

MATLAB统计分析与应用

——判别分析

主讲人：谢中华

SSstudy.com

科学软件学习网

主要内容

- 距离判别
- 贝叶斯判别
- Fisher判别

第一节 距离判别

一、基本思想

距离判别的最直观的想法是计算样品到第 i 类总体的平均数的距离，哪个距离最小就将它判归哪个总体，所以，我们首先考虑的是是否能够构造一个恰当的距离函数，通过样本与某类别之间距离的大小，判别其所属类别。

二、距离判别的MATLAB函数

➤ classify函数

调用格式：

class = classify(sample,training,group)

class = classify(sample,training,group,type)

class = classify(sample,training,group,type,prior)

[class,err] = classify(...)

[class,err,POSTERIOR] = classify(...)

[class,err,POSTERIOR,logp] = classify(...)

[class,err,POSTERIOR,logp,coeff] = classify(...)

三、距离判别案例

【例14.1-1】中小企业的破产模型

对21个破产的企业收集它们在破产前两年的年度财务数据，同时对25个财务良好的企业也收集同一时期的数据。

数据涉及4个变量：

x_1 = 现金流量/总债务，

x_2 = 净收入/总资产，

x_3 = 流动资产/流动债务，

x_4 = 流动资产/净销售额。

数据列于下表，1组为破产企业，2组为非破产企业。现有4个未判企业，它们的相关数据列于表中最后4行，试根据距离判别法，对这4个未判企业进行判别。

企业编号	组 别	x1	x2	x3	x4	
1	1	-0.45		-0.41	1.09	0.45
2	1	-0.56		-0.31	1.51	0.16
3	1	0.06		0.02	1.01	0.40
4	1	-0.07		-0.09	1.45	0.26
5	1	-0.10		-0.09	1.56	0.67
6	1	-0.14		-0.07	0.71	0.28
7	1	0.04		0.01	1.50	0.71
8	1	-0.07		-0.06	1.37	0.40
9	1	0.07		-0.01	1.37	0.34
10	1	-0.14		-0.14	1.42	0.43
11	1	-0.23		-0.30	0.33	0.18
12	1	0.07		0.02	1.31	0.25
13	1	0.01		0.00	2.15	0.70
14	1	-0.28		-0.23	1.19	0.66
15	1	0.15		0.05	1.88	0.27
16	1	0.37		0.11	1.99	0.38
17	1	-0.08		-0.08	1.51	0.42
18	1	0.05		0.03	1.68	0.95
19	1	0.01		0.00	1.26	0.60
20	1	0.12		0.11	1.14	0.17
21	1	-0.28		-0.27	1.27	0.51
22	2	0.51		0.10	2.49	0.54
23	2	0.08		0.02	2.01	0.53
24	2	0.38		0.11	3.27	0.35
25	2	0.19		0.05	2.25	0.33
26	2	0.32		0.07	4.24	0.63
27	2	0.31		0.05	4.45	0.69
28	2	0.12		0.05	2.52	0.69
29	2	-0.02		0.02	2.05	0.35
30	2	0.22		0.08	2.35	0.40
31	2	0.17		0.07	1.80	0.52
32	2	0.15		0.05	2.17	0.55
33	2	-0.10		-0.01	2.50	0.58
34	2	0.14		-0.03	0.46	0.26
35	2	0.14		0.07	2.61	0.52
36	2	0.15		0.06	2.23	0.56
37	2	0.16		0.05	2.31	0.20

.....

MATLAB程序

```
sampladata = xlsread('classify.xls','C2:F51');
```

```
training = xlsread('classify.xls','C2:F47');
```

```
group = xlsread('classify.xls','B2:B47');
```

```
obs = [1 : 50];
```

```
% 距离判别 , 返回判别结果向量C和误判概率err
```

```
[C,err] = classify(sampladata,training,group,'mahalanobis');
```

```
headline = {'观测序号','所在组别','判归组别'};
```

```
group = [group;NaN;NaN;NaN;NaN];
```

```
[headline;num2cell([obs,group,C])] % 查看判别结果
```

```
err % 查看误判概率
```

【例14.1-2】 手写体数字识别

随着计算机技术的高速发展，利用计算机进行模式识别得以实现，例如指纹识别、数字识别等。本例借助Matlab编程实现手写体数字（0-9）的识别，识别原理为距离判别。

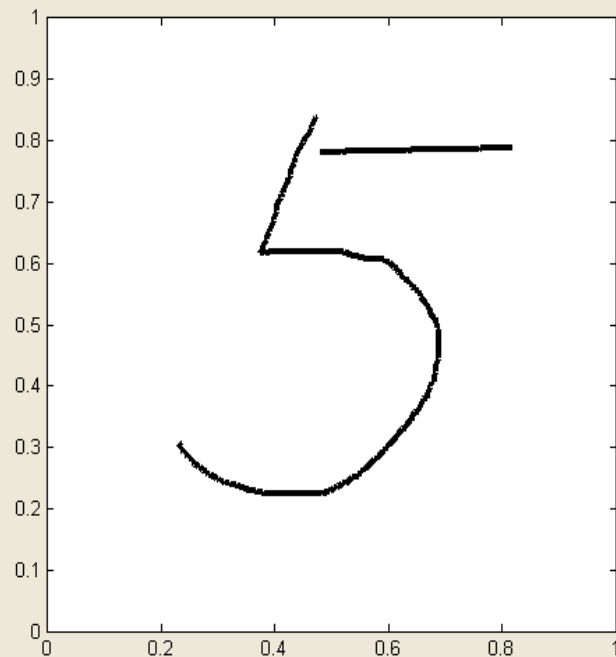
每个数字选取3张图片作为样本，如图



重新做一张0-9数字的手写图片，编程作自动识别。

手写体数字识别

帮助 关于



输出识别结果

5

导入样本图片

选择要识别的图片

☒ 选择手绘图片

☐ 选择外部图片

选择绘图颜色

黑色

开始识别

清除

退出

第二节 贝叶斯判别

一、基本思想

距离判别没有考虑人们对研究对象已有的认知，而这种已有的认知可能会对判别的结果产生影响。贝叶斯（ Bayes ）判别则用一个先验概率来描述这种已有的认知，然后通过样本来修正先验概率，得到后验概率，最后基于后验概率进行判别。

二、贝叶斯判别的MATLAB函数

- NaiveBayes : 创建朴素贝叶斯分类器
- fit : 根据训练样本创建朴素贝叶斯分类器
- predict : 对待判样品进行分类

三、贝叶斯判别案例

【例14.2-1】蠓虫的分类识别问题

生物学家试图对两种蠓虫（Af与Apf）进行鉴别，依据的资料是触角和翅膀的长度，已经测得9只Af和6只Apf的数据如下：

p_1 : 9只Af

触角长：1.24 1.36 1.38 1.38 1.38 1.40 1.48 1.54 1.56

翅膀长：1.72 1.74 1.64 1.82 1.90 1.70 1.82 1.82 2.08

p_2 : 6只Apf

触角长：1.14 1.18 1.20 1.26 1.28 1.30

翅膀长：1.78 1.96 1.86 2.00 2.00 1.96

问题：

- (1) 如何凭借原始资料(15 对数据,称之为样本)制定一种方法,正确区分两类蠓虫;
- (2) 依据确立的方法,对以下三个样本：
 $(1.24, 1.80)$, $(1.28, 1.84)$, $(1.40, 2.04)$ 加以识别

蠓虫观测数据

观测序号	触角长	翅膀长	原属类别
1	1.24	1.72	Af
2	1.36	1.74	Af
3	1.38	1.64	Af
4	1.38	1.82	Af
5	1.38	1.9	Af
6	1.4	1.7	Af
7	1.48	1.82	Af
8	1.54	1.82	Af
9	1.56	2.08	Af
10	1.14	1.78	Apf
11	1.18	1.96	Apf
12	1.2	1.86	Apf
13	1.26	2	Apf
14	1.28	2	Apf
15	1.3	1.96	Apf
16	1.24	1.8	unknown
17	1.28	1.84	unknown
18	1.4	2.04	unknown

MATLAB程序

```
[xdata,textdata] = xlsread('NaiveBayes.xls');
```

```
training = xdata(1:15,2:3);
```

```
group = textdata(2:16,4);
```

```
ObjBayes = NaiveBayes.fit(training, group);
```

```
sampladata = xdata(:,2:3);
```

```
% 利用所创建的朴素贝叶斯分类器对象对训练样本进行判别
```

```
pre_group = ObjBayes.predict(sampladata);
```

```
headline = {'观测序号','触角长','翅膀长','原属组别','判归组别'};
```

```
group = textdata(2:end,4);
```

```
[headline;num2cell(xdata),group,pre_group] % 查看判别结果
```

第四节 Fisher判别（典型判别）

一、基本思想

先降维，用 p 维向量 $\mathbf{x} = (x_1, x_2, \dots, x_p)'$ 的少数几个线性组合（称为线性判别式或典型变量） $y_1 = \mathbf{a}'_1 \mathbf{x}, y_2 = \mathbf{a}'_2 \mathbf{x}, \dots, y_r = \mathbf{a}'_r \mathbf{x}$,
（一般 r 明显小于 p ）来代替原始的 p 个变量，
然后作距离判别。

二、Fisher判别的MATLAB函数

1. 自编fisher函数

调用格式：

`class = fisher(sampledata,training,group)`

`class = fisher(sampledata,training,group,contri)`

`[class, TabCan] = fisher(...)`

`[class, TabCan, TabL] = fisher(...)`

`[class, TabCan, TabL, TabCon] = fisher(...)`

`[class, TabCan, TabL, TabCon, TabM] = fisher(...)`

`[class, TabCan, TabL, TabCon, TabM, TabG] = fisher(...)`

`[class, TabCan, TabL, TabCon, TabM, TabG, trainscore] = fisher(...)`

三、Fisher判别案例

【例14.3-1】 例14.2-1续 蠓虫的分类识别问题

```
[xdata,textdata] = xlsread('NaiveBayes.xls');  
training = xdata(1:15,2:3);  
group = textdata(2:16,4);  
sampledata = xdata(:,2:3);  
% 调用自编fisher函数进行判别  
pre_group = fisher(sampledata,training,group);  
headline = {'观测序号','触角长','翅膀长','原属组别','判归组别'};  
group = textdata(2:end,4);  
[headline,num2cell(xdata),group,pre_group] % 查看判别结果
```

【例14.3-2】 Fisher于1936年发表的鸢尾花

(Iris) 数据被广泛的作为判别分析的例子。数据是对3种鸢尾花：刚毛鸢尾花 (setosa类)、变色鸢尾花 (versicolor类) 和弗吉尼亚鸢尾花

(virginica类) 各抽取一个容量为50的样本，测量其花萼长 x_1 、花萼宽 x_2 、花瓣长 x_3 、花瓣宽 x_4 ，单位为cm。数据保存在MATLAB中统计工具箱文件夹下的文件fisheriris.mat中。现有10个未知类别的鸢尾花数据，如下表所示。试把文件fisheriris.mat中的数据作为训练样本，利用fisher判别对表中待判样品进行判别。

待判鸢尾花数据

观测序号	花萼长 x_1	花萼宽 x_2	花瓣长 x_3	花瓣宽 x_4
1	5.8	2.7	1.8	0.73
2	5.6	3.1	3.8	1.8
3	6.1	2.5	4.7	1.1
4	6.1	2.6	5.7	1.9
5	5.1	3.1	6.5	0.62
6	5.8	3.7	3.9	0.13
7	5.7	2.7	1.1	0.12
8	6.4	3.2	2.4	1.6
9	6.7	3	1.9	1.1
10	6.8	3.5	7.9	1

MATLAB程序

load fisheriris

```
x = [5.8 2.7 1.8 0.73  
      5.6 3.1 3.8 1.8  
      6.1 2.5 4.7 1.1  
      6.1 2.6 5.7 1.9  
      5.1 3.1 6.5 0.62  
      5.8 3.7 3.9 0.13  
      5.7 2.7 1.1 0.12  
      6.4 3.2 2.4 1.6  
      6.7 3 1.9 1.1  
      6.8 3.5 7.9 1 ];
```

MATLAB程序

% 利用fisher函数进行判别，返回各种结果，其中ts为判别式得分

```
[outclass,TabCan,TabL,TabCon,TabM,TabG,ts] = fisher(x,meas,species);
```

```
ts1 = ts(ts(:,1) == 1,:); % setosa类的判别式得分
```

```
ts2 = ts(ts(:,1) == 2,:); % versicolor类的判别式得分
```

```
ts3 = ts(ts(:,1) == 3,:); % virginica类的判别式得分
```

```
plot(ts1(:,2),ts1(:,3),'o') % setosa类的判别式得分的散点图
```

```
hold on
```

```
plot(ts2(:,2),ts2(:,3), 'r*') % versicolor类的判别式得分的散点图
```

```
plot(ts3(:,2),ts3(:,3), 'gp') % virginica类的判别式得分的散点图
```

```
legend('setosa类','versicolor类','virginica类'); %加标注框
```

```
xlabel('第一判别式得分'); %给X轴加标签
```

```
ylabel('第二判别式得分'); %给Y轴加标签
```

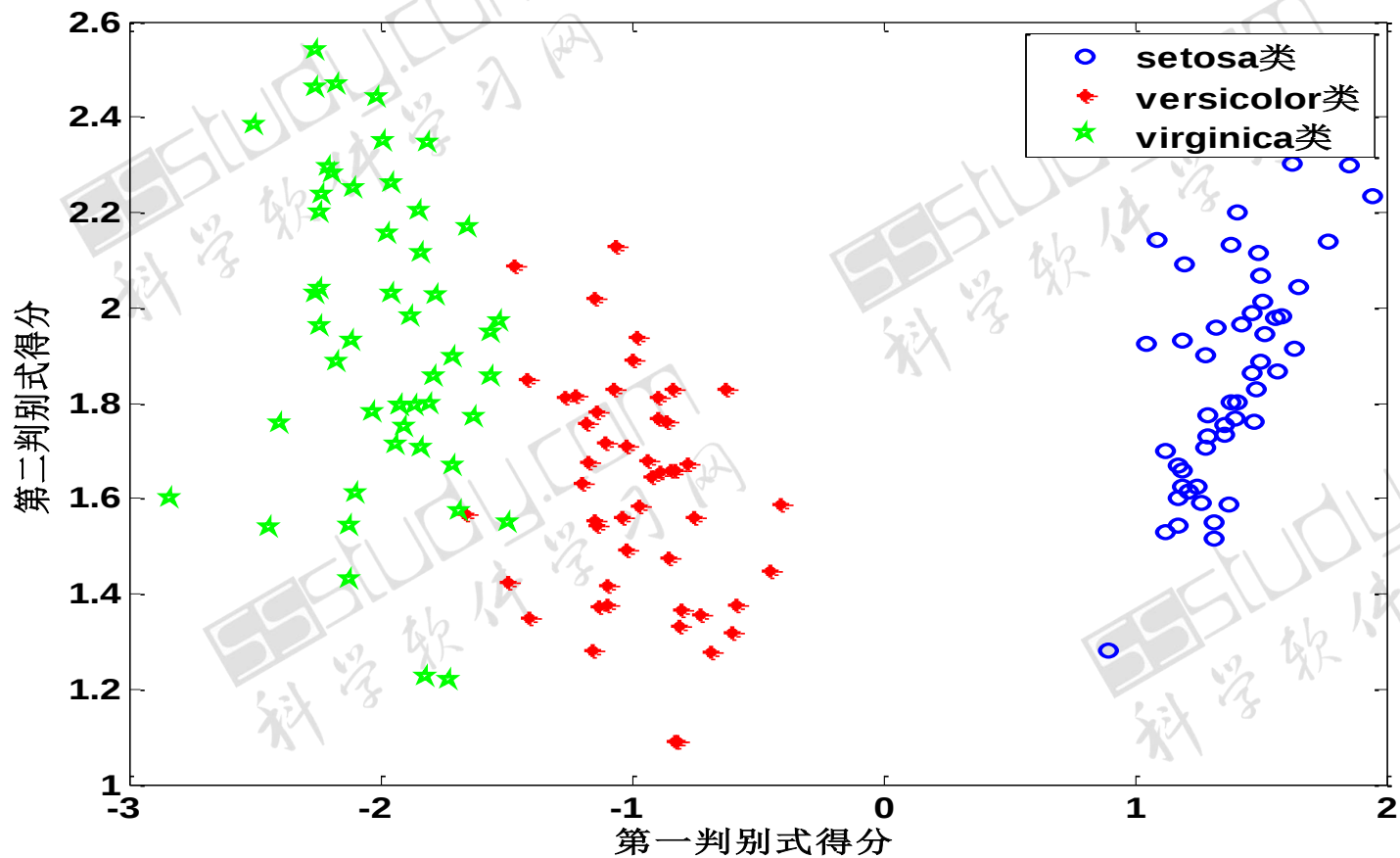
```
headline = {'观测序号','x1','x2','x3','x4','判归组别'};
```

```
[headline;num2cell([(1:10)',x]),outclass] % 查看结果
```

判别结果

'观测序号'	'x1'	'x2'	'x3'	'x4'	'判归组别'
[1]	[5.8000]	[2.7000]	[1.8000]	[0.7300]	'setosa'
[2]	[5.6000]	[3.1000]	[3.8000]	[1.8000]	'versicolor'
[3]	[6.1000]	[2.5000]	[4.7000]	[1.1000]	'versicolor'
[4]	[6.1000]	[2.6000]	[5.7000]	[1.9000]	'virginica'
[5]	[5.1000]	[3.1000]	[6.5000]	[0.6200]	'versicolor'
[6]	[5.8000]	[3.7000]	[3.9000]	[0.1300]	'setosa'
[7]	[5.7000]	[2.7000]	[1.1000]	[0.1200]	'setosa'
[8]	[6.4000]	[3.2000]	[2.4000]	[1.6000]	'versicolor'
[9]	[6.7000]	[3]	[1.9000]	[1.1000]	'setosa'
[10]	[6.8000]	[3.5000]	[7.9000]	[1]	'virginica'

判别式得分散点图



Thank You