第六章 判别分析

6.1 距离判别

6.2 Fisher判别

6.3 Bayes判别

第六章 判别分析

判别分析（discriminant analysis）是根据所研究的个体的观测指标来推断该个体所属类型的一种统计方法，在自然科学和社会科学的研究中经常会碰到这种统计问题。例如在地质找矿中要根据某异常点的地质结构、化探和物探的各项指标来判断该异常点属于哪一种矿化类型；医生要根据某人的各项化验指标的结果来判断该人属于什么病症；调查了某地区的土地生产率、劳动生产率、人均收入、费用水平、农村工业比重等指标，来确定该地区属于哪一种经济类型地区等等。

第六章 判别分析

该方法起源于1921年Pearson的种族相似系数法，1936年Fisher提出线性判别函数，并形成把一个样本归类到两个总体之一的判别法。

判别问题用统计的语言来表达，就是已有q个总体 $X_1,X_2,\cdots,X_q$ ，它们的分布函数分别为$F_1(x),F_2(x),\cdots,F_q(x)$，每个$F_i(x)$都是$p$维函数。对于给定的样本$x$，要判断它来自哪一个总体？当然，应该要求判别准则在某种意义下是最优的，例如错判的概率最小或错判的损失最小等。我们仅介绍最基本的几种判别方法，即距离判别，Bayes判别和Fisher判别。

为了作出判别，应有一个一般规则，依据 x 的值，便可以根据该规则作出判断，称这样的规则为判别规则。判别规则往往通过函数表达，这些函数称为判别函数，记为 $W(x)$。

6.1 距离判别

假定已有 r 类 $A_1, A_2, \cdots, A_r$，问待判定对象 $x = [x_1, x_2, \cdots, x_m]^T$ 属于 $A_i (i = 1, 2, \cdots, r)$ 的哪一类?

距离判别法就是建立待判定对象 x 到 A_i 的距离 $d(x, A_i)$，然后根据距离最近原则进行判别，即判别函数 $W(i, x) = d(x, A_i)$。若 $W(k, x) = \min\{W(i, x) \mid i = 1, 2, \cdots, r\}$，则 $x \in A_k$。

距离 $d(x, A_i)$ 一般采用 Mahalanobis 距离（马氏距离）。

第六章 判别分析

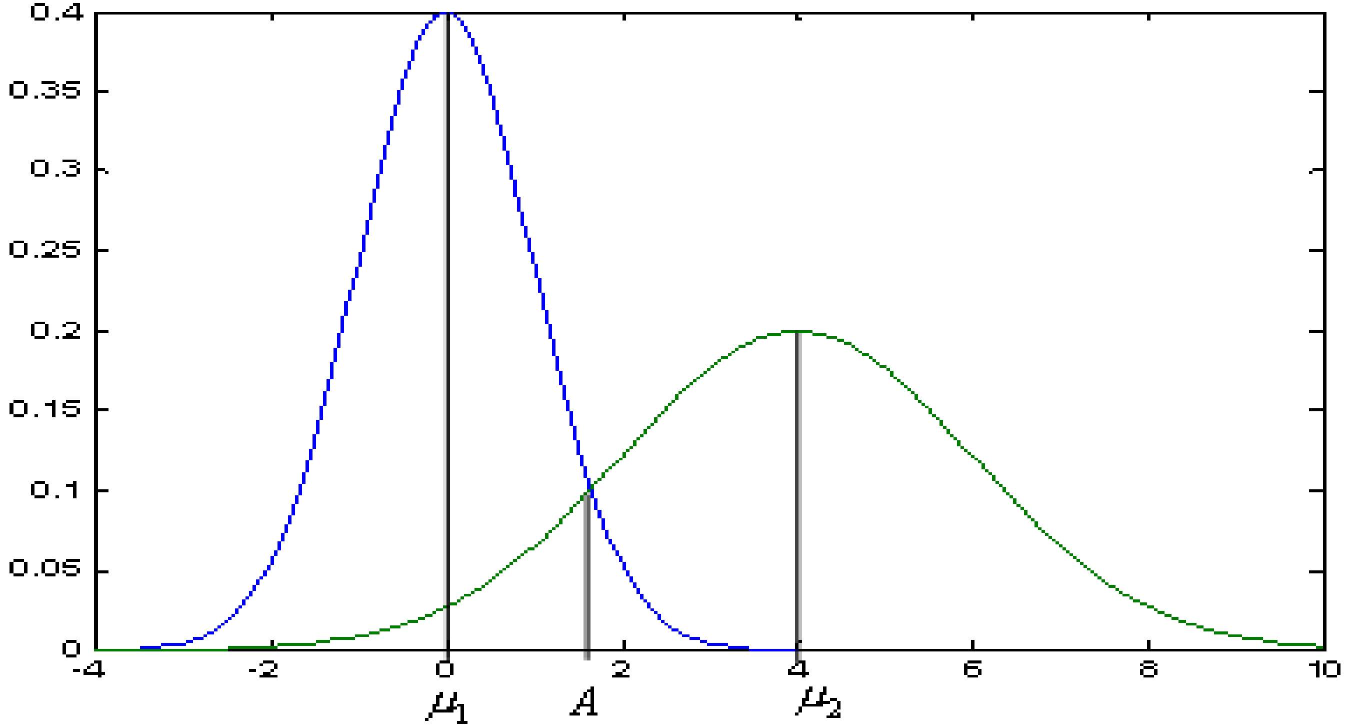
6.1 距离判别

1. Mahalanobis距离的概念

欧式距离是高维空间中两点之间的距离，它计算简单、应用广泛，但是会受到变量之间相关性的影响，当体现单一特征的多个变量参与计算时会影响结果的准确性，同时它对每个维度上的误差都同等对待，一定程度上放大了较大变量误差在距离测度中的作用。

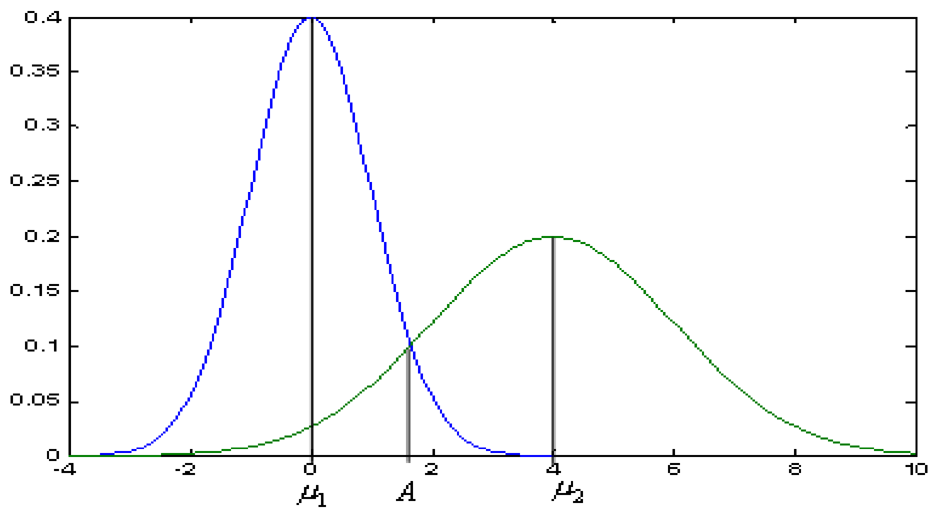
通常定义的距离是 Euclid 距离（简称欧氏距离）。但在统计分析与计算中，Euclid 距离就不适用了，看一下下面的例子。

第六章 判别分析



第六章 判别分析

为简单起见，考虑一维 $p=1$ 的情况。设 $X \sim N(0,1)$，$Y \sim N(4,2^2)$。从图上来看，A点距X的均值$\mu_1=0$较近，距Y的均值$\mu_2=4$较远。但从概率角度来分析问题，情况并非如此。经计算，A点的x值为1.66，也就是说，A点距$\mu_1=0$是$1.66\sigma_1$，而A点距$\mu_2=4$却只有$1.17\sigma_2$，因此，应该认为A点距μ_2更近一点。



第六章 判别分析

定义 设x,y是从均值为μ，协方差为Σ的总体A中抽取的样本，则总体A内两点x与y的 Mahalanobis 距离（简称马氏距离）定义为

$$d(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)},$$

定义样本x与总体A的 Mahalanobis 距离为

$$d(x,A)=\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}.$$

特点:

- 它不受量纲的影响，两点之间的马氏距离与原始数据的测量单位无关;
- 由标准化数据和中心化数据(即原始数据与均值之差）计算出的二点之间的马氏距离相同;
- 可以排除变量之间的相关性的干扰;
- 要求总体样本数大于样本的维数，否则得到的总体样本协方差矩阵逆矩阵不存在。

2.距离判别的判别准则和判别函数

在这里讨论两个总体的距离判别，分协方差相同和协方差不同两种进行讨论。

设总体A和B的均值向量分别为μ_1和μ_2，协方差阵分别为Σ_1和Σ_2，今给一个样本x，要判断x来自哪一个总体。

(1) 首先考虑协方差相同，即

$$\mu_1 \neq \mu_2, \quad \Sigma_1 = \Sigma_2 = \Sigma.$$

要判断x来自哪一个总体，需要计算x到总体A和B Mahalanobis 距离$d(x,A)$和$d(x,B)$，然后进行比较，若$d(x,A) \le d(x,B)$，则判定x属于A；否则判定x来自B。由此得到如下判别准则

$$x \in \begin{cases} A, d(x,A) \le d(x,B), \\ B, d(x,A) > d(x,B). \end{cases}$$

现在引进判别函数的表达式，考察$d^2(x,A)$与$d^2(x,B)$之间的关系，有

$$d^2(x,B)-d^2(x,A)=(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)$$

$$=2(x-\bar{\mu})^T\Sigma^{-1}(\mu_1-\mu_2),$$

其中$\bar{\mu}=\dfrac{\mu_1+\mu_2}{2}$。

$$d^2(x,B)-d^2(x,A)=2(x-\bar{\mu})^T\Sigma^{-1}(\mu_1-\mu_2) \qquad \bar{\mu}=\frac{\mu_1+\mu_2}{2}$$

令

$$w(x)=(x-\bar{\mu})^T\Sigma^{-1}(\mu_1-\mu_2),$$

称$w(x)$为两总体距离的判别函数，因此判别准则变为

$$x\in\begin{cases}A, w(x)\geq 0,\\B, w(x)<0.\end{cases}$$

(2) 再考虑协方差不同的情况，即

$$\mu_1 \neq \mu_2, \quad \Sigma_1 \neq \Sigma_2,$$

对于样本x，在方差不同的情况下，判别函数为

$$w(x) = (x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1).$$

3.距离判别的具体计算

假定已有r类判别对象$A_1, A_2, \cdots, A_r$，每一类A_i由m个指标的n_i个样本确定，即A_i类有样本值矩阵

$$A_i = \begin{bmatrix} a_{11}^{(i)} & a_{12}^{(i)} & \cdots & a_{1m}^{(i)} \\ a_{21}^{(i)} & a_{22}^{(i)} & \cdots & a_{2m}^{(i)} \\ \vdots & \vdots & & \vdots \\ a_{n_i 1}^{(i)} & a_{n_i 2}^{(i)} & \cdots & a_{n_i m}^{(i)} \end{bmatrix} = \begin{bmatrix} (a_1^{(i)})^T \\ (a_2^{(i)})^T \\ \vdots \\ (a_{n_i}^{(i)})^T \end{bmatrix},$$

其中，A_i矩阵的第k行是A_i的第k个样本点的观测值向量。问待判定对象$x = [x_1, x_2, \cdots, x_m]^T$属于$A_i (i = 1, 2, \cdots, r)$的哪一类?

3.sklearn.neighbors 模块的 KNeighborsClassifier 函数

sklearn.neighbors 模块的 KNeighborsClassifier 函数实现距离判别法的分类，其调用格式为：

KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None)

其中，第一个参数 n_neighbors 指定分类的类别数；algorithm 的取值可以为：'auto', 'ball_tree', 'kd_tree', 'brute'；metric 的默认取值为'minkowski'，即默认的距离为欧氏距离，metric 的取值及其含义见表 11.1。

字符串	含义
'euclidean'	x,y的欧氏距离：$\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2}$
'manhattan'	x,y的曼哈顿距离：$\sum_{i=1}^{m}\lvert x_i-y_i\rvert$
'chebyshev'	x,y的切比雪夫距离：$\max\{\lvert x_i-y_i\rvert, i=1,2,\cdots,m\}$
'minkowski'	x,y的闵可夫斯基距离：$\sqrt[p]{\sum_{i=1}^{m}\lvert x_i-y_i\rvert^p}$，$p=1$为曼哈顿距离，$p=2$为欧氏距离
'wminkowski'	x,y的带权重闵可夫斯基距离：$\sqrt[p]{\sum_{i=1}^{m}\left(w_i\lvert x_i-y_i\rvert\right)^p}$，其中$w=[w_1,w_2,\cdots,w_m]$为权重
'seuclidean'	标准化欧氏距离，即各指标变量的数据都标准化为均值为 0，标准差为 1
'mahalanobis'	x,y的马氏距离：$\sqrt{(x-y)^T\Sigma^{-1}(x-y)}$，$\Sigma$为样本的协方差矩阵。当$r$个类别的总体相互独立时，$\Sigma$为单位阵，此时马氏距离等同于欧氏距离

例 6.1 1989 年国际大学生数学建模竞赛 A 题：蠓虫分类，蠓虫是一种昆虫，分为很多类型，其中有一种名为 Af，是能传播花粉的益虫；另一种名为 Apf，是会传播疾病的害虫。这两种类型的蠓虫在形态上十分相似，很难区别。现测得 9 只 Af 和 6 只 Apf 蠓虫的触角长度和翅膀长度数据。

Af：(1.24,1.27)，(1.36,1.74)，(1.38,1.64)，(1.38,1.82)，(1.38,1.90)，(1.40,1.70)，(1.48,1.82)，(1.54,1.82)，(1.56,2.08)；

Apf：(1.14,1.78)，(1.18,1.96)，(1.20,1.86)，(1.26,2.00)，(1.28,2.00)，(1.30,1.96)。

若两类蠓虫协方差矩阵相等，试判别(1.24,1.80)，(1.28,1.84)与(1.40,2.04)三只蠓虫属于哪一类。

程序设计如下；

```
#程序文件 Pex6_1.py
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
x0=np.array([[1.24,1.27], [1.36,1.74], [1.38,1.64], [1.38,1.82], [1.38,1.90], [1.40,1.70],
        [1.48,1.82], [1.54,1.82], [1.56,2.08], [1.14,1.78], [1.18,1.96], [1.20,1.86],
        [1.26,2.00], [1.28,2.00], [1.30,1.96]])  #  输入已知样本数据
x=np.array([[1.24,1.80], [1.28,1.84], [1.40,2.04]])  # 输入待判样本点数据
```

```
g=np.hstack([np.ones(9),2*np.ones(6)])  #g  为已知样本数据的类别标号
v=np.cov(x0.T)  #  计算协方差
knn=KNeighborsClassifier(2,metric='mahalanobis',metric_params={'V':
v}) #马氏距离分类
knn.fit(x0,g); pre=knn.predict(x); print("马氏距离分类结果：",pre)
print("马氏距离已知样本的误判率为：",1-knn.score(x0,g))
knn2=KNeighborsClassifier(2)  #  欧氏距离分类
knn2.fit(x0,g);
pre2=knn2.predict(x); print("欧氏距离分类结果：",pre2)
print("欧氏距离已知样本的误判率为：",1-knn2.score(x0,g))
```

程序运行结果如下：

马氏距离分类结果： [2. 2. 1.]

马氏距离已知样本的误判率为： 0.0

欧氏距离分类结果： [2. 1. 2.]

欧氏距离已知样本的误判率为： 0.0

从程序运行结果看，使用马氏距离分类时，把前两个样本点判为 Apf，第三个样本点判为 Af；使用欧氏距离分类时，把第一个和第三个样本点判为 Apf，第二个样本点判为 Af，但两种分类法对已知样本点的误判率都为 0，但我们倾向于使用马氏距离进行分类。

例 6.2 从健康人群、硬化症患者和冠心病患者中分别随机选取 10 人、6 人和 4 人，考察了他们各自心电图的 5 个不同指标（记作 x_1, x_2, x_3, x_4, x_5）如表 11.2 所示，试对两个待判样品作出判断。

序号	x_1	x_2	x_3	x_4	x_5	类型
1	8.11	261.01	13.23	5.46	7.36	1
2	9.36	185.39	9.02	5.66	5.99	1
3	9.85	249.58	15.61	6.06	6.11	1
4	2.55	137.13	9.21	6.11	4.35	1
5	6.01	231.34	14.27	5.21	8.79	1
6	9.46	231.38	13.03	4.88	8.53	1
7	4.11	260.25	14.72	5.36	10.02	1
8	8.90	259.51	14.16	4.91	9.79	1
9	7.71	273.84	16.01	5.15	8.79	1
10	7.51	303.59	19.14	5.7	8.53	1
11	6.8	308.9	15.11	5.52	8.49	2
12	8.68	258.69	14.02	4.79	7.16	2
13	5.67	355.54	15.13	4.97	9.43	2
14	8.1	476.69	7.38	5.32	11.32	2
15	3.71	316.12	17.12	6.04	8.17	2
16	5.37	274.57	16.75	4.98	9.67	2
17	5.22	330.34	18.19	4.96	9.61	3
18	4.71	331.47	21.26	4.3	13.72	3
19	4.71	352.5	20.79	5.07	11	3
20	3.36	347.31	17.9	4.65	11.19	3
21	8.06	231.03	14.41	5.72	6.15	待判
22	9.89	409.42	19.47	5.19	10.49	待判

把表中的数据保存到 Excel 文件 data62.xlsx 中，文件没有表头，总共 22 行，7 列数据。

```
#程序文件 Pex6_2.py
import numpy as np
import pandas as pd
from sklearn.neighbors import KNeighborsClassifier
a=pd.read_excel("data62.xlsx",header=None)
b=a.values
x0=b[:-2,1:-1].astype(float)  # 提取已知样本点的观测值
y0=b[:-2,-1].astype(int)
```

```
x=b[-2:,1:-1]  #  提取待判样本点的观察值
v=np.cov(x0.T)  #  计算协方差
knn=KNeighborsClassifier(3,metric='mahalanobis',metric_params={'V': v}) #马氏距离分类
knn.fit(x0,y0); pre=knn.predict(x); print("分类结果：",pre)
print("已知样本的误判率为：",1-knn.score(x0,y0))
```

程序运行结果如下：

分类结果： [1   1]

已知样本的误判率为： 0.15000000000000002

即样品 1 和样品 2 都属于第 1 类。

已知样本的误判率为 15%，是比较高的。我们把可以使用的距离判别都测试了一遍，马氏距离的误判率是最低的。