第 3 章 聚类分析

二、计算题

1.

（1）用欧氏距离计算样品之间的距离

$$d_{12}=\sqrt{(1-2)^2+(0-1)^2}=\sqrt{2}$$
$$d_{13}=\sqrt{(1-3)^2+(0-2)^2}=2\sqrt{2}$$
$$d_{14}=\sqrt{(1-4)^2+(0-2)^2}=\sqrt{13}$$
$$d_{15}=\sqrt{(1-3)^2+(0-5)^2}=\sqrt{29}$$
$$d_{23}=\sqrt{(2-3)^2+(1-2)^2}=\sqrt{2}$$
$$d_{24}=\sqrt{(2-4)^2+(1-2)^2}=\sqrt{5}$$
$$d_{25}=\sqrt{(2-3)^2+(1-5)^2}=\sqrt{17}$$
$$d_{34}=\sqrt{(3-4)^2+(2-2)^2}=1$$
$$d_{35}=\sqrt{(3-3)^2+(2-5)^2}=3$$
$$d_{45}=\sqrt{(4-3)^2+(2-5)^2}=\sqrt{10}$$

计算 5 个样品两两之间的距离 d_{ij}（采用欧氏距离），记为距离矩阵 D=（d_{ij}）$_{n\times n}$

	②	③	④	⑤
①	$\sqrt{2}$	$2\sqrt{2}$	$\sqrt{13}$	$\sqrt{29}$
②		$\sqrt{2}$	$\sqrt{5}$	$\sqrt{17}$
③			1	3
④				$\sqrt{10}$

（2）

1）最短距离法对这 5 位销售员进行分类

合并距离最小的两类为新类，按顺序定为第⑥类，$d_{34}=1$ 为最小，⑥={3,4}

计算新类⑥与各当前类的距离，

$$d_{61}=\min\{d_{13},d_{14}\}=2\sqrt{2}$$
$$d_{62}=\min\{d_{23},d_{24}\}=\sqrt{2}$$
$$d_{65}=\min\{d_{35},d_{45}\}=3$$

得到距离矩阵：

	②	⑤	⑥
①	$\sqrt{2}$	$\sqrt{29}$	$2\sqrt{2}$
②		$\sqrt{17}$	$\sqrt{2}$
⑤			3

合并距离最小的两类为新类,$d_{12}=\sqrt{2}$为最小，⑦={1,2}

$$d_{57}=\min\{d_{15},d_{25}\}=\sqrt{17}$$
$$d_{67}=\min\{d_{16},d_{26}\}=\sqrt{2}$$

得到距离矩阵：

	⑥	⑦
⑤	3	$\sqrt{17}$
⑥		$\sqrt{2}$

合并距离最小的两类为新类，$d_{67}=\sqrt{2}$为最小，⑧={⑥，⑦}

$$d_{85}=\min\{d_{56},d_{57}\}=3$$

最后将⑤与⑧合并，得到类 9. 聚类过程完毕，按照聚类过程及并类距离画聚类谱系图。

2）最长距离法对这 5 位销售员进行分类：

合并距离最小的两类为新类，按顺序定为第⑥类，$d_{34}=1$ 为最小，⑥={3,4}

计算新类⑥与各当前类的距离，

$$d_{61}=\max\{d_{13},d_{14}\}=\sqrt{13}$$
$$d_{62}=\max\{d_{23},d_{24}\}=\sqrt{5}$$
$$d_{65}=\max\{d_{35},d_{45}\}=\sqrt{10}$$

得距离矩阵：

	②	⑤	⑥
①	$\sqrt{2}$	$\sqrt{29}$	$\sqrt{13}$
②		$\sqrt{17}$	$\sqrt{5}$
⑤			$\sqrt{10}$

合并距离最小的两类为新类，$d_{12}=\sqrt{2}$为最小，⑦={1,2}

$$d_{57}=\max\{d_{15},d_{25}\}=\sqrt{29}$$
$$d_{67}=\min\{d_{16},d_{26}\}=\sqrt{13}$$

得到距离矩阵：

	⑥	⑦
⑤	$\sqrt{10}$	$\sqrt{29}$
⑥		$\sqrt{13}$

合并距离最小的两类为新类，$d_{56}=\sqrt{10}$为最小，⑧={⑤，⑥}

$$d_{87}=\max\{d_{57},d_{67}\}=\sqrt{29}$$

最后将⑦与⑧合并，得到类 9. 聚类过程完毕，按照聚类过程及并类距离画聚类谱系图。

2.

1）合并距离最小的两类为新类，按顺序定为第⑥类，$d_{12}=5.2$ 为最小，⑥={1,2}

计算新类⑥与各当前类的距离，

$$d_{63}=\min\{d_{13},d_{23}\}=8.6$$
$$d_{64}=\min\{d_{14},d_{24}\}=27.4$$
$$d_{65}=\min\{d_{15},d_{25}\}=26.1$$

得到距离矩阵:

	④	⑤	⑥
③	7.5	15.1	8.6
④		6.5	27.4
⑤			26.1

合并距离最小的两类为新类，$d_{45}=6.5$ 为最小，⑦={4,5}

$$d_{37}=\min\{d_{34},d_{35}\}=7.5$$
$$d_{67}=\min\{d_{46},d_{56}\}=26.1$$

距离矩阵:

	⑥	⑦
③	8.6	7.5
⑥		26.1

$d_{37}=7.5$ 为最小，⑧={3,7}

$$d_{86}=\min\{d_{36},d_{67}\}=8.6$$

最后将⑥与⑧合并，得到类 9. 聚类过程完毕，

分类结果:　　第一类:①②　第二类:③④⑤

3.

最短距离法:

$$d_{G_1G_2}=min\{d_{13},d_{14},d_{15},d_{23},d_{24},d_{25}\}=2.45$$

最长距离法:

$$d_{G_1G_2}=max\{d_{13},d_{14},d_{15},d_{23},d_{24},d_{25}\}=8.42$$

重心法:

$$\overline{X}_{G_1}\left(\frac{5.54+3.12}{2},\ \frac{7.88+3.90}{2}\right)=(4.33,\ 5.89)$$

$$\overline{X}_{G_2}\left(\frac{1.96+1.18+1.65}{3},\ \frac{1.74+0.68+0.86}{3}\right)=(1.597,\ 1.093)$$

$$d_{G_1G_2}=\sqrt{(4.33-1.597)^2+(5.89-1.093)^2}=5.521$$

类平均法:

$$d^2_{G_1G_2}=\frac{1}{2\times3}(7.11^2+8.42^2+8.03^2+2.45^2+3.76^2+3.38^2)=36.249$$

所以类平均法的距离为 6.021

第四章 判别分析

1.

1) $\hat{\Sigma}=\frac{n_1S_1+n_2S_2}{n_1+n_2-2}=\begin{bmatrix}9.118 & 4.194\\ 4.194 & 3.979\end{bmatrix}$

2) $\hat{\Sigma}^{-1}=\begin{bmatrix}0.213 & -0.224\\ -0.224 & 0.488\end{bmatrix}$

$$y(x)=\left(\overline{x}^{(1)}-\overline{x}^{(2)}\right)'\hat{\Sigma}^{-1}\left(x-\frac{\overline{x}^{(1)}+\overline{x}^{(2)}}{2}\right)$$

$$=-1.86x_1+2.426x_2+28.576$$

3) $y(x)=-1.86\times20.12+2.426\times5.11+28.576=3.5295>0$

因此这个样品属于第一类

2.得到线性判别函数

$$f_1(x) = -28.70x_1 + 28.70x_2 + 37.50x_3 + 34.30x_4 + 67.80x_5 - 280.66$$
$$f_2(x) = -9.70x_1 + 3.91x_2 + 27.20x_3 + 27.30x_4 + 19.20x_5 - 47.56$$
$$f_3(x) = -4.91x_1 + 3.91x_2 + 27.20x_3 + 27.30x_4 + 19.20x_5 - 32.94$$
$$f_l(x) = \max_{1\le i\le 3} f_i(x), \text{则} x \epsilon G_l$$

第五章 主成分分析

1.

1)求特征值及其对应的单位特征向量

$$|S - \lambda I| = 0$$

$$\begin{vmatrix} 90-\lambda & 48 \\ 48 & 45-\lambda \end{vmatrix} = 0$$

$$(90-\lambda)(45-\lambda) - 48^2 = 0 \Longrightarrow \lambda^2 - 135\lambda + 1746 = 0$$

计算得到 S 的两个特征根：λ_1=120.51， λ_2=14.49

λ_1的单位特征向量: $(S-\lambda_1 I)\alpha_1 = 0 \qquad \alpha_1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$

$$\begin{pmatrix} 90-120.51 & 48 \\ 48 & 45-120.51 \end{pmatrix}\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = 0 \quad \text{且} {a_{11}}^2 + {a_{21}}^2 = 1$$

$$\Longrightarrow \quad \alpha_1 = \begin{pmatrix} 0.8439 \\ 0.5364 \end{pmatrix}$$

λ_2的单位特征向量：

$$(S-\lambda_2 I)\alpha_1 = 0 \qquad \alpha_1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$$

$$\begin{pmatrix} 90-14.49 & 48 \\ 48 & 45-14.49 \end{pmatrix}\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = 0 \quad \text{且} {a_{12}}^2 + {a_{22}}^2 = 1$$

$$\Longrightarrow \quad \alpha_2 = \begin{pmatrix} -0.5364 \\ 0.8439 \end{pmatrix}$$

2)主成分表达式:

第一主成分 $y_1 = 0.8439(x_1 - 134) + 0.5364(x_2 - 92)$

第二主成分 $y_2 = -0.5364(x_1 - 134) + 0.8439(x_2 - 92)$

3)第一主成分方差贡献率:$\frac{\lambda_1}{\lambda_1+\lambda_2}$=89.27%

第二主成分方差贡献率: $\frac{\lambda_2}{\lambda_1+\lambda_2}$=10.73%

4)解释主成分的实际意义

在y_1的表达式中, x_1与x_2前面的系数都为正，因此可以表示 x_1与x_2的加权和，表示叶片的面积大小， x_1长度的权重为 0.84, x_2宽度的权重 0.54，说明y_1表示叶片面积大小，其中长度对面积的影响程度更大。当 x_1与x_2都较大时，主成分y_1的得分较高，表明该叶片面积较大.

在y_2的表达式中, x叶片长度前面系数为负， x_2叶片宽度前面的系数为正，因此y_2表示叶片形态，当某叶片长度偏长、宽度偏窄时， y_2得分较小，表明该叶子越长，当y_2接近

于 0，则该叶子形态较匀称.

5)

$(x_1, x_2) = (146,94)$

$y_1 = 11.1996$，$y_2 = -4.749$ ，该叶片由于x_1与x_2都大于平均数，且y_1较大，说明该叶片面积较大，同时y_2略小于 0，说明该叶片形态偏细长.

6)协方差矩阵

$$Var(Y) = \begin{bmatrix} 120.51 & 0 \\ 0 & 14.49 \end{bmatrix}$$

7)主成分与每一个原变量的相关系数与方差贡献

第一主成分$r_{1Y_1} = a_{11}\sqrt{\frac{\lambda_1}{S_{11}}} = 0.8439 \times \sqrt{\frac{120.51}{90}}$=0.9765

$r_{2Y_1} = a_{21}\sqrt{\frac{\lambda_1}{S_{22}}} = 0.5364 \times \sqrt{\frac{120.51}{45}}$=0.8778

第二主成分$r_{1Y_2} = a_{21}\sqrt{\frac{\lambda_2}{S_{22}}} = -0.5364 \times \sqrt{\frac{14.49}{90}}$=0.9765

$r_{2Y_2} = a_{22}\sqrt{\frac{\lambda_2}{S_{22}}} = 0.8439 \times \sqrt{\frac{14.49}{45}}$=0.4789

y_1对x_1的方差贡献，${r_{1Y_1}}^2 = {a_{11}}^2\frac{\lambda_1}{S_{11}} = 0.9536$

y_1对x_2的方差贡献，${r_{2Y_1}}^2 = {a_{21}}^2\frac{\lambda_1}{S_{22}} = 0.7705$

y_2对x_1的方差贡献，${r_{1Y_2}}^2 = {a_{12}}^2\frac{\lambda_2}{S_{11}} = 0.0463$

y_2对x_2的方差贡献，${r_{2Y_1}}^2 = {a_{22}}^2\frac{\lambda_2}{S_{22}} = 0.2293$

2.

$\overline{\boldsymbol{x}} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \end{bmatrix} = \begin{bmatrix} 156.67 \\ 47.58 \end{bmatrix}$ $\quad \boldsymbol{S} = \begin{bmatrix} 7.697 & 3.361 \\ 3.361 & 3.538 \end{bmatrix}$

$$|S - \lambda I| = 0$$

$\begin{vmatrix} 7.697 - \lambda & 3.361 \\ 3.361 & 3.538 - \lambda \end{vmatrix}$=0

$$(7.697 - \lambda)(3.538 - \lambda) - 3.361^2 = 0 \Longrightarrow \lambda^2 - 11.235\lambda + 15.934 = 0$$

计算得到 S 的两个特征根：λ_1=9.5699 , λ_2=1.6651

λ_1的单位特征向量： $(S - \lambda_1 I)\alpha_1 = 0$ $\quad \alpha_1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$

$$\begin{pmatrix} 7.697 - 9.5699 & 3.361 \\ 3.361 & 3.538 - 9.5699 \end{pmatrix}\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}=0 \quad 且{a_{11}}^2 + {a_{21}}^2 = 1$$

$$\Longrightarrow \quad \alpha_1 = \begin{pmatrix} 0.874 \\ 0.487 \end{pmatrix}$$

$(S - \lambda_2 I)\alpha_1 = 0$ $\quad \alpha_1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$

λ_2的单位特征向量：

$\begin{pmatrix} 7.697-1.6651 & 3.361 \\ 3.361 & 3.538-1.6651 \end{pmatrix}\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$=0 且$a_{12}{}^2+a_{22}{}^2=1$

$$\Rightarrow \quad \alpha_2=\begin{pmatrix} -0.487 \\ 0.874 \end{pmatrix}$$

第一主成分 $y_1=0.874(x_1-156.67)+0.487(x_2-47.58)$

第二主成分 $y_2=-0.487(x_1-156.67)+0.874(x_2-47.58)$

第一主成分方差贡献率：$\frac{\lambda_1}{\lambda_1+\lambda_2}$=85.2%

第二主成分方差贡献率：$\frac{\lambda_2}{\lambda_1+\lambda_2}$=14.8%

根据第一主成分的方差贡献率，主成分降维的效果较好，近一个主成分就反映了原两个变量85.2%的信息.

在y_1的表达式中，x_1与x_2前面的系数都是正的，因此可以表示x_1与x_2的加权和，表示学生的体型，当某一位同学的身高与体重都比较大时,y_1的得分较高，将 12 位同学的数据代入,得出y_1并进行排序，比较学生体型胖瘦的排序.

在y_2的表达式中，x_1身高前面系数为负，x_2体重前面的系数为正，因此y_2表示学生的体型特征，当某位同学身高较高，体重较小，则y_2得分低表明该同学身材呈细长，反之则为矮胖. 当y_2接近于 0 时,表明该同学身材比较匀称.

第六章 因子分析

1.

1)

$$\begin{cases} X_1=0.857F_1-0.011F_2+0.205F_3+\varepsilon_1 \\ X_2=0.841F_1+0.321F_2-0.102F_3+\varepsilon_2 \\ X_3=0.847F_1-0.120F_2+0.323F_3+\varepsilon_3 \\ X_4=0.901F_1+0.281F_2-0.027F_3+\varepsilon_4 \\ X_5=0.899F_1+0.215F_2-0.019F_3+\varepsilon_5 \\ X_6=-0.313F_1+0.839F_2+0.305F_3+\varepsilon_6 \\ X_7=-0.666F_1+0.062F_2+0.679F_3+\varepsilon_7 \\ X_8=0.575F_1-0.580F_2+0.367F_3+\varepsilon_8 \end{cases}$$

2)

$$h_1{}^2=0.857^2+0.011^2+0.205^2=0.777$$
$$h_2{}^2=0.841^2+0.321^2+0.102^2=0.821$$
$$h_3{}^2=0.847^2+0.120^2+0.323^2=0.836$$
$$h_4{}^2=0.901^2+0.281^2+0.027^2=0.891$$
$$h_5{}^2=0.899^2+0.215^2+0.019^2=0.855$$
$$h_6{}^2=0.313^2+0.839^2+0.305^2=0.895$$
$$h_7{}^2=0.666^2+0.062^2+0.679^2=0.908$$
$$h_8{}^2=0.575^2+0.580^2+0.367^2=0.802$$

$h_i{}^2$表示 3 个公共因子共同反映x_i信息的多少。

这 8 个变量 $h_i{}^2$共同度都较大，说明因子分析效果比较好.

3)

$$g_1{}^2=0.857^2+0.841^2+0.847^2+0.901^2+0.899^2+0.313^2+0.666^2+0.575^2=4.65$$

$$g_2{}^2 = 0.011^2 + 0.321^2 + 0.120^2 + 0.281^2 + 0.215^2 + 0.839^2 + 0.062^2 + 0.580^2 = 1.29$$
$$g_3{}^2 = 0.205^2 + 0.102^2 + 0.323^2 + 0.027^2 + 0.019^2 + 0.305^2 + 0.679^2 + 0.367^2 = 0.85$$

2.

1)

$$\begin{cases} X_1 = 0.321F_1 + 0.123F_2 + 0.916F_3 + \varepsilon_1 \\ X_2 = 0.088F_1 + 0.064F_2 + 0.926F_3 + \varepsilon_2 \\ X_3 = 0.165F_1 + 0.320F_2 + 0.833F_3 + \varepsilon_3 \\ X_4 = -0.030F_1 + 0.906F_2 + 0.184F_3 + \varepsilon_4 \\ X_5 = 0.231F_1 + 0.914F_2 + 0.155 + \varepsilon_5 \\ X_6 = 0.145F_1 + 0.952F_2 + 0.105F_3 + \varepsilon_6 \\ X_7 = 0.944F_1 + 0.112 + 0.212F_3 + \varepsilon_7 \\ X_8 = 0.963F_1 - 0.078F_2 + 0.209F_3 + \varepsilon_8 \\ X_9 = 0.976F_1 + 0.144F_2 + 0.115F_3 + \varepsilon_9 \end{cases}$$

2)

由于 $X_7\ X_8 X_9$在F_1上的因子载荷值较大，所以F_1可命名为企业经营能力因子

由于 $X_4\ X_5 X_6$在F_1上的因子载荷较大，所以F_2可命名为盈利能力因子

由于$X_1\ X_2 X_3$在F_1上的因子载荷较大，所以F_3可命名为负债偿还能力因子

3)

$$h_1{}^2 = 0.321^2 + 0.123^2 + 0.916^2 = 0.957$$
$$h_2{}^2 = 0.088^2 + 0.064^2 + 0.926^2 = 0.937$$

$h_i{}^2$表示 3 个公共因子共同反映x_i信息的多少，指这三个公共因子对变量x_i的方差所做的贡献，$h_1{}^2$和$h_2{}^2$这 2 个变量共同度都较大，说明因子分析效果比较好.