2018 年全国大学生数学建模竞赛暨美赛培训
聚类分析与判别分析

厦门大学2016 级各学院

数学建模团队：谭忠教授 ；助教：陈小伟，姜小蒙，姚瑶，余娇妍

要求：（1）必须用TEX输入编辑后将$TEXPDF$以及图表一并发邮件提交给$ztan85@163.com$及$sxjm004@163.com$，压缩包及邮件主题名为“编号+姓名+专业+第*次作业”；

（2）必须抄题，以免判错。

1. 5位代理商对某种产品的四种指标评分如下：

	x_1	x_2	x_3	x_4
1	2	4	6	32
2	5	2	5	38
3	3	3	7	30
4	1	2	3	16
5	4	3	2	30

其中，x_1, x_2, x_3为态度测度，共有1 7的分值，x_4为兴趣测度，取值为11 40. 求出其绝对值距离矩阵，平方和距离矩阵.

2. 检测某类产品的重量，抽了六个样品，每个样品只测了一个指标，分别为1，2，3，6，9，11.试用最短距离法，重心法进行聚类分析.

3. 某店五个售货员的销售量x_1与教育水平x_2之间的评分表如下，试用最短距离法作聚类分析

	x_1	x_2
1	1	1
2	1	2
3	6	3
4	8	2
5	8	0

4. 下面给出七个样品两两之间的欧式距离矩阵

$$D=\begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0 & & & & & & \\ 2 & 4 & 0 & & & & & \\ 3 & 7 & 3 & 0 & & & & \\ 4 & 12 & 8 & 5 & 0 & & & \\ 5 & 18 & 14 & 11 & 6 & 0 & & \\ 6 & 19 & 15 & 12 & 7 & 1 & 0 & \\ 7 & 21 & 17 & 14 & 9 & 3 & 2 & 0 \end{pmatrix}$$

试分别用最小距离法，最大距离法，重心距离法进行聚类，并画出谱系图.

5. 华北五站（北京、天津、营口、太原、石家庄）1968年

（及1969年）7 8月份降水量(Y)作预报.

(1) 根据专业的统计分析Y主要取决于下列因子：

X_1：上海4月份平均气温，

X_2：北京三月份降水总量，

X_3：5月份地磁Ci指数，

X_4：4月份500mbW环流型日数

(2) 1961-1967年的历史数据如下：

时间	Y/mm	x_1	x_2	x_3	x_4
1961	410	14.8	20.1	0.69	13
1962	255	12.5	2.3	0.36	4
1963	527	14.5	12.4	0.69	12
1964	510	16.4	10.6	0.58	26
1965	226	12.2	0.3	0.35	4
1966	456	13.8	12.3	0.42	23
1967	389	13.6	7.7	0.82	25
1968		13.7	0.6	0.68	12.5
1969		14.2	16.5	0.65	15

6. 下表是15个上市公司2001年的一些主要财务指标，使用系统聚类法和K－均值法分别对这些公司进行聚类，并对结果进行比较分析. 其中，x_1：公司编号，x_2：净资产收益率，x_3：每股净利润，x_4：总资产周转率，x_5：资产负债率，x_6：流动负债比率，x_7：每股净资产，x_8：净利润增长率，x_9：总资产增长率

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	11.09	0.21	0.05	96.98	70.53	1.86	-44.04	81.99
2	11.96	0.59	0.74	51.78	90.73	4.95	7.02	16.11
3	0	0.03	0.03	181.99	100	-2.98	103.33	21.18
4	11.58	0.13	0.17	46.07	92.18	1.14	6.55	-56.32
5	-6.19	-0.09	0.03	43.3	82.24	1.52	-1713.5	-3.36
6	10	0.47	0.48	68.4	86	4.7	-11.56	0.85
7	10.49	0.11	0.35	82.98	99.87	1.02	100.23	30.32
8	11.12	-1.69	0.12	132.14	100	-0.66	-4454.39	-62.75
9	3.41	0.04	0.2	67.86	98.51	1.25	-11.25	-11.43
10	1.16	0.01	0.54	43.7	100	1.03	-87.18	-7.41
11	30.22	0.16	0.4	87.36	94.88	0.53	729.41	-9.97
12	8.19	0.22	0.38	30.31	100	2.73	-12.31	-2.77
13	95.79	-5.2	0.5	252.34	99.34	-5.42	-9816.52	-46.82
14	16.55	0.35	0.93	72.31	84.05	2.14	115.95	123.41
15	-24.18	-1.16	0.79	56.26	97.8	4.81	-533.89	-27.74

7. 下表是某年我国16个地区农民支出情况的抽样调查数据，每

个地区调查了反映每人平均生活消费支出情况的六个经济指标. 试通过统计分析软件用不同的方法进行聚类分析，并比较何种方法与人们观察到的实际情况较接近.

地区	食品	衣着	燃料	住房	交通和通讯	娱乐教育文化
北京	190.33	43.77	9.73	60.54	49.01	9.04
天津	135.2	36.4	10.47	44.16	36.49	3.94
河北	95.21	22.83	9.3	22.44	22.81	2.8
山西	104.78	25.11	6.4	9.89	18.17	3.25
内蒙	128.41	27.63	8.94	12.58	23.99	2.27
辽宁	145.68	32.83	17.79	27.29	39.09	3.47
吉林	159.37	33.38	18.37	11.81	25.29	5.22
黑龙江	116.22	29.57	13.24	13.76	21.75	6.04
上海	221.11	38.64	12.53	115.65	50.82	5.89
江苏	144.98	29.12	11.67	42.6	27.3	5.74
浙江	169.92	32.75	12.72	47.12	34.35	5
安徽	135.11	23.09	15.62	23.54	18.18	6.39
福建	144.92	21.26	16.96	19.52	21.75	6.73
江西	140.54	21.5	17.64	19.19	15.97	4.94
山东	115.84	30.26	12.2	33.6	33.77	3.85
河南	101.18	23.26	8.46	20.2	20.5	4.3

8. 设三个总体的分布分别为：G1为$N(2,0.5^2)$，G2为$N(0,2^2)$，G3为$N(3,1^2)$. 试问样品$x=2.5$应判归哪一类?(1) 按距离准则；(2) 按Bayes准则($q_1=q_2=q_3=\frac{1}{3}, L(j|i)=\begin{cases}1 & i\neq j\\ 0 & i=j\end{cases}$)

9. 设有两个正态总体G1和G2，已知($m=2$)，

$$\mu_1=\begin{pmatrix}10\\15\end{pmatrix},\mu_2=\begin{pmatrix}20\\25\end{pmatrix},\Sigma_1=\begin{pmatrix}18&12\\12&32\end{pmatrix},\Sigma_2=\begin{pmatrix}20&-7\\-7&5\end{pmatrix}$$

先验概率$q_1=q_2$，而$L(2|1)=10$，$L(1|2)=75$，试问样品$x_{(1)}=\begin{pmatrix}20\\20\end{pmatrix},x_{(2)}=\begin{pmatrix}15\\20\end{pmatrix}$各应归判哪一类？

10. 某公司为掌握其新产品的动向，向12个代理商做调查，要他们对产品给予评估(对产品式样、包装及耐久性，用10 分制打分，高分表示性能良好，低分则较差)并说明是否购买. 调查结果如下，试做fisher判别

		式样	包装	耐久性
购买组样品	1	9	8	7
	2	7	6	6
	3	10	7	8
	4	8	4	5
	5	9	9	7
	6	8	6	7
	7	7	5	6
非购买组样品	1	4	4	4
	2	3	6	6
	3	6	3	3
	4	2	4	5
	5	1	2	2

11. 某超市经销十种品牌的饮料，其中有四种畅销，三种滞销，三种平销. 下表是这十种品牌饮料的销售价格（元）和顾客对各种饮料的口味评分、信任度评分的平均数.

销售情况	产品序号	销售价格	口味评分	信任度评分
畅销	1	2.2	5	8
	2	2.5	6	7
	3	3.0	3	9
	4	3.2	8	6
平销	5	2.8	7	6
	6	3.5	8	7
	7	4.8	9	8
滞销	8	1.7	3	4
	9	2.2	4	2
	10	2.7	4	3

(1) 根据数据建立贝叶斯判别函数，并根据此判别函数对原样本进行回判.

(2) 现有一新品牌的饮料在该超市试销，其销售价格为3.0，顾客对其口味的评分平均为8，信任评分平均为5，试预测该饮料的销售情况.

12. 银行的贷款部门需要判别每个客户的信用好坏（是否未履行还贷责任），以决定是否给予贷款. 可以根据贷款申请人的年龄（X_1）、受教育程度（X_2）、现在所从事工作的年数（X_3）、未变更住址的年数（X_4）、收入（X_5）、负债收入比例（X_6）、信用卡债务

（X_7）、其它债务（X_8）等来判断其信用情况. 下表是从某银行的客户资料中抽取的部分数据，(1) 根据样本资料分别用距离判别法、Bayes判别法和Fisher 判别法建立判别函数和判别规则. (2) 某客户的如上情况资料为（53，1，9，18，50，11.20，2.02，3.58），对其进行信用好坏的判别.

目前信用好坏	客户序号	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
已履行还贷责任	1	23	1	7	2	31	6.60	0.34	1.71
	2	34	1	17	3	59	8.00	1.81	2.91
	3	42	2	7	23	41	4.60	0.94	0.94
	4	39	1	19	5	48	13.10	1.93	4.36
	5	35	1	9	1	34	5.00	0.40	1.30
未履行还贷责任	6	37	1	1	3	24	15.10	1.80	1.82
	7	29	1	13	1	42	7.40	1.46	1.65
	8	32	2	11	6	75	23.30	7.76	9.72
	9	28	2	2	3	23	6.40	0.19	1.29
	10	26	1	4	3	27	10.50	2.47	0.36

13. 人文发展指数是联合国开发计划署于1990 年5 月发表的第一份《人类发展报告》中公布的. 该报告建议，目前对人文发展的衡量应当以人生的三大要素为重点，衡量人生三大要素的指示指标分别要用出生时的预期寿命、成人识字率和实际人均GDP，将以上三个指示指标的数值合成为一个复合指数，即为人文发展指数. 资料来源：UNDP《人类发展报告》1995 年. 今从1995 年世界各国人文发展指数的排序中，选取高发展水平、中等发展水平的国家各五个作为两组样品，另选四个国家作为待判样品作距离判别分析.

类别	序号	国家名称	出生时预期寿命	成人识字率	人均GDP
第一类	1	美国	76	99	5374
（高发展	2	日本	79.5	99	5359
水平国	3	瑞士	78	99	5372
家）	4	阿根廷	72.1	95.9	5242
	5	阿联酋	73.8	77.7	5370
第二类	6	保加利亚	71.2	93	4250
（中等发	7	古巴	75.3	94.9	3412
展水平	8	巴拉圭	70	91.2	3390
国家）	9	格鲁吉亚	72.8	99	2300
	10	南非	62.9	80.6	3799
	11	中国	68.5	79.3	1950
待判	12	罗马尼亚	69.9	96.9	2840
样品	13	希腊	77.6	93.8	5233
	14	哥伦比亚	69.3	90.3	5158

14. 根据经验，今天与昨天的温度差X_1及今天的压温差（压强与温度之差）X_2是预报明天下雨或不下的两个重要因素. 现有一批已收集的数据资料如下表. 今测得$x_1 = 8.1, x_2 = 2.0$，试问预报明天下雨还是不下？

雨	天	非雨	天
X_1	X_2	X_1	X_2
-1.9	3.2	0.2	0.2
-6.9	10.4	-0.1	7.5
5.2	2.0	0.4	14.6
5.0	2.5	2.7	8.3
7.3	0.0	2.1	0.8
6.8	12.7	-4.6	4.3
0.9	-15.4	-1.7	10.9
-12.5	-2.5	-2.6	13.1
1.5	1.3	2.6	12.8
3.8	6.8	-2.8	10.0

15. 为了更深入地了解我国人口的文化程度状况，现利用1990年全国人口普查数据对全国30个省、直辖市、自治区进行聚类分析，原始数据如下表. 分析选用了三个指标：大学以上文化程度的人口比例（DXBZ)、初中文化程度的人口比例（CZBZ)、文盲半文盲人口比例（WMBZ）来反映较高、中等、较低的文化程度人口的状况.

（1）计算样本的Euclid距离，分别用最长距离法、均值法、重心法和Ward 法作聚类分析，并画出相应的谱系图. 如果将所有样本分为四类，试写出各种方法的分类结果.

（2）用动态规划方法分4类，写出相应的分类结果.

地区	DXBZ	CZBZ	WMBZ	地区	DXBZ	CZBZ	WMBZ
北京	9.30	30.55	8.70	河南	0.85	26.55	16.15
天津	4.67	29.38	8.92	湖北	1.57	23.16	15.79
河北	0.96	24.69	15.21	湖南	1.14	22.57	12.10
山西	1.38	29.24	11.30	广东	1.34	23.04	10.45
内蒙古	1.48	25.47	15.39	广西	0.79	19.14	10.61
辽宁	2.60	32.32	8.81	海南	1.24	22.53	13.97
吉林	2.15	26.31	10.49	四川	0.96	21.65	16.24
黑龙江	2.14	28.46	10.87	贵州	0.78	14.65	24.27
上海	6.53	31.59	11.04	云南	0.81	13.85	25.44
江苏	1.47	26.43	17.23	西藏	0.57	3.85	44.43
浙江	1.17	23.74	17.46	陕西	1.67	24.36	17.62
安徽	0.88	19.97	24.43	甘肃	1.10	16.85	27.93
福建	1.23	16.87	15.63	青海	1.49	17.76	27.70
江西	0.99	18.84	16.22	宁夏	1.61	20.27	22.06
山东	0.98	25.18	16.87	新疆	1.85	20.66	12.75