蠓虫的分类

——判别分析与聚类分析

主讲：刘琼荪

实际应用问题

1、蠓虫的分类问题

两种蠓Af和Apf根据它们触角长度和翼长加以区分。假定已知类别的部分样本数据，即 9只Af蠓虫和 6只Apf蠓虫的数据。

若给定一只蠓虫，如何正确地区分它属于哪一族？

已知蠓虫类别的数据

Af	触角	1.24	1.36	1.38	1.38	1.38	1.4	1.48	1.54	1.56
	翼长	1.72	1.74	1.64	1.82	1.9	1.7	1.82	1.82	2.08

Apf	触角	1.14	1.18	1.2	1.26	1.28	1.3			
	翼长	1.78	1.96	1.86	2.0	2.0	1.96			

未知类别的三个样本数据：

（1.24，1.8）、（1.28，1.84）、（1.4，2.04）

2、DNA序列的分类模型

假定已知两组人工已分类的DNA序列（20个已知类别的人工制造的序列），其中序列标号1—10为A类，11-20为B类。要求我们从中提取已经分类了的DNA序列片段的特征和构造分类方法，并且还要衡量所用分类方法的好坏，从而构造或选择一种较好的分类方法。测试对象是20个未标明类别的人工序列（标号21—40）和182个自然DNA序列。例如A类：

a1='aggcacggaaaaacgggaataacggaggaggacttggcacggcattaca
cggaggacgaggtaaaggaggcttgtctacggccggaagtgaaggggatatg
accgcttgg';

b1='gttagatttaacgtttttatggaatttatggaattataaatttaaaaa
tttatatttttaggtaagtaatccaacgttttattacttttaaaattaaa
tatttatt';

……

需要进行特征提取，将字符转换成数据。上述两个问题极其相似，都属于分类问题。

有关的分类方法有：判别分析、聚类分析、神经网络分析、粗集理论……

现代统计分析方法与应用

方法概述

- 判别分析
- 主成分分析
- 因子分析
- 聚类分析

判别分析

目的：对某一种研究对象的归属作出判断。

例如：在经济学中，根据人均国民收入、人均消费水平、人均住房面积等多种指标去判定一个国家的经济发展程度所属类型（高、中、低等）。

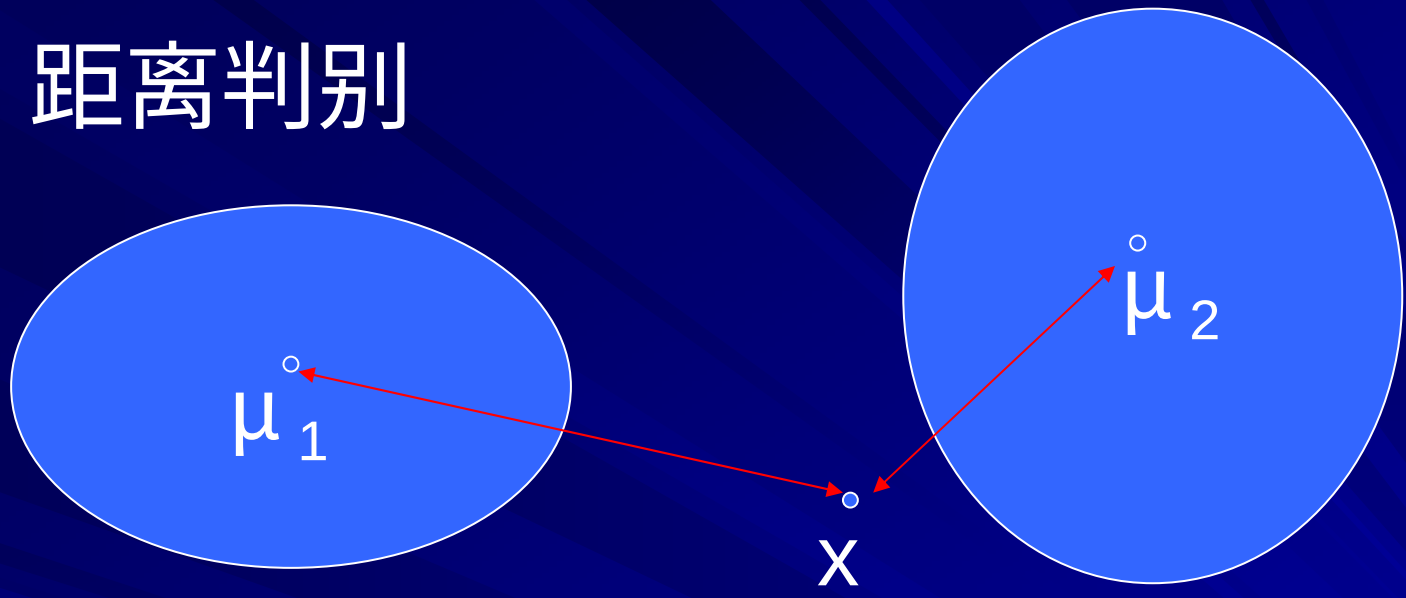
判别分析的统计模型描述

设有k个类别G1，G2，...，Gk (总体)，对任意样品x，希望建立一个准则能判定它属于哪个总体？

关键是建立什么样的判别准则，判断x的归属问题。

要求建立的准则在某中意义下是最优的。例如概率最小或错判损失最小等。

1、距离判别



$X=\{x_1, x_2, \ldots, x_n\}$

$\mu_1=\{a_1,\ldots,a_n\}$, $\mu_2=\{b_1,\ldots,b_n\}$

$d^2(x,G_1)=(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)$

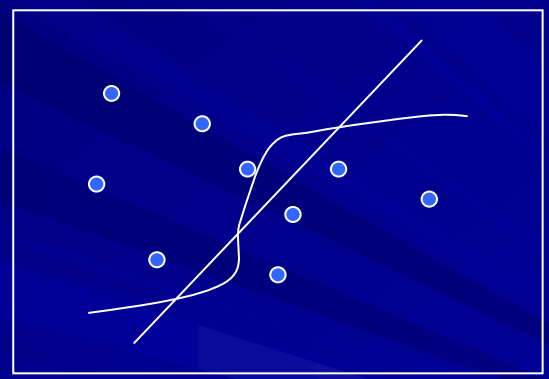
$d^2(x,G_2)=(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)$

其中Σ_1，Σ_2分别为协方差矩阵

假设$\Sigma_1=\Sigma_2=\Sigma$，可以证明

$d^2(x,G_1)-d^2(x,G_2)=-2(x-(\mu_1+\mu_2)/2)'\Sigma^{-1}(\mu_1-\mu_2)$

记为w(x)，显然 w(x)是x的线性函数。



判别规则如下：

① $x\in G_1$　　$w(x)>0$

② $x\in G_2$　　$w(x)<0$

③ 待判　　$w(x)=0$　　（线性判别法）

实际问题中， $\mu_1, \mu_2, \Sigma_1, \Sigma_2$往往是未知的，它们可以用各总体的训练样本作估计。

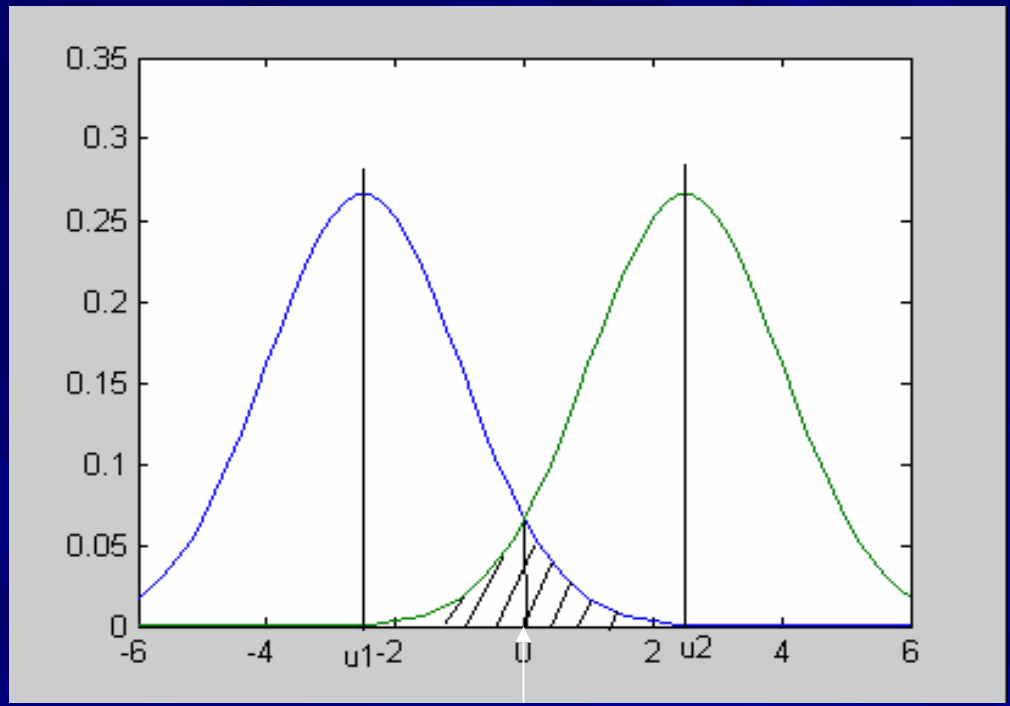
判别准则的评价

当判别准则提出后，还应该研究其优良性。这里我们主要考虑误判概率。

判别情况分析

在正态性的假定下，误判概率为图中阴影部分的面积。

如何计算？



阈值 $\overline{\mu}$

阈值点的选择极为重要。注意：如果两个总体靠得很近，则无论用何种办法，误判的概率都很大。

误判率回代估计法

设G_1, G_2为两个总体，$x^{(1)}$,$x^{(2)}$分别是来自两个总体的样本，其样本容量分别是n_1,n_2。以全体训练样本，逐个代入已建立的判别准则中判别其归属，这个过程称为回判。回判结果如下表：

回判情况 / 实际归类	G_1	G_2
G_1	n_{11}	n_{12}
G_2	n_{21}	n_{22}

其中n_{12}表示属于G_1的样品误判为G_2的个数，则总的误判个数为$n_{12}+n_{21}$。误判率回代估计：

$$\hat{a} = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

误判率的交叉确认估计

1）从总体G_1的容量为n_1的训练样本中，剔除其中一个样品,用剩余的n_1-1的训练样本和总体G_2的n_2个训练样板一起建立判别函数；

2）用建立的判别函数对删除的样本作判别；

3）重复以上步骤，直到n_1个训练样本依次被剔除，又进行判别，其误判样品个数记为n_{12}*。

4）对总体G_2的训练样本重复1）2）3），其误判样品个数为n_{21}*。

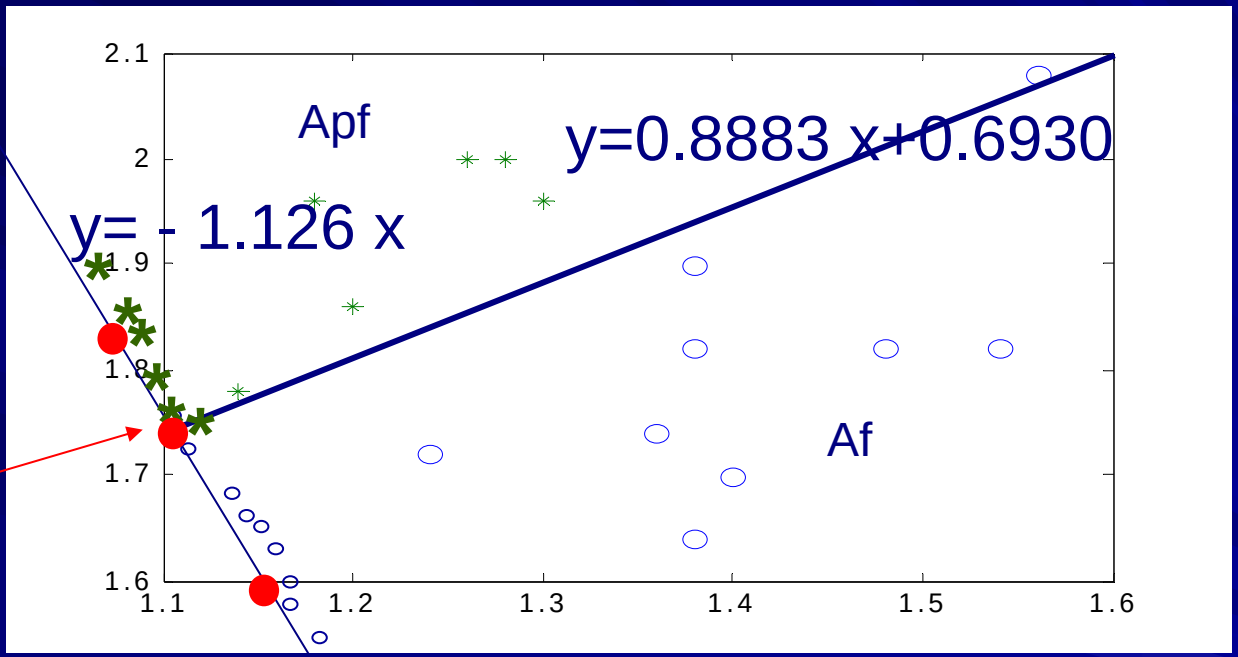
$$\hat{a} = \frac{n^*_{12} + n^*_{21}}{n_1 + n_2}$$

2、Fisher判别

判别思想：

通过将多维数据投影到某个方向上。投影的原则是将总体与总体之间尽可能分开，再选择合适的判别规则，将待判的样品进行分类判别。

Fisher判别方法的图形解释



蠓虫分类的散点图

Fisher判别方法概述

欲寻找线性函数 $y = a'x$, 使得来自两个总体的数据间的距离大，而来自同一个总体数据间的变异小。可以证明：

$$a=(\mu_1-\mu_2)'\,\Sigma^{-1},\ \text{其中}\ \Sigma_1=\Sigma_2=\Sigma$$

Fisher判别方法概述

判别规则：

当 $y \geq m$时，判 $x \in G_1$

当 $y < m$时，判 $x \in G_2$

其中，m是两个总体均值在投影方向上的中点

即 $$m = \frac{a\mu_1 + a\mu_2}{2} = \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$$

以蠓虫分类问题，用Fisher判别方法编程如：fisher1.m(结果：三个新的数据属于Af类)

3、Bayes判别概述

前面讨论的两种方法没有考虑两个总体的分布这个信息，也没有考虑错判所造成的损失。Bayes判别是考虑了这两种因素的判别方法。

建立一种错判损失函数，从而计算总的平均损失，Bayes判别准则：使总的平均损失达到极小。

聚类分析

基本概念

聚类分析(Cluster Analysis)是研究“物以类聚”的一种方法。

根据一批样品的多个观测指标，具体找出能够度量样品或指标之间相似程度的统计量，以这些统计量为划分类型的依据，将相似程度较大的样品（指标）聚合为一类。

方法概述

√

- 系统聚类法
- 动态聚类法
- 图论聚类法
- 模糊聚类法
- 有序聚类法

数据结构

P个指标 $x_1, x_2, \cdots, x_p$

n个样本

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

系统聚类法

1、对样品进行聚类

将样品间的“靠近”程度由某种距离来刻画。常见的距离有欧氏、马氏等，如：

$$d_{ij}^{\ 2}=[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2]^{\frac{1}{2}},\ d_{ij}^{\ q}=[\sum_{k=1}^{p}(x_{ik}-x_{jk})^q]^{\frac{1}{q}}$$

Minkowski

当$q=1$时，$$d_{ij}=\sum_{k=1}^{p}|x_{ik}-x_{jk}|$$

$$d_{ij}^{\ 2}(M)=(x_i-x_j)'\Sigma^{-1}(x_i-x_j)$$

马氏

2、对指标进行聚类

对指标之间的“靠近‘程度往往用相似系数来刻画。

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{[\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2]^{\frac{1}{2}}}$$

系统聚类法（Hierarchical Clustering）的计算步骤：

1）计算n个样品两两间的距离$\{d_{ij}\}$，记D

2）构造n个类，每个类只包含一个样品；

3）合并距离最近的两类为一新类；

4）计算新类与当前各类的距离；若类的个数等于1，转到5）；否则回3）；

5）画聚类图；

6）决定类的个数和类；

Matlab软件对系统聚类法的实现

cluster	从连接输出(linkage)中创建聚类
clusterdata	从数据集合(x)中创建聚类
dendrogram	画系统树状图
linkage	连接数据集中的目标为二元群的层次树
pdist	计算数据集合中两两元素间的距离(向量)
squareform	将距离的输出向量形式定格为矩阵形式
zscore	对数据矩阵 X 进行标准化处理

各种命令解释

1、T = clusterdata(X, cutoff)

其中X为数据矩阵，cutoff是创建聚类的临界值。即表示欲分成几类。

以上语句等价与以下几句命令：

```
Y=pdist(X,'euclid')
Z=linkage(Y,'single')
T=cluster(Z,cutoff)
```

以上三组命令更加灵活，可以自由选择各种方法！

2、T = cluster(Z, cutoff)

从逐级聚类树中构造聚类，其中Z是由语句likage产生的(n-1)×3阶矩阵，cutoff是创建聚类的临界值。

3、 Z = linkage(Y)

Z = linkage(Y, 'method')

创建逐级聚类树，其中Y是由语句pdist产生的n(n-1)/2 阶向量，’method’表示用何方法，默认值是欧氏距离（single)。有’complete’——最长距离法；‘average’——类平均距离；‘centroid’——重心法 ；‘ward‘——递增平方和等。

4、 Y = pdist(X)
Y = pdist(X, 'metric')

计算数据集X中两两元素间的距离， ‘metric’表示使用特定的方法，有欧氏距离‘euclid’、标准欧氏距离‘SEuclid’、马氏距离‘mahal’、明可夫斯基距离'Minkowski‘ 等

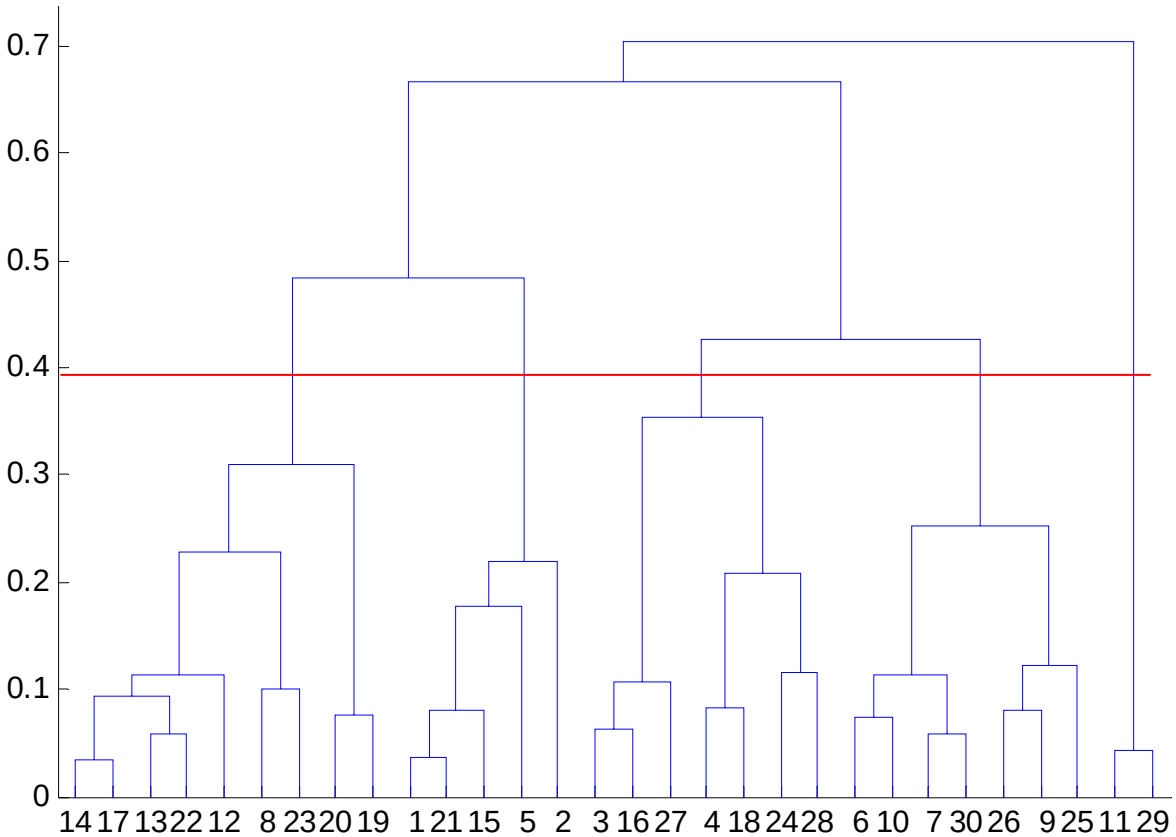
5、 H = dendrogram(Z)
H = dendrogram(Z, p)

由linkage产生的数据矩阵z画聚类树状图。P是结点数，默认值是30。

例：一段程序（julei1.m)

```
X=[7.90 39.77 8.49 12.94 19.27 11.05 2.04 13.29;
   7.68 50.37 11.35 13.3 19.25 14.59 2.75 14.87;
   9.42 27.93 8.20 8.14 16.17 9.42 1.55 9.76;
   9.16 27.98 9.01 9.32 15.99 9.10 1.82 11.35;
   10.06 28.64 10.52 10.05 16.18 8.39 1.96 10.81];
BX=zscore(X);  % 标准化数据矩阵
Y=pdist(X)    % 用欧氏距离计算两两之间的距离
D=squareform(Y)   % 欧氏距离矩阵
Z = linkage(Y)    % 最短距离法
T = cluster(Z,3)  等价于   { T=clusterdata(X,3) }
find(T==3)   % 第3类集合中的元素
[H,T]=dendrogram(Z)     % 画聚类图
```

聚类树状图分析


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
14 17 13 22 12 8 23 20 19 1 21 15 5 2 3 16 27 4 18 24 28 6 10 7 30 26 9 25 11 29

注意

不同的分类方法有不同的分类效果！

究竟采用哪一种分类好呢？

一种方法是根据分类问题本身的知识来决定取舍；

另一种方法是将几种方法的共性取出来，有争议的样本放在一边。