8 & 3 & 0 & \dots & \dots \\ X5 & 9 & 8 & 3 & 4 & 0 & \dots \\ X6 & 6 & 5 & 5 & 2 & 1 & 0 \end{bmatrix},$$

若类之间连接应用最小距离方法，最先聚类的是 X5, X6 (1分)。

**评测结果** 答案正确 (1分)

测试点得分	序号	结果	得分
	0	答案正确	1

4-11 进行系统聚类分析时，计算初始6个样本 (X1...X6) 的距离矩阵为：

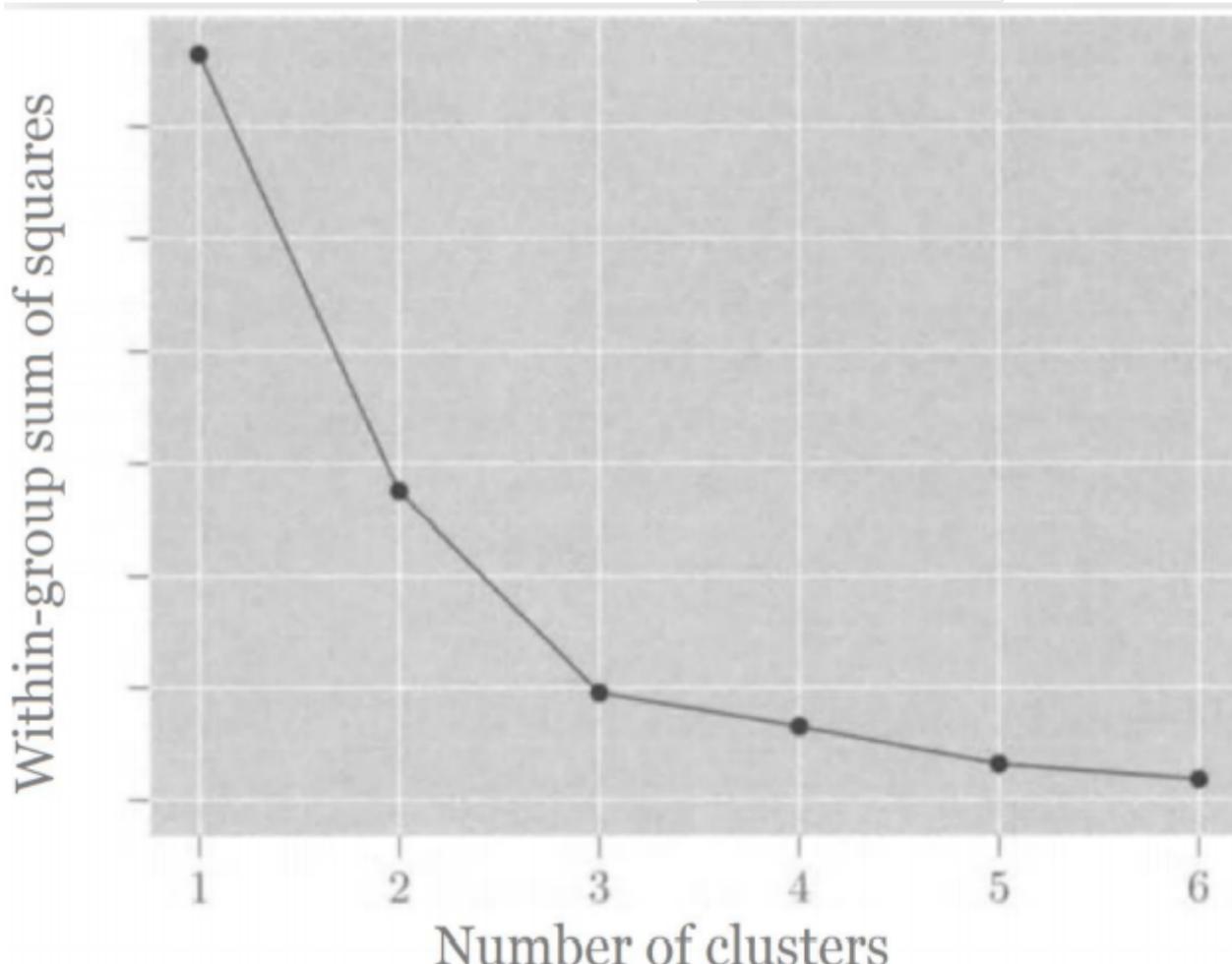
$$D_0 = \begin{bmatrix} \dots & X1 & X2 & X3 & X4 & X5 & X6 \\ X1 & 0 & \dots & \dots & \dots & \dots & \dots \\ X2 & 3 & 0 & \dots & \dots & \dots & \dots \\ X3 & 5 & 6 & 0 & \dots & \dots & \dots \\ X4 & 6 & 8 & 3 & 0 & \dots & \dots \\ X5 & 9 & 8 & 3 & 4 & 0 & \dots \\ X6 & 6 & 5 & 5 & 2 & 1 & 0 \end{bmatrix},$$

若类之间连接应用最小距离方法，假设将X5和X6聚为一类定义为X7，则X7与X1的距离d(7,1)=6 (1分)。

**评测结果** 答案正确 (1分)

测试点得分	序号	结果	得分
	0	答案正确	1

4-12 进行K-均值聚类时，碎石图如图所示，则最优的分类数为 3 (1分)。



**评测结果** 答案正确 (1分)

测试点得分	序号	结果	得分
	0	答案正确	1

4-13 聚类分析中，**模糊聚类** (1分)方法的基本思想是通过优化目标函数得到每个样本点对所有类中心的隶属度，从而对样本进行自动分类。

**评测结果** 答案正确 (1分)**测试点得分**

序号	结果	得分
0	答案正确	1

4-14 Q型聚类法是按样品进行聚类, R型聚类法是按变量进行聚类, 9. Q型聚类相似度统计量是距离 (1分), 而R型聚类统计量通常采用相似系数 (1分)。

**评测结果** 答案正确 (2分)**测试点得分**

序号	结果	得分
0	答案正确	1
1	答案正确	1

4-15

常用的Minkowski距离公式为  $d(\mathbf{x}, \mathbf{y}) = \left[ \sum_{i=1}^p |x_i - y_i|^q \right]^{1/q}$ , 当q=1时, 它表示曼哈顿距离 (1分)

**评测结果** 答案正确 (1分)**测试点得分**

序号	结果	得分
0	答案正确	1

4-16

常用的Minkowski距离公式为  $d(\mathbf{x}, \mathbf{y}) = \left[ \sum_{i=1}^p |x_i - y_i|^q \right]^{1/q}$ , 当q=2时, 它表示欧式距离 (1分)

**评测结果** 答案正确 (1分)**测试点得分**

序号	结果	得分
0	答案正确	1

4-17

常用的Minkowski距离公式为  $d(\mathbf{x}, \mathbf{y}) = \left[ \sum_{i=1}^p |x_i - y_i|^q \right]^{1/q}$ , 当q趋于正无穷时, 它表示切比雪夫距离 (1分)。

**评测结果** 答案正确 (1分)**测试点得分**

序号	结果	得分
0	答案正确	1

4-18 主成分分析 (1分)是利用降维的思想, 在损失很少信息的前提下把多个指标转化为几个综合指标的多元统计方法。

**评测结果** 答案正确 (1分)**测试点得分**

序号	结果	得分
0	答案正确	1

4-19 主成分分析通常把转化生成的综合指标称之为成分, 其中每个成分都是原始变量的线性组合 (1分), 且各个成分之间互不相关, 这就使得成分比原始变量具有某些更优越的性能。

**评测结果** 答案正确 (1分)**测试点得分**

序号	结果	得分
0	答案正确	1

4-20 主成分分析中我们所说的保留原始变量尽可能多的信息, 也就是指的生成的较少的综合变量的方差和尽可能接近于原始变量方差 (1分)的总和。

**评测结果** 答案正确 (1分)

测试点得分

序号	结果	得分
0	答案正确	1



该文档是极速PDF编辑器生成,  
如果想去掉该提示,请访问并下载:  
<http://www.jisupdfeditor.com/>

4-21 主成分分析中可以利用 协方差矩阵|相关矩阵 (1分)求解主成分。

评测结果 答案正确 (1分)

序号	结果	得分
0	答案正确	1

4-22 对多元数据X ( $x_1, x_2, x_3, x_4$ ) 进行了主成分分析, 样本的特征值 $\lambda_1=2.857, \lambda_2=0.809, \lambda_3=0.702, \lambda_4=0.025$ , 则第一主成分的方差贡献率是  
65 (1分)% (保留百分数的整数位)。

评测结果 答案正确 (1分)

序号	结果	得分
0	答案正确	1

4-23 对多元数据X ( $x_1, x_2, x_3, x_4, x_5$ ) 进行了主成分分析, 样本的特征值 $\lambda_1=2.857, \lambda_2=0.809, \lambda_3=0.609, \lambda_4=0.521, \lambda_5=0.203$ 对应特征向量 $p_1=(0.464, 0.457, 0.470, 0.421, 0.421), p_1=(0.240, 0.509, 0.260, -0.526, -0.582)$ , 则第一主成分Y1的计算公式是  $Y_1=0.464x_1+0.457x_2+0.470x_3+0.421x_4+0.421x_5$  (1分)。

评测结果 答案正确 (1分)

序号	结果	得分
0	答案正确	1

4-24 在进行主成分分析得出协方差阵或是相关阵发现最小特征根接近于零时, 意味着中心化以后的原始变量之间存在 多重共线性 (1分), 即原始变量存在着不可忽视的重叠信息。

评测结果 答案正确 (1分)

序号	结果	得分
0	答案正确	1

## 编程题

得分: 95 总分: 95

7-1 求随机矩阵的特征值和特征向量 (15分)

输入25个整型数据, 然后将其转换为5X5的矩阵, 求其特征值及特征向量, 正确的输入格式请见样例, 若输入的格式不是int型或者输入个数不对, 输出: 输入有错!

输入样例:

1 2 3 4 5 6 7 8 9 10 1 2 1 3 14 12 12 14 1 4 5 7 8 9 23

输出样例:

```
[38.06636747 -7.45341787  4.24478256 -0.35682312 -1.50090904]
 [[ 0.18862655  0.26052599 -0.10151955  0.70437575  0.65107271]
 [ 0.46119015  0.36485524 -0.66904382 -0.70374152  0.18805252]
 [ 0.3246663 -0.04587277  0.45913749 -0.01841258 -0.73179827]
 [ 0.40776924 -0.88072456 -0.48315401  0.08433305 -0.03269975]
 [ 0.69284898  0.14569421  0.31277647  0.03393505  0.06436303]]
```

编译器 PYTHON3

代码

```
import numpy as np

try:
    a = input()
    n = []
    num = [int(n) for n in a.split(' ')]
    if len(num) != 25:
        print('输入有错!')
        exit(0)
    for i in range(0, len(num), 5):
```



```

x = num[i:i+5]
n.append(x)
n = np.matrix(n)
w, v = np.linalg.eig(n)
print(w, end=' \n ')
print(v)
except ValueError:
    print("输入有错! ")

```

**评测结果** 答案正确 (15 分)

测试点得分	测试点	结果	得分	耗时	内存
	0	答案正确	5	285.00 ms	14256 KB
	1	答案正确	5	237.00 ms	14260 KB
	2	答案正确	5	305.00 ms	14148 KB

## 7-2 求相关系数矩阵 (5分)

有两组数据，分别是31个地区城镇居民年人均可支配收入和年人均消费性支出数据，人均可支配收入为：15637.84 11467.16 7951.31 7902.86 8122.99 8007.56 7840.61 7470.71 16682.82 10481.93 14546.38 7511.43 11175.37 7559.64 9437.8 7704.9 8022.75 8617.48 13627.65 8689.99 7735.78 9220.96 7709.87 7322.05 8870.88 9106.07 7492.47 7376.74 7319.67 7217.87 7503.42，其对应人均消费性支出数据为：12200.4 8802.44 5819.18 5654.15 6219.26 6543.26 6068.99 5567.53 12631.03 7332.26 10636.14 5711.33 8161.15 5337.84 6673.75 5294.19 6398.52 6884.61 10694.79 6445.73 5802.4 7973.05 6371.14 5494.45 6837.01 8338.21 6233.07 5937.3 5758.95 5821.38 5773.62，请求人均可支配收入和年人均消费性支出的相关矩阵。

### 输入格式：

请在这里写输入格式。例如：输入在一行中给出2个绝对值不超过1000的整数A和B。

### 输出格式：

请在这里描述输出格式。例如：对每一组输入，在一行中输出A+B的值。

### 输入样例：

在这里给出一组输入。例如：

```

15637.84 11467.16 7951.31 7902.86 8122.99 8007.56 7840.61 7470.71 16682.82 10481.93 14546.38 7511.43 11175.37
7559.64 9437.8 7704.9 8022.75 8617.48 13627.65 8689.99 7735.78 9220.96 7709.87 7322.05 8870.88 9106.07 7492.47
7376.74 7319.67 7217.87 7503.42 12200.4 8802.44 5819.18 5654.15 6219.26 6543.26 6068.99 5567.53 12631.03 7332.26
10636.14 5711.33 8161.15 5337.84 6673.75 5294.19 6398.52 6884.61 10694.79 6445.73 5802.4 7973.05 6371.14 5494.45
6837.01 8338.21 6233.07 5937.3 5758.95 5821.38 5773.62
2 31

```

### 输出样例：

在这里给出相应的输出。例如：

```

[[1.          0.97510266]
 [0.97510266 1.        ]]

```

**编译器** PYTHON3

**代码**

```

import numpy as np

s1 = input()
s2 = input()
n = []
num = [float(n) for n in s1.split()]
for i in range(0, len(num), 31):
    x = num[i:i+31]
    n.append(x)
n = np.matrix(n)
nn = np.corrcoef(n)
print(nn)

```

**评测结果** 答案正确 (5 分)

测试点得分	测试点	结果	得分	耗时	内存
	0	答案正确	5	234.00 ms	14288 KB

## 7-3 分类判别 (10分)

已知有两类数据和二者的先验概率，已知 $P(w1)=0.6$ ,  $P(w2)=0.4$ 。 w1中数据点的坐标对应如下：  $x1 = 0.23, 1.52, 0.65, 0.77, 1.05, 1.19, 0.29, 0.25, 0.66, 0.56, 0.90, 0.13, -0.54, 0.94, -0.21, 0.05, -0.08, 0.73, 0.33, 1.06, -0.02, 0.11, 0.31, 0.66$   $y1 = 2.34, 2.19, 1.67, 1.63, 1.78, 2.01, 2.06, 2.12, 2.47, 1.51, 1.96, 1.83, 1.87, 2.29, 1.77, 2.39, 1.56, 1.93, 2.20, 2.45, 1.75, 1.69, 2.48, 1.72$  W2中数据点的坐标对应如下：  $x2 = 1.40, 1.23, 2.08, 1.16, 1.37, 1.18, 1.76, 1.97, 2.41, 2.58, 2.84, 1.95, 1.25, 1.28, 1.26, 2.01, 2.18, 1.79, 1.33, 1.15, 1.70, 1.59, 2.93, 1.46$   $y2 = 1.02, 0.96, 0.91, 1.49, 0.82, 0.93, 1.14, 1.06, 0.81, 1.28, 1.46, 1.43, 0.71, 1.29, 1.37, 0.93, 1.22, 1.18, 0.87, 0.55, 0.51, 0.99, 0.91, 0.71$

## 输入格式：

## 输出格式：

## 输入样例：

在这里给出一组输入。例如：

```
0.23,1.52,0.65,0.77,1.05,1.19,0.29,0.25,0.66,0.56,0.90,0.13,-0.54,0.94,-0.21,0.05,-0.08,0.73,0.33,1.06,-0.02,0.11,0.31,0.66  
2.34,2.19,1.67,1.63,1.78,2.01,2.06,2.12,2.47,1.51,1.96,1.83,1.87,2.29,1.77,2.39,1.56,1.93,2.20,2.45,1.75,1.69,2.48,1.72  
1.40,1.23,2.08,1.16,1.37,1.18,1.76,1.97,2.41,2.58,2.84,1.95,1.25,1.28,1.26,2.01,2.18,1.79,1.33,1.15,1.70,1.59,2.93,1.46  
1.02,0.96,0.91,1.49,0.82,0.93,1.14,1.06,0.81,1.28,1.46,1.43,0.71,1.29,1.37,0.93,1.22,1.18,0.87,0.55,0.51,0.99,0.91,0.71  
1,1.5
```

## 输出样例：

在这里给出相应的输出。例如：

该点属于第二类

编译器 PYTHON3

代码

```
from numpy import *  
import numpy as np  
import math  
  
x1 = array([float(i) for i in input().split(',')])  
y1 = array([float(i) for i in input().split(',')])  
w1 = mat(vstack((x1,y1)))  
x2 = array([float(i) for i in input().split(',')])  
y2 = array([float(i) for i in input().split(',')])  
w2 = mat(vstack((x2, y2)))  
  
mean1 = np.mean(w1, 1)  
mean2 = np.mean(w2, 1)  
  
dimens1, nums1 = w1.shape[:2]  
samples_mean1 = w1 - mean1  
s_in1 = 0  
for i in range(nums1):  
    x = samples_mean1[:, i]  
    s_in1 += dot(x, x.T)  
  
dimens2, nums2 = w2.shape[:2]  
samples_mean2 = w2 - mean2  
s_in2 = 0  
for i in range(nums2):  
    x = samples_mean2[:, i]  
    s_in2 += dot(x, x.T)  
  
s = s_in1 + s_in2  
  
s_t = s.I  
  
w = dot(s_t, mean1 - mean2)  
  
w_new = w.T  
m1_new = dot(w_new,mean1)  
m2_new = dot(w_new,mean2)  
pw1 = 0.6  
pw2 = 0.4  
w0 = m1_new*pw1+m2_new*pw2
```

```

x = mat(array([float(i) for i in input().split(',')])).reshape(2,1)
y_i = w_new * x[:,0]
if y_i > w0:
    print('该点属于第一类')
else:
    print('该点属于第二类')

```

**评测结果** 答案正确 (10分)

测试点得分	测试点	结果	得分	耗时	内存
	0	答案正确	5	283.00 ms	14300 KB
	1	答案正确	5	319.00 ms	14324 KB

#### 7-4 写出贡献率最大的主成分线性方程 (10分)

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

有一组数据 Data = [2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2, 1, 1.5, 1.1], 现在要用主成分做降维处理，且原始数据在降维前需采用均值标准化后才参与计算，最后给出特征值最大

的第一主成分的变换公式。

#### 输入格式:

请在这里写输入格式。例如：输入在一行中给出2个绝对值不超过1000的整数A和B。

#### 输出格式:

请在这里描述输出格式。例如：对每一组输入，在一行中输出A+B的值。

#### 输入样例:

在这里给出一组输入。例如：

```
2.5,2.4,0.5,0.7,2.2,2.9,1.9,2.2,3.1,3.0,2.3,2.7,2,1.6,1,1.1,1.5,1.6,1.1,0.9
10,2
```

#### 输出样例:

在这里给出相应的输出。例如：

```
第1主成分=-0.67787*(x1-1.81)-0.73518*(x2-1.91)
```

**编译器** PYTHON3

**代码**

```

import numpy as np

a = input()
b = input()

bb = np.array([int(i) for i in b.split(',')])  

num = np.array([float(i) for i in a.split(',')]).reshape(bb[0], bb[1])

x = (num - np.mean(num)) / np.std(num)
x1 = np.cov(x.T)
w, v = np.linalg.eig(x1)
if w[0] > w[1]:
    index = 0
else:
    index = 1
print("第1主成分={:.5f}*(x1-{:.2f}){:+.5f}*(x2-{:.2f})".format(v[0][index], np.mean(num, axis=0)[0], v[1][index], np.mean(num, axis=0)[1]))

```

**评测结果** 答案正确 (10分)

测试点得分	测试点	结果	得分	耗时	内存
	0	答案正确	5	351.00 ms	14284 KB
	1	答案正确	5	342.00 ms	14184 KB

## 7-5 K均值聚类的实现 (15分)

已知一批数据，有两个变量：销售额和利润，有8个样本：

公司名称	A	B	C	D	E	F	G	H
销售额	174	90	54	161	86	24	37	33
利润	54	28	15	133	9	19	7	86

采用K均值聚类算法对样本进行聚类，判断公司A所在类，并输出该类的中心坐标。

### 输入格式：

第一行输入所有的变量值（中间以空格分隔），例如：174 54 90 28 54 15 161 133 86 9 24 19 37 7 33 86；第二行输入样本和变量的个数（中间以空格分隔），例如：8 2。第三行输入聚类的个数，例如：3。

### 输出格式：

输出最终判断结果（保留小数点后两位）。例如：A公司所在类的中心为：167.50,93.50。

### 输入样例：

```
174 54 90 28 54 15 161 133 86 9 24 19 37 7 33 86  
8 2  
3
```

```
174 54 90 28 54 15 161 133 86 9 24 19 37 7 33 86  
8 2  
4
```

```
174 54 90 28 54 15 161 133 86 9 24 19 37 7 33 86  
8 2  
2
```

### 输出样例：

```
A公司所在类的中心为：167.50,93.50。
```

```
A公司所在类的中心为：174.00,54.00。
```

```
A公司所在类的中心为：167.50,93.50。
```

### 编译器 PYTHON3

#### 代码

```
import numpy as np  
from sklearn.cluster import KMeans  
  
a = input()  
b = input()  
c = input()  
t1 = np.array([float(i) for i in a.split(' ')])  
t2, t3 = np.array([int(i) for i in b.split(' ')])  
t4 = int(c)  
n = np.array(t1).reshape(t2, t3)  
  
kmeans = KMeans(n_clusters=t4)  
kmeans.fit(n)  
m = kmeans.labels_[0]  
  
print("A公司所在类的中心为: {:.2f},{:.2f}。".format(kmeans.cluster_centers_[m, 0], kmeans.cluster_centers_[m, 1]))
```

### 评测结果 答案正确 (15 分)

测试点得分	测试点	结果	得分	耗时	内存
	0	答案正确	5	846.00 ms	49760 KB
	1	答案正确	5	792.00 ms	49740 KB
	2	答案正确	5	787.00 ms	49932 KB

## 7-6 针对变量的系统聚类实现 (15分)

比较10种咖啡豆的质量，由5名专业人员对每种咖啡的香气、甜感、酸质、醇厚度、风味、余韵和平衡7项指标进行打分，最低分1分，最高分为10分，得到每种咖啡的每项指标的平均得分。

咖啡	香气	甜感	酸质	醇厚度	风味	余韵	平衡
1	4.65	6.32	4.87	4.88	6.73	7.45	8.1
2	8.42	6.45	7.5	4.22	6.11	4.6	4.68
3	6.65	7.56	8.23	8.54	6.81	7.32	4.5
4	6.85	4.03	4.12	6.27	7.8	7.95	7.2
5	6.31	7.52	5.01	6.21	4.95	4.43	6.72
6	7.6	8.01	8.12	6.52	7.42	4.5	6.85
7	4.15	4.12	6.13	7.8	7.95	7.88	6.31
8	7.52	4.15	6.52	4.02	4.03	6.51	7.2
9	8.31	8.26	6.27	7.1	4.12	6.33	4.11
10	4.14	6.36	7.18	8.26	7.98	6.06	6.95

采用系统聚类法对该数据的变量进行聚类（划分为2类），选择最大距离法，距离判断指标选择相关系数，判断香气和酸质是否属于一类，输出结果。

### 输入格式：

第一行输入所有的变量值（中间以空格分隔）例如：4.65 6.32 4.87 4.88 6.73 7.45 8.1 8.42 6.45 7.5 4.22 6.11 4.6 4.68 6.65 7.56 8.23 8.54 6.81 7.32 4.5 6.85 4.03 4.12 6.27 7.8 7.95 7.2 6.31 7.52 5.01 6.21 4.95 4.43 6.72 7.6 8.01 8.12 6.52 7.42 4.5 6.85 4.15 4.12 6.13 7.8 7.95 6.31 7.52 4.15 6.52 4.02 4.03 6.51 7.2 8.31 8.26 6.27 7.1 4.12 6.33 4.11 4.14 6.36 7.18 8.26 7.98 6.06 6.95。第二行输入样本和变量的个数（中间以空格分隔），例如：10 7。第三行输入聚类的个数，例如：2。

### 输出格式：

输出最终判断结果。例如：香气和酸质属于一类或者香气和酸质不属于一类。

### 输入样例：

```
4.65 6.32 4.87 4.88 6.73 7.45 8.1 8.42 6.45 7.5 4.22 6.11 4.6 4.68 6.65 7.56 8.23 8.54 6.81 7.32 4.5 6.85 4.03  
4.12 6.27 7.8 7.95 7.2 6.31 7.52 5.01 6.21 4.95 4.43 6.72 7.6 8.01 8.12 6.52 7.42 4.5 6.85 4.15 4.12 6.13 7.8 7.95  
7.88 6.31 7.52 4.15 6.52 4.02 4.03 6.51 7.2 8.31 8.26 6.27 7.1 4.12 6.33 4.11 4.14 6.36 7.18 8.26 7.98 6.06 6.95  
10 7  
2
```

```
4.65 6.32 4.87 4.88 6.73 7.45 8.1 8.42 6.45 7.5 4.22 6.11 4.6 4.68 6.65 7.56 8.23 8.54 6.81 7.32 4.5 6.85 4.03  
4.12 6.27 7.8 7.95 7.2 6.31 7.52 5.01 6.21 4.95 4.43 6.72 7.6 8.01 8.12 6.52 7.42 4.5 6.85 4.15 4.12 6.13 7.8 7.95  
7.88 6.31 7.52 4.15 6.52 4.02 4.03 6.51 7.2 8.31 8.26 6.27 7.1 4.12 6.33 4.11 4.14 6.36 7.18 8.26 7.98 6.06 6.95  
10 7  
3
```

```
4.65 6.32 4.87 4.88 6.73 7.45 8.1 8.42 6.45 7.5 4.22 6.11 4.6 4.68 6.65 7.56 8.23 8.54 6.81 7.32 4.5 6.85 4.03  
4.12 6.27 7.8 7.95 7.2 6.31 7.52 5.01 6.21 4.95 4.43 6.72 7.6 8.01 8.12 6.52 7.42 4.5 6.85 4.15 4.12 6.13 7.8 7.95  
7.88 6.31 7.52 4.15 6.52 4.02 4.03 6.51 7.2 8.31 8.26 6.27 7.1 4.12 6.33 4.11 4.14 6.36 7.18 8.26 7.98 6.06 6.95  
10 7  
4
```

### 输出样例：

```
香气和酸质属于一类。
```

```
香气和酸质属于一类。
```

```
香气和酸质不属于一类。
```

### 编译器

PYTHON3

### 代码

```
import numpy as np  
from sklearn.cluster import AgglomerativeClustering  
  
temp = np.array([float(i) for i in input().split(' ')])  
n_samplesj, n_features = np.array([int(i) for i in input().split(' ')])  
X = np.array(temp).reshape(n_samplesj, n_features)  
n_clusters = int(input())  
hc=AgglomerativeClustering(n_clusters = n_clusters, affinity = 'correlation', linkage = 'complete')  
hc.fit(X.T)  
hcl=hc.labels_
```

```

if hcl[0]==hcl[2]:
    print("香气和酸质属于一类。")
else:
    print("香气和酸质不属于一类。")

```

**评测结果** 答案正确 (15 分)

测试点得分	测试点	结果	得分	耗时	内存
	0	答案正确	5	855.00 ms	50060 KB
	1	答案正确	5	830.00 ms	50080 KB
	2	答案正确	5	892.00 ms	50132 KB

### 7-7 实现基于相关阵的主成分分析并输出相关参数 (10分)

某面馆有各种种类的汤面，为了得知受欢迎程度，进行了在【面】、【汤】、【配料】3个维度的打分。

编号	面	汤	配料
0	2	1	5
1	2	3	4
2	4	1	3
3	5	4	5
4	3	2	5
5	3	4	2
6	3	5	5
7	1	4	3
8	5	2	3
9	1	2	3

利用主成分分析法对数据挖掘：①求解数据的相关矩阵；②基于相关矩阵对数据进行主成分分析（求解相关阵的特征值）；③输出面和汤的相关系数；第一主成和第二主成分的方差贡献率，以及两者的累计方差贡献率。

#### 输入格式：

第一行输入所有的变量值（中间以空格分隔）例如：2 1 5 2 3 4 4 1 3 5 4 5 3 2 5 3 4 2 3 5 5 1 4 3 5 2 3 1 2 3。第二行输入样本和变量的个数（中间以空格分隔），例如：10 3。

#### 输出格式：

输出面和汤的相关系数；第一主成和第二主成分的方差贡献率，以及两者的累计方差贡献率。

#### 输入样例：

```
2 1 5 2 3 4 4 1 3 5 4 5 3 2 5 3 4 2 3 5 5 1 4 3 5 2 3 1 2 3
10 3
```

```
2 1 5 2 3 4 4 1 3 5 4 5 3 2 5 3 4 2 3 5 5 1 4 3 5 2 3 1 2 3
15 2
```

#### 输出样例：

对面和汤打分的相关系数为-0.01。

第一主成的方差贡献率为37.52%，第二主成分的方差贡献率为33.56%，前两个主成分的累计贡献率为71.08%。

对面和汤打分的相关系数为-0.17。

第一主成的方差贡献率为58.36%，第二主成分的方差贡献率为41.64%，前两个主成分的累计贡献率为100.00%。

**编译器** PYTHON3

**代码**

```

import numpy as np
from sklearn.decomposition import PCA
from sklearn.preprocessing import scale

a = input()
b = input()

bb = np.array([int(i) for i in b.split(' ')])
num = np.array([int(i) for i in a.split(' ')]) . reshape(bb[0], -1)

```



```

x_cor = np.corrcoef(num.T)
w, v = np.linalg.eig(x_cor)
print("对面和汤打分的相关系数为{:.2f}。".format(x_cor[0][1]))

x = scale(num)
pca = PCA(n_components=2)
pca.fit(x)
print("第一主成的方差贡献率为 {:.2%} , 第二主成分的方差贡献率为 {:.2%}, 前两个主成分的累计贡献率为 {:.2%}。".format(pca.explained_variance_ratio_[0], pca.explained_variance_ratio_[1],
(pca.explained_variance_ratio_[0]+pca.explained_variance_ratio_[1])))

```

**评测结果 答案正确 (10 分)**

测试点得分	测试点	结果	得分	耗时	内存
	0	答案正确	5	1447.00 ms	47236 KB
	1	答案正确	5	1382.00 ms	47260 KB

### 7-8 实现矩阵运算，基于标准化数据主成分分析并输出第一主成分系数 (15分)

已知一个10\*2维的矩阵M：

$$\begin{bmatrix} 10.39 & 7.91 \\ 7.64 & 6.94 \\ 6.38 & 5.22 \\ 9.54 & 8.57 \\ 6.24 & 10.51 \\ 20.81 & 17.96 \\ 19.02 & 17.48 \\ 19.58 & 18.43 \\ 17.52 & 17.02 \\ 13.7 & 20.04 \end{bmatrix} \dots$$

实现矩阵M与该矩阵转置的矩阵乘法运算（叉乘），获得n\*n方阵N；② 数据进行Z-score标准化后进行主成分分析（或者直接利用相关阵进行主成分分析）；③输出贡献率最大的主成分（第一主成分）线性方程。

#### 输入格式：

第一行输入所有的变量值（中间以空格分隔）例如：10.39 7.91 7.64 6.94 6.38 5.22 9.54 8.57 6.24 10.51 20.81 17.96 19.01 17.48 19.58 18.43 17.52 17.02 13.7 20.04。

第二行输入样本和变量的个数（中间以空格分隔），例如：10 2。

第三行输入主成分的个数，例如：1。

#### 输出格式：

输出最终判断结果。例如：第1主成分线性方程系数为： -0.21,-0.16,-0.13,-0.2,-0.19,-0.44,-0.41,-0.43,-0.39,-0.38。

#### 输入样例：

```

10.39 7.91 7.64 6.94 6.38 5.22 9.54 8.57 6.24 10.51 20.81 17.96 19.01 17.48 19.58 18.43 17.52 17.02 13.7 20.04
10 2
1

10.39 7.91 7.64 6.94 6.38 5.22 9.54 8.57 6.24 10.51 20.81 17.96 19.01 17.48 19.58 18.43 17.52 17.02 13.7 20.04
10 2
2

10.39 7.91 7.64 6.94 6.38 5.22 9.54 8.57 6.24 10.51 20.81 17.96 19.01 17.48 19.58 18.43 17.52 17.02 13.7 20.04
5 4
1

```

#### 输出样例：

第1主成分线性方程系数为： -0.21,-0.16,-0.13,-0.2,-0.19,-0.44,-0.41,-0.43,-0.39,-0.38。

第2主成分线性方程系数为： -0.27,-0.07,-0.13,-0.1,0.49,-0.3,-0.16,-0.11,-0.04,0.73。

第1主成分线性方程系数为： -0.26,-0.24,-0.45,-0.6,-0.56。

编译器

PYTHON3

代码

```

import numpy as np
from sklearn.decomposition import PCA

```

```

a = input()
b = input()
c = int(input())

bb = np.array([int(i) for i in b.split(' ')])
num = np.mat([float(i) for i in a.split(' '')]).reshape(bb[0], bb[1])

N = num*(num.T)
x = (N - np.mean(N)) / np.std(N)

pca = PCA(n_components = bb[1])
pca.fit(x)
print("第{}主成分线性方程系数为: ".format(c), end=' ')
for i in pca.components_[c-1]:
    if i == (pca.components_[c-1])[bb[0]-1]:
        print("{}".format(round(i, 2)), end='。')
    else:
        print("{}".format(round(i, 2)), end=',')

```

**评测结果** 答案正确 (15 分)

测试点得分	测试点	结果	得分	耗时	内存
	0	答案正确	5	1323.00 ms	46992 KB
	1	答案正确	5	1406.00 ms	47060 KB
	2	答案正确	5	1304.00 ms	47016 KB

