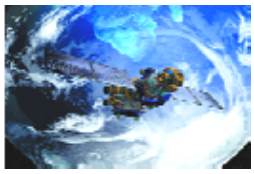
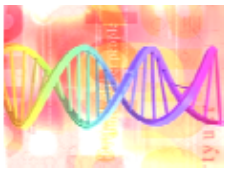





Introduction to Computational Biology

《计算生物学导论》

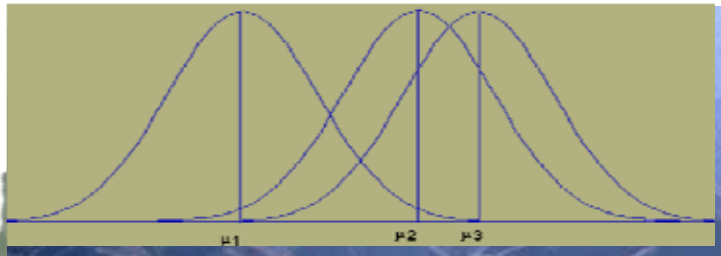
第六章

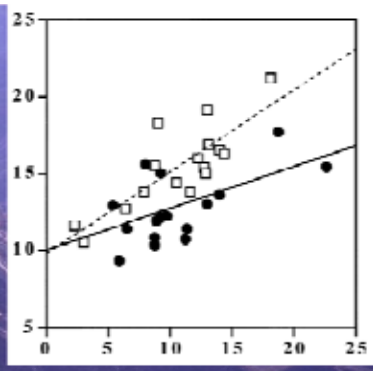
计算生物学的多元统计方法

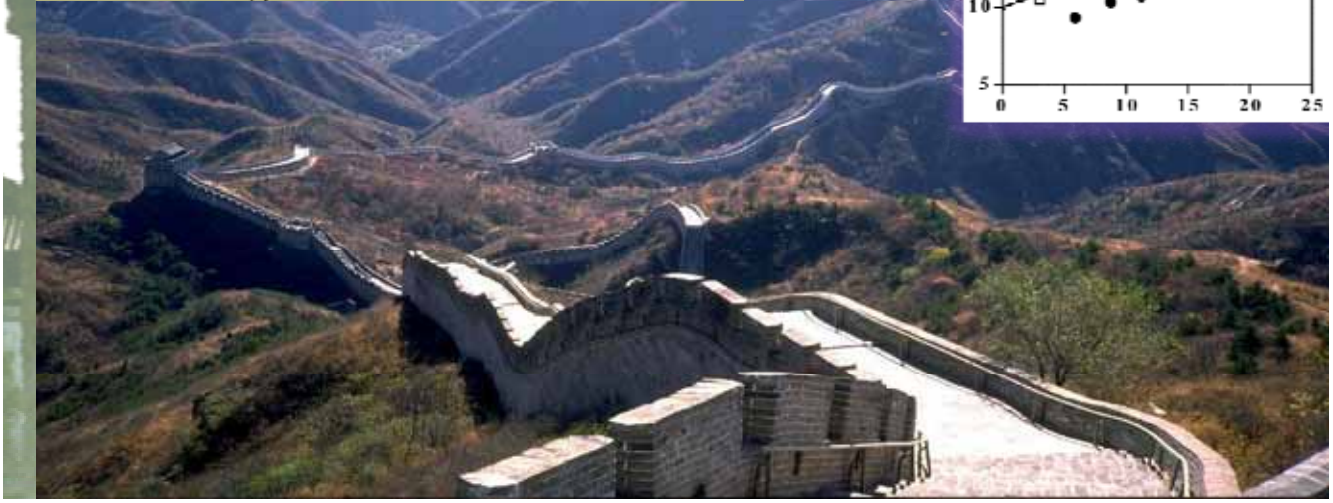

数据，尤其是大量的数据，通常不能提供信息。而统计学家的目的是揭示这些数据所包含的信息

——John Tabak

(《*Probability & Statistics—The Science of Uncertainty*》in 2004)







生命科学

生命现象
多样性
重复性
复杂性
随机性

统计分析方法

多元统计分析方法

多元回归分析方法
多元判别分析方法
聚类分析方法
主成分分析方法
相关性分析方法
……

Peking University
Introduction to Computational Biology

随机现象的多变量、多因素

战争的胜负
经济的衰退与复苏
医学病症诊断
生态环境
生物的进化
……

运用数理统计方法研究多变量、多因素问题
→ 多元统计分析理论和方法

多元统计分析

研究多元变量的统计规律性，是一元统计学的推广，同时又有多元随机变量特有的问题。

多元统计分析的主要研究内容和方法

1928年，Wishart 《多元正态总体样本协方差矩阵的精确分布》

1、降维问题（简化数据结构）

（1） 将某些较复杂的数据结构通过变量变换等方法使相互依赖的变量变成互不相关的变量

（2） 把高维空间的数据投影到低维空间，使问题得到简化同时损失的信息不太多。

主成分分析
因子分析
对应分析

Peking University
Introduction to Computational Biology

2、归类问题

对所考察的观测样本（或变量）按照相似程度进行分类、归类

聚类分析
判别分析

3、变量间的相互联系

（1） 相互依赖关系：分析一个或几个变量的变化是否依赖于另一些变量的变化。建立变量间的定量关系，并用于预测或控制

回归分析

（2） 变量间的相互关系：分析两组变量间的相互关系

典型相关性分析

Peking University
Introduction to Computational Biology

4、多元数据的统计推断

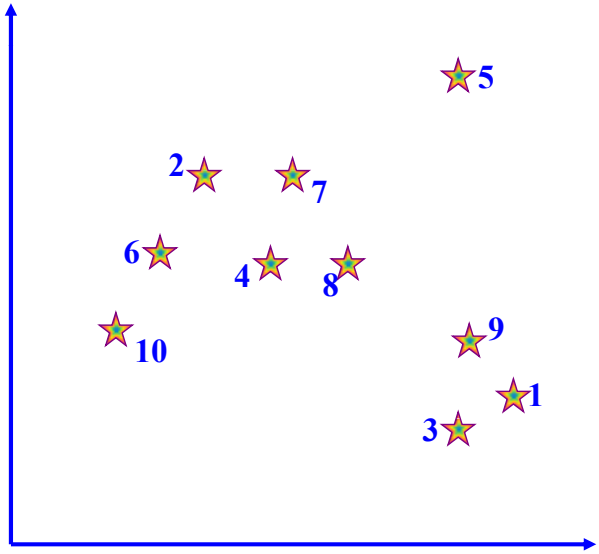
参数估计
假设检验

5、多元统计分析的数学理论基础

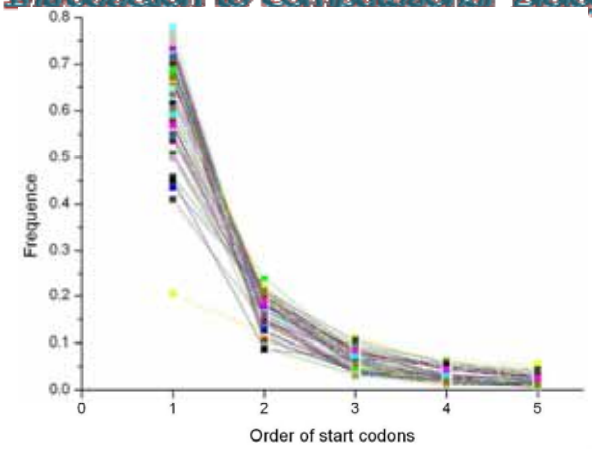
多维随机向量
多维正态随机向量
多元统计量

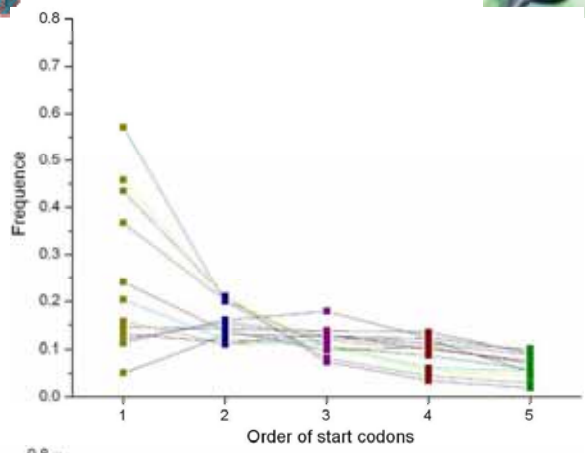
多元统计数据的图表示法

散点图

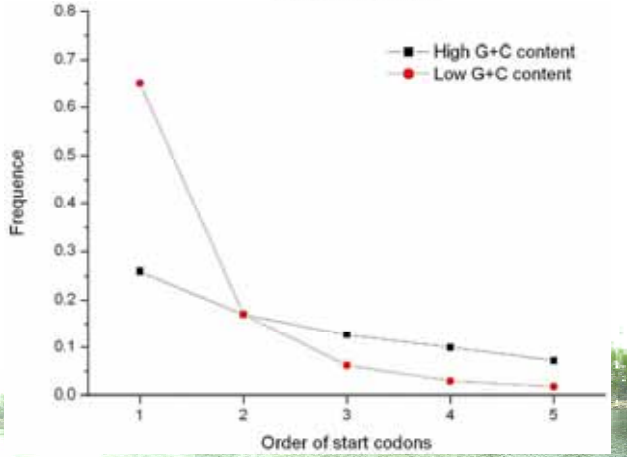


Peking University
Introduction to Computational Biology

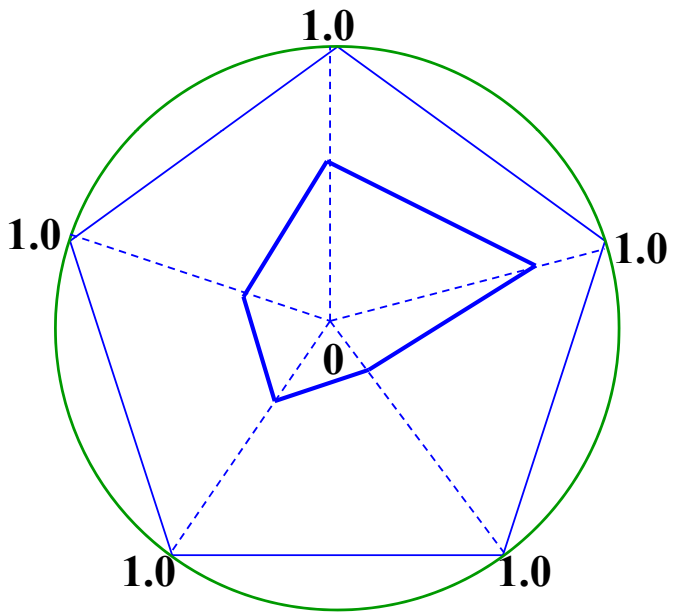




轮廓图



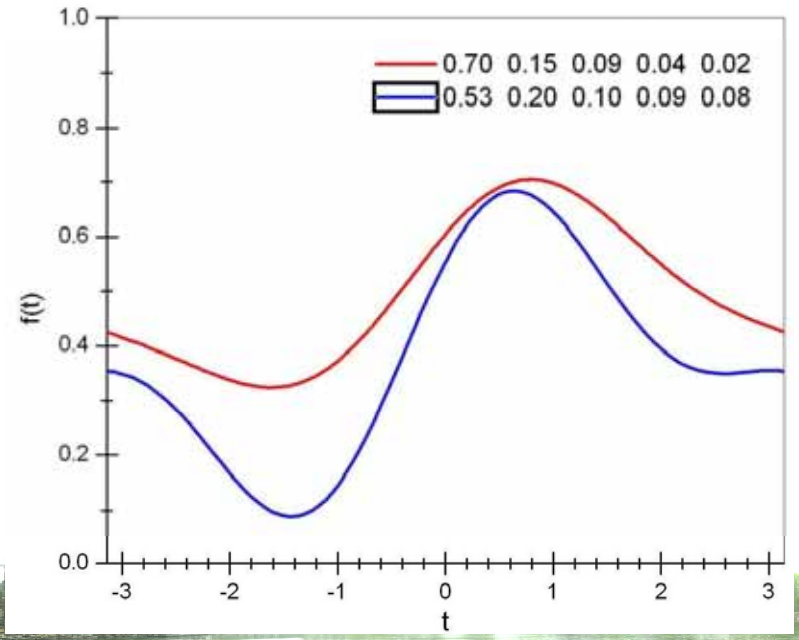
雷达图



Peking University
Introduction to Computational Biology

调和曲线图

$$f_X(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \ldots$$



§6.1

回归分析方法
(Regression analysis)

Peking University
Introduction to Computational Biology

生命活动和过程中不同现象之间的关系

变量与变量的关系：
确定性关系 → 函数关系
$U=IR$
$v=gt$

变量与变量的关系：
非确定性关系 → ……

统计相关
（具有统计规律）
$Y=f(x_1, x_2, \ldots, x_n)+\varepsilon$

回归分析方法

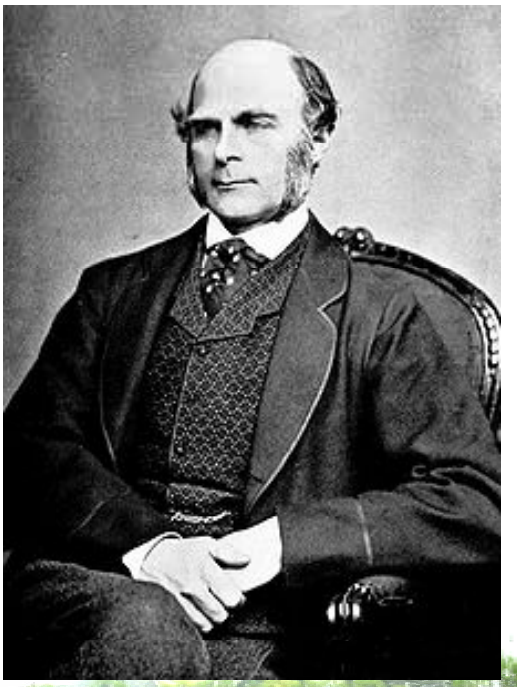
回归分析的基本问题

✓ **Regression:** the relation between selected values of *x* and observed values of *y* (from which the most probable value of *y* can be predicted for any value of *x*)

（“regression”来源于F. Galton，他曾对亲子间的身高做研究，发现父母的身高虽然会遗传给子女，但子女的身高却有逐渐“回归到中等即人的平均值”的现象。）

✓ 寻求表达量Y与$x_1, x_2, \ldots, x_n$的相关关系的经验回归方程，简称回归方程

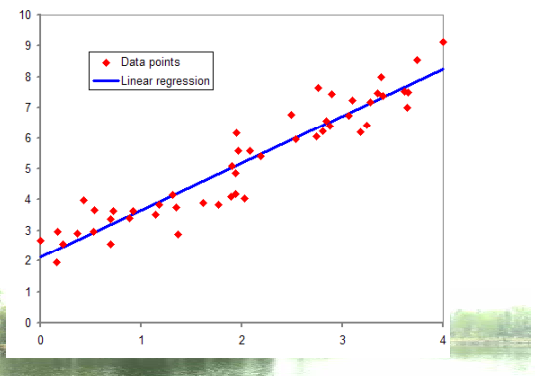
✓ 因果关系

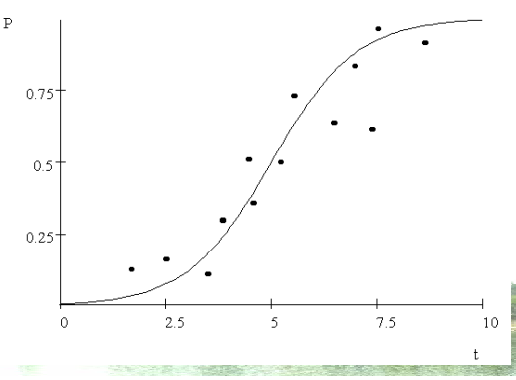


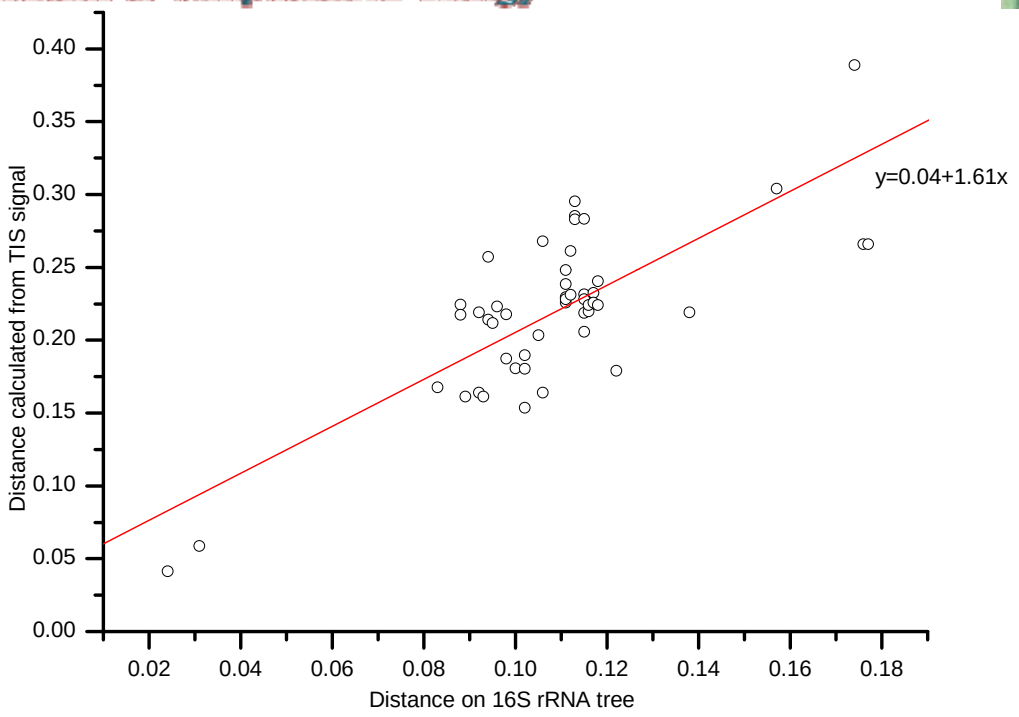
regression toward the mean

Peking University
Introduction to Computational Biology

- 利用回归方程，在一定可靠度的要求下，预估当自变量$x_1, x_2, \ldots, x_n$取确定值时，随机变量Y的取值，称为预测问题；
- 为使Y在给定的范围内取值，利用回归方程，控制自变量$x_1, x_2, \ldots, x_n$的取值范围，称为控制问题。
- 一元回归问题、多元回归问题
- 多因变量回归问题
- 线性回归问题、非线性回归问题







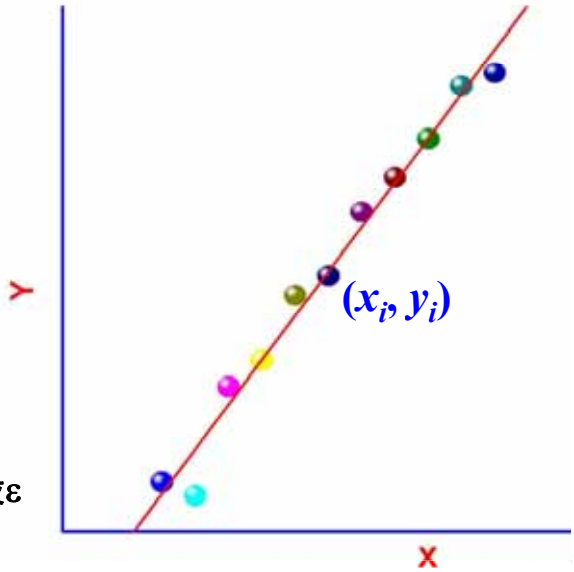
Mutation rate of translation initiation signals.
The Y axis shows the change rate between TA signal and SD signal usage. The X axis shows the evolutionary distance of 16S rRNA. Each point denotes an *Actinobacterial* genome, and all the distances are calculated by comparing with *Streptomyces. Coelicolor* A3(2).

Peking University
Introduction to Computational Biology

6.1.1 一元线性回归问题

x：可控制或可精确观测得到的数据的变量；
Y：与x具有相关关系的随机变量。

环境湿度——x_i (i=1, 2, …, n)
细菌生长数量——y_i (i=1, 2, …, n)
数据对（样本值）：(x_i, y_i) i=1, 2, …, n
→散点图(Scatter Graph)



不妨假定Y与x具有线性相关关系：

$$Y = a + bx + \varepsilon$$

其中，ε是数学期望为0的随机变量，假设ε满足正态分布，于是：

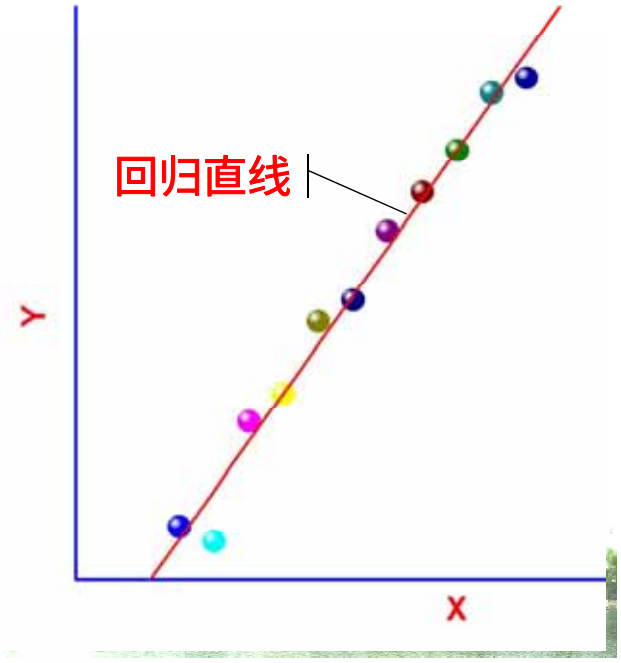
$$E(Y) = a + bx$$

根据样本(x_i, y_i), i=1, 2, …, n对系数a、b作出估计，并求得E(Y)的估计值：

回归值 $\hat{y} = \hat{a} + \hat{b}x$ 回归系数

称为一元线性回归方程。

回归直线
回归值
回归系数



求回归方程的两个基本步骤：

1. 求a、b的估计值，从而求出线性回归方程；
2. 作线性相关性检验。

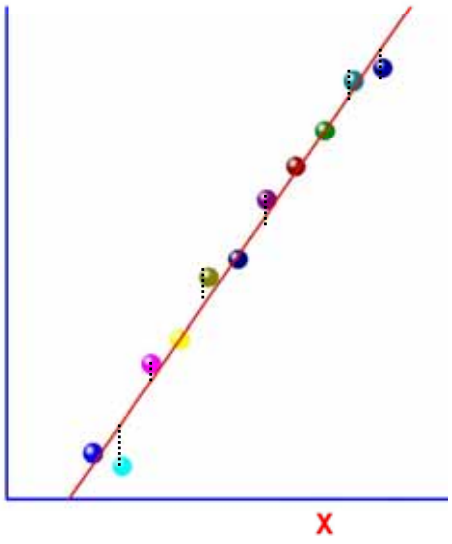
Peking University
Introduction to Computational Biology

1. a、b的最小二乘法估计

直线L：y=a+bx
样本点(x_i, y_i), i=1, 2, …, n

定义离差平方和为：

$$Q(a,b) = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$



$Q(a, b)$表示点(x_i, y_i), i=1, 2, …, n与直线L的偏离程度。

满足：

$$Q(\hat{a}, \hat{b}) = \min \quad Q(a, b)$$

的$\hat{a}, \hat{b}$称为a, b的最小二乘估计值。

根据多元函数达到极值的条件，令：

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}[y_i - (a + bx_i)] = 0 \\ \dfrac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}[y_i - (a + bx_i)]x_i = 0 \end{cases}$$

化为方程组：

$$\begin{cases} na + n\bar{x}b = n\bar{y} \\ n\bar{x}a + \left(\sum_{i=1}^{n} x_i^2\right)b = \sum_{i=1}^{n} x_i y_i \end{cases}$$

其中：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

可以证明（略），当x_i不全相同时，上述方程组有且存在唯一解。

Peking University
Introduction to Computational Biology

解得：

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = S_{xy} / S_{xx}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

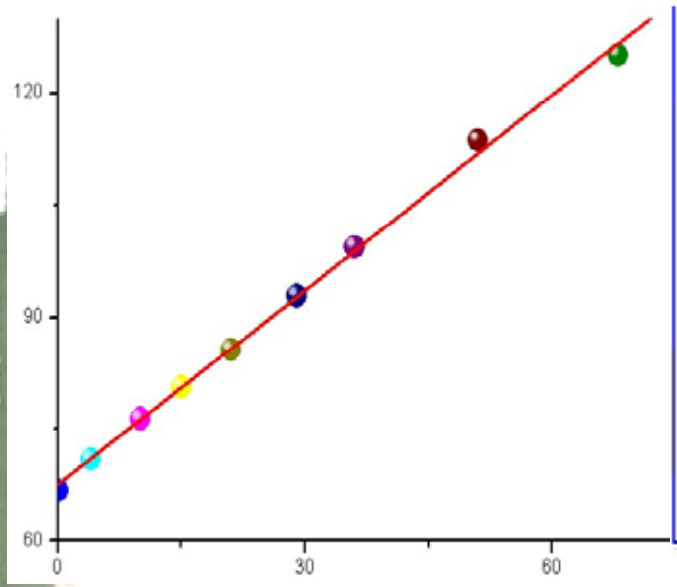
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

可以证明（略），$\hat{a}, \hat{b}$ 是a, b的最小方差无偏估计。

线性回归方程可改写为： $\hat{y} = \bar{y} + \hat{b}(x - \bar{x})$

2. 线性相关性检验



Peking University
Introduction to Computational Biology

运用方差分析。考虑样本离差平方和（总和）：

$$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})\right]^2$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

$$S_{yy}=U+Q$$

$$\hat{y}=\bar{y}+\hat{b}(x-\bar{x})$$

$$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=(\hat{b})^2S_{xx}=\frac{S_{xy}^2}{S_{xx}}$$

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

U：回归值的离差平方和，由n个x_i的离散性通过x对Y的相关关系造成
称为回归平方和（回归和）

Q：x对Y的非线性影响以及试验的随机误差（数据的随机涨落）造成
称为剩余平方和（余和）

（1）． r检验法

考虑回归和U相对于总和S_{yy}的比：

$$\frac{U}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} \leq 1$$

定义：

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

称为**相关系数**。

相关系数r：|r|≤1

|r|越大，线性相关关系越显著；

r=0，Y与x不存在线性相关关系；

|r|=1，Y与x完全线性相关（完全正/负相关）

Peking University
Introduction to Computational Biology

采用相关系数r为统计量，当：

$$|r| > r_{\alpha}(n-2)$$

（n：数据点数目）

时，认为在显著性水平α下，线性回归显著。

相关系数临界值$r_{\alpha}(n\text{-}2)$表

n-2 \ α	0.10	0.05	0.02	0.01	0.001
1	0.98769	0.99692	0.999507	0.999877	0.9999988
…					
7	0.5822	0.6664	0.7498	0.7977	0.8982
8	0.5494	0.6319	0.7155	0.7646	0.8721
…					
100	0.1638	0.1946	0.2301	0.2540	0.3211

（2）. F检验法

计算F值：

$$F = \frac{U}{Q/(n-2)}$$

数据点数目

显然，F值越大，U在总和中所占比例越大，回归性也越显著。

当：

$$F > F_{1-\alpha}(1, n-2)$$

时，认为在显著性水平α下，线性回归显著。

查表：F分布表（略）

3. 例子

x_i	0	4	10	15	21	29	36	51	68
y_i	66.7	71.0	76.3	80.6	85.7	92.9	99.4	113.6	125.1

根据散点图，确定回归方程形式

$$\hat{y} = \hat{a} + \hat{b}x$$

计算得到：

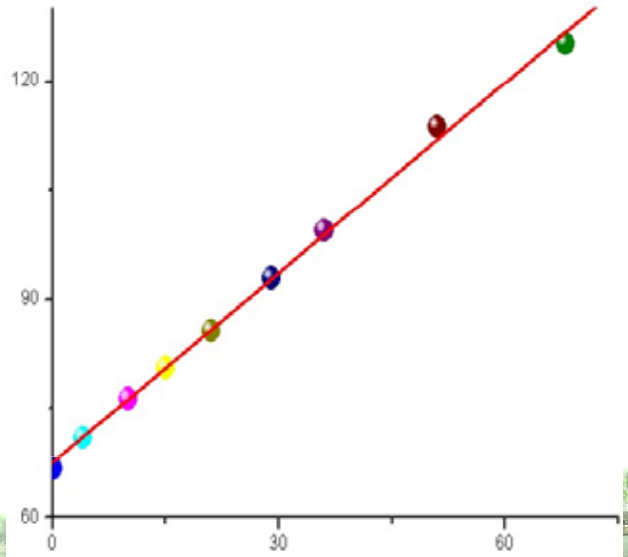
$\bar{x} = 26.0$　　$\bar{y} = 90.1$

$S_{xy} = 3534.8$　　$S_{xx} = 4060$

$S_{yy} = 3084$

$\hat{b} = 0.8706$　　$\hat{a} = 67.5078$

$$\hat{y} = 67.5078 + 0.8706x$$



线性相关性检验：

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.99896$$

查表得：

$r_{0.01}(7) = 0.7977$　　$r_{0.001}(7) = 0.8982$

显然，在显著性水平α=0.001下，Y与x的线性相关关系高度显著。

Peking University
Introduction to Computational Biology

6.1.2 可线性化的曲线回归

方法：变量替换

1. 双曲线型

$$y = a + \frac{b}{x}$$

令　$u = \frac{1}{x}$，　得到

$$y = a + bu$$

$$\frac{1}{y} = a + \frac{b}{x}$$

令　$u = \frac{1}{x}, v = \frac{1}{y}$　得到

$$v = a + bu$$

2. 指数曲线型

$$y = ae^{bx}$$

若a>0，则令v=lny，得到：

$$v = \ln a + bx$$

若a<0，则令v=ln(-y），得到：

$$v = \ln(-a) + bx$$

3. 幂函数型

$$y = ax^{b} \qquad x > 0$$

若a>0，则令v=lny，u=lnx，得到（a<0情况类推）：

$$v = \ln a + bu$$

Peking University
Introduction to Computational Biology

4. 对数曲线型

$$y = a + b\log x$$

令u=logx，得到：

$$y = a + bu$$

$$\log y = a + bx$$

令v=logy，得到：

$$v = a + bx$$

$$\log y = a + b\log x$$

令u=logx， v=logy，得到：

$$v = a + bu$$

5. S曲线型

$$y = \frac{1}{a + be^{-x}}$$

令：

$$u = e^{-x} \qquad v = 1/y$$

得到：

$$v = a + bu$$

Peking University
Introduction to Computational Biology

6.1.3 多元线性回归问题

$x_1, x_2, \ldots, x_r$：r个可控制或可精确观测得到的数据的变量；
Y：与$x_1, x_2, \ldots, x_r$具有相关关系的随机变量。

不妨假定Y与$x_1, x_2, \ldots, x_r$具有线性相关关系：

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_r x_r + \varepsilon$$

其中，ε是数学期望为0的随机误差，且满足正态分布。
对于n组样本观察值（n>r）：
$x_{i1}, x_{i2}, \ldots, x_{ir}$ (i=1, 2, …, n)
y_i (i=1, 2, …, n)

多元线性回归模型为：

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_r x_{ir} + \varepsilon_i$$
$$E(\varepsilon_i) = 0 \qquad i = 1, 2, \ldots, n$$

其中，ε_i互不相关。

记

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \cdots \\ b_n \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1r} \\ 1 & x_{21} & \cdots & x_{2r} \\ \cdots & \cdots & & \cdots \\ 1 & x_{n1} & \cdots & x_{nr} \end{bmatrix}$$

多元线性回归模型可写成：

$$\begin{cases} \mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \\ E(\boldsymbol{\varepsilon}) = \mathbf{0} \end{cases}$$

Peking University
Introduction to Computational Biology

1. 回归系数b的最小二乘估计

定义离差平方和：

$$Q(\mathbf{b}) = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_r x_{ir}) \right]^2$$

定义：在多元线性回归模型中，若存在b的估计值$\hat{\mathbf{b}}$，则对于任意一组实数$b_0, b_1, \ldots, b_r$构成的向量，都成立不等式

$$Q(\hat{\mathbf{b}}) \le Q(\mathbf{b})$$

称$\hat{\mathbf{b}}$是b的最小二乘估计。

定理：在多元线性回归模型中，设矩阵X列线性无关，则唯一存在b的最小二乘估计 $\hat{\mathbf{b}}$

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

2. 计算回归系数 $\hat{b}$

令 $\frac{\partial Q}{\partial b_j} = 0, \quad j = 0,1,2,..., r$，得到方程组：

$$b_0 n + b_1 \sum_{i=1}^{n} x_{i1} + ... + b_r \sum_{i=1}^{n} x_{ir} = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_{ij} + b_1 \sum_{i=1}^{n} x_{ij} x_{i1} + ... + b_r \sum_{i=1}^{n} x_{ij} x_{ir} = \sum_{i=1}^{n} x_{ij} y_i$$

$$j = 1,2,..., r$$

问题：求解上述方程组。

Peking University
Introduction to Computational Biology

$$\hat{b}_0 = \bar{y} - \sum_{j=1}^{r} \bar{x}_j \hat{b}_j$$

$$\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ ... \\ \hat{b}_r \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & ... & l_{1r} \\ l_{21} & l_{22} & ... & l_{2r} \\ ... & ... & & ... \\ l_{r1} & l_{r2} & ... & l_{rr} \end{bmatrix}^{-1} \begin{bmatrix} l_{1y} \\ l_{2y} \\ ... \\ l_{ry} \end{bmatrix}$$

其中：

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad \bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \qquad j = 1,2,..., r$$

$$l_{ij} = \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

$$l_{iy} = \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(y_k - \bar{y})$$

3. 显著性检验

S_{yy}：样本离差平方和
U：回归平方和（回归和）
Q：剩余平方和（余和）

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = U + Q$$

$$U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$Q = S_{yy} - U$$

$$r = \sqrt{\frac{U}{S_{yy}}}$$

$$F = \frac{U / r}{Q /(n - r - 1)}$$

Peking University
Introduction to Computational Biology

6.1.4 逐步回归问题

理想的多元回归效果

- 选取对Y有显著关联的自变量$x_1, x_2, \ldots, x_k$进行回归，剔除关联较小的自变量；
- 对于相互关联很强的自变量$x_i, x_j, \ldots, x_k$，只要从中选取一个对Y有显著关联的自变量进行回归；

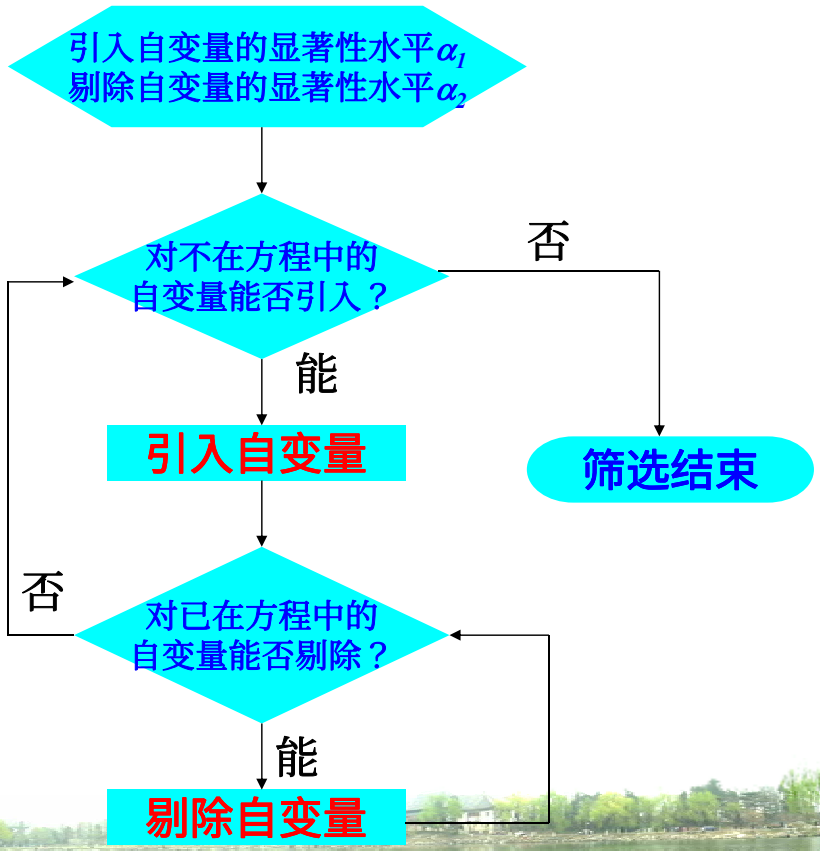
逐步回归

基于逐步筛选法的逐步回归

- 逐个引入自变量，每次引入对Y影响最显著的自变量，并对原有变量逐个进行检验（如F检验），把变为影响不显著的自变量逐个剔除。最终得到的回归方程既不漏掉对Y影响显著的自变量，又不包含对Y影响不显著的变量。

Peking University
Introduction to Computational Biology

逐步回归的基本步骤



§6.2

判别分析方法
(Discriminant analysis)

Peking University
Introduction to Computational Biology

判别分析

- 用于判别样品所属类型的统计分析方法
- 基因识别：根据某一DNA序列的核苷酸组分、功能信号特征等指标，判别是否编码蛋白序列？
- 药物设计：通过对先导化合物的一系列指标（溶解性、代谢稳定性、毒性等）的计算预测，综合判别化合物的类药性
- 医学诊断：某一病人肺部存在阴影，判别:
  肺结核？良性肿瘤？肺癌？
- 人类考古学：根据头盖骨的特征，判别：民族、性别、生活年代？
- 股票分析预测：
- 气象分析预测：
- 自然灾害分析预测：
- ……

- 判别分析问题的数学描述

设有k类n维的总体$G_1, G_2, \ldots, G_k$，

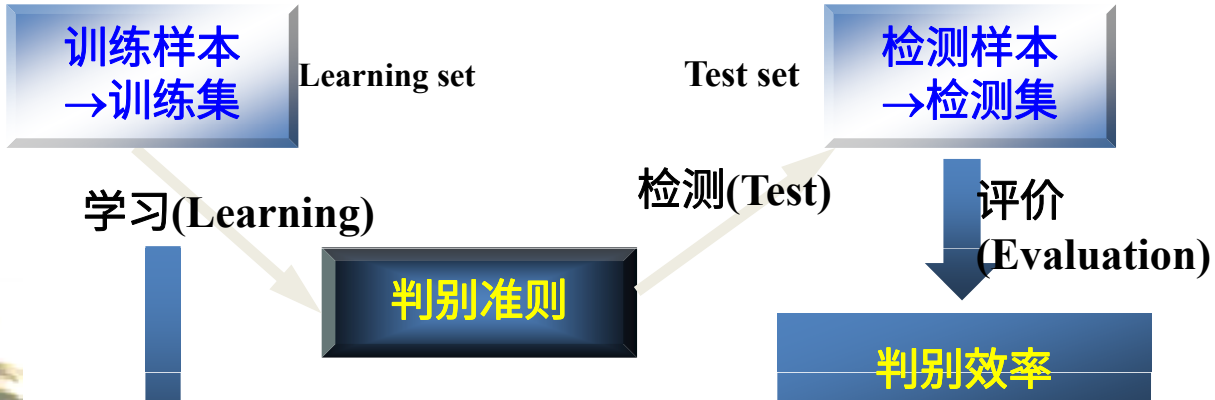
(1). 它们的分布特征已知，可以表示为$F_1(x), F_2(x), \ldots, F_k(x)$

(2). 或者知道来自各个总体的样本（训练样本）。

对于给定的一个未知样品X（检测样本），判别X属于哪类总体。

- 多元的、复杂的、高度综合的统计分析问题

Peking University
Introduction to Computational Biology



Fisher判别法
距离判别法
Bayes判别法
逐步判别法
……

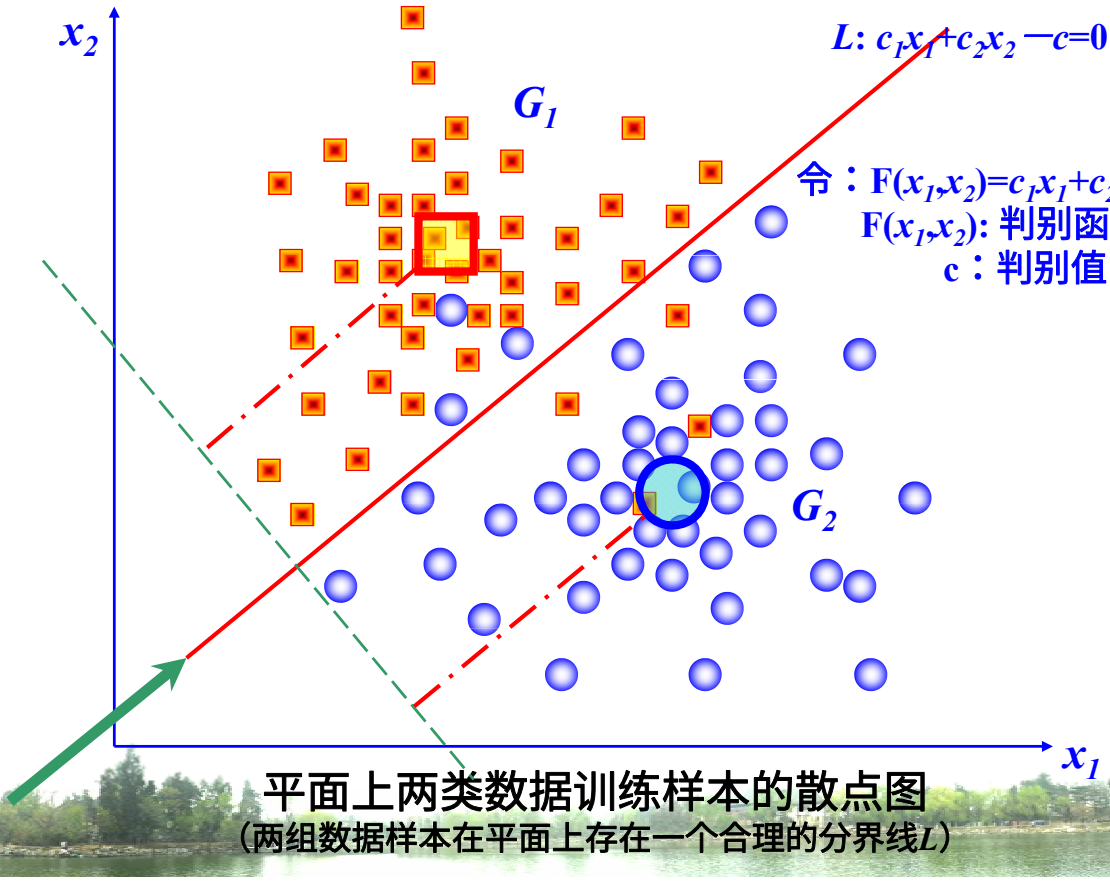
判别分析的原理

6.2.1 Fisher线性判别法

Fisher判别的基本思想

将 *k*组*n*维的数据投影到某一个方向，使得投影后的组与组之间尽可能地分开。

Peking University
Introduction to Computational Biology



平面上两类数据训练样本的散点图

（两组数据样本在平面上存在一个合理的分界线*L*）

已知：数据属性有n个，每个数据点为n维向量X：

$$X(x_1, x_2, ..., x_n)$$

已知总体数据分为两类： G_1和G_2 ，总体G_1有p个样本，总体G_2有q个样本。

		属 性 （分量）			
		1	2	…	n
总体G_1 (i=1, …, p)	1 $X_1^{(1)}$	$x_{11}^{(1)}$	$x_{12}^{(1)}$	…	$x_{1n}^{(1)}$
	…	…	…	…	…
	i $X_i^{(1)}$	$x_{i1}^{(1)}$	$x_{i2}^{(1)}$	…	$x_{in}^{(1)}$
	…	…	…	…	…
	p $X_p^{(1)}$	$x_{p1}^{(1)}$	$x_{p2}^{(1)}$	…	$x_{pn}^{(1)}$
总体G_2 (i=1, …, q)	1 $X_1^{(2)}$	$x_{11}^{(2)}$	$x_{12}^{(2)}$	…	$x_{1n}^{(2)}$
	…	…	…	…	…
	i $X_i^{(2)}$	$x_{i1}^{(2)}$	$x_{i2}^{(2)}$	…	$x_{in}^{(2)}$
	…	…	…	…	…
	q $X_q^{(2)}$	$x_{q1}^{(2)}$	$x_{q2}^{(2)}$	…	$x_{qn}^{(2)}$

目标：求解在n维空间中总体G_1和总体G_2的最优分界平面。

Peking University
Introduction to Computational Biology

定义线性判别函数为：

$$F(x_1, x_2, ..., x_n) = C_1x_1 + C_2x_2 + ... + C_nx_n$$

其中$\{C_i\}$ (i = 1, 2, …, n)为常数（待定系数）。

若判别值为 C ， 对于任何未知数据点$X(x_1, x_2, …, x_n)$，代入判别函数，依据F $(x_1, x_2, …, x_n)$与C值的比较，可以判别点X属于哪一类。

1、确定待定系数C_i (i = 1, 2, …, n)

2、确定判别值C

1、确定待定系数C_i

将类G_1的p个点、类G_2的q个点分别代入判别函数：

$$y_i^{(1)} = C_1 x_{i1}^{(1)} + C_2 x_{i2}^{(1)} + \ldots + C_n x_{in}^{(1)} \qquad i = 1,\ldots, p$$

$$y_i^{(2)} = C_1 x_{i1}^{(2)} + C_2 x_{i2}^{(2)} + \ldots + C_n x_{in}^{(2)} \qquad i = 1,\ldots, q$$

记

$$\bar{x}_i^{(1)} = \frac{1}{p}\sum_{k=1}^{p} x_{ki}^{(1)} \qquad i = 1,2,\ldots, n \qquad\qquad \bar{x}_i^{(2)} = \frac{1}{q}\sum_{k=1}^{q} x_{ki}^{(2)} \qquad i = 1,2,\ldots, n$$

$$\bar{y}^{(1)} = \frac{1}{p}\sum_{i=1}^{p} y_i^{(1)} \qquad\qquad \bar{y}^{(2)} = \frac{1}{q}\sum_{i=1}^{q} y_i^{(2)}$$

$$\bar{y}^{(1)} = C_1 \bar{x}_1^{(1)} + C_2 \bar{x}_2^{(1)} + \ldots + C_n \bar{x}_n^{(1)}$$

$$\bar{y}^{(2)} = C_1 \bar{x}_1^{(2)} + C_2 \bar{x}_2^{(2)} + \ldots + C_n \bar{x}_n^{(2)}$$

Peking University
Introduction to Computational Biology

令：

$$\delta_A = (\bar{y}^{(1)} - \bar{y}^{(2)})^2$$

δ_A是关于待定系数$\{C_i\}$的函数，与G_1和G_2两类点的几何中心的距离相关。显然，判别函数F $(x_1, x_2, \ldots, x_n)$应该使δ_A值越大越好。

令：

$$\delta_B = \sum_{i=1}^{p}\left(y_i^{(1)} - \bar{y}^{(1)}\right)^2 + \sum_{i=1}^{q}\left(y_i^{(2)} - \bar{y}^{(2)}\right)^2$$

δ_B也是关于待定系数$\{C_i\}$的函数，与G_1和G_2两类点的相对于各自几何中心的离差相关。显然，判别函数F $(x_1, x_2, \ldots, x_n)$应该使δ_B值越小越好。

Peking University
Introduction to Computational Biology

构造函数I：

$$I = I(C_1, C_2, \ldots, C_n) = \frac{\delta_A}{\delta_B} = \frac{\left(\bar{y}^{(1)} - \bar{y}^{(2)}\right)^2}{\sum_{i=1}^{p}\left(y_i^{(1)} - \bar{y}^{(1)}\right)^2 + \sum_{i=1}^{q}\left(y_i^{(2)} - \bar{y}^{(2)}\right)^2}$$

选择合适的待定系数C_i (i = 1, 2, …, n)，
使得函数I(C_1, C_2, …, C_n)达到极大值。

$$\frac{\partial I}{\partial C_i} = 0 \qquad i = 1, 2, \ldots, n$$

$$\ln I = \ln \frac{\delta_A}{\delta_B} = \ln \delta_A - \ln \delta_B$$

$$\Rightarrow \quad \frac{\partial}{\partial C_i}(\ln I) = \frac{1}{\delta_A}\frac{\partial \delta_A}{\partial C_i} - \frac{1}{\delta_B}\frac{\partial \delta_B}{\partial C_i} = 0 \qquad i = 1,2,\ldots,n$$

$$\Rightarrow \quad \frac{1}{I}\frac{\partial \delta_A}{\partial C_i} = \frac{\partial \delta_B}{\partial C_i} = 0 \qquad i = 1,2,\ldots,n$$

$$\delta_A = (\bar{y}^{(1)} - \bar{y}^{(2)})^2 = \left[\sum_{k=1}^{n} C_k (\bar{x}_k^{(1)} - \bar{x}_k^{(2)})\right]^2$$

$$\delta_B = \sum_{j=1}^{p}\left(y_j^{(1)} - \bar{y}^{(1)}\right)^2 + \sum_{j=1}^{q}\left(y_j^{(2)} - \bar{y}^{(2)}\right)^2$$

$$= \sum_{j=1}^{p}\left[\sum_{k=1}^{n} C_k (x_{jk}^{(1)} - x_k^{(1)})\right]^2 + \sum_{j=1}^{q}\left[\sum_{k=1}^{n} C_k (x_{jk}^{(2)} - x_k^{(2)})\right]^2$$

$$\Rightarrow \quad \frac{1}{I}\frac{\partial \delta_A}{\partial C_i} = \frac{2}{I}(\bar{y}^{(1)} - \bar{y}^{(2)})(\bar{x}_i^{(1)} - \bar{x}_i^{(2)})$$

$$\frac{\partial \delta_B}{\partial C_i} = 2\sum_{k=1}^{p}(y_k^{(1)} - \bar{y}^{(1)})(x_{ki}^{(1)} - \bar{x}_i^{(1)}) + 2\sum_{k=1}^{q}(y_k^{(2)} - \bar{y}^{(2)})(x_{ki}^{(2)} - \bar{x}_i^{(2)})$$

$$= 2\sum_{k=1}^{p}\sum_{j=1}^{n} C_j (x_{kj}^{(1)} - \bar{x}_j^{(1)})(x_{ki}^{(1)} - \bar{x}_i^{(1)}) +$$

$$2\sum_{k=1}^{q}\sum_{j=1}^{n} C_j (x_{kj}^{(2)} - \bar{x}_j^{(2)})(x_{ki}^{(2)} - \bar{x}_i^{(2)})$$

$$= 2\sum_{j=1}^{n} S_{ij} C_j$$

$$S_{ij} = \sum_{k=1}^{p}(x_{ki}^{(1)} - \bar{x}_i^{(1)})(x_{kj}^{(1)} - \bar{x}_j^{(1)}) + \sum_{k=1}^{q}(x_{ki}^{(2)} - \bar{x}_i^{(2)})(x_{kj}^{(2)} - \bar{x}_j^{(2)})$$

$$S_{11}C_1 + S_{12}C_2 + \ldots + S_{1n}C_n = \lambda(\bar{x}_1^{(1)} - \bar{x}_1^{(2)})$$
$$S_{21}C_1 + S_{22}C_2 + \ldots + S_{2n}C_n = \lambda(\bar{x}_2^{(1)} - \bar{x}_2^{(2)})$$
$$\ldots\ldots$$
$$S_{n1}C_1 + S_{n2}C_2 + \ldots + S_{nn}C_n = \lambda(\bar{x}_n^{(1)} - \bar{x}_n^{(2)})$$

$$\lambda = \frac{1}{I}(\bar{y}^{(1)} - \bar{y}^{(2)})$$

消去非零的因子λ，得到求解待定系数(C_1, C_2, …, C_n)的线性方程组：

$$S_{11}C_1 + S_{12}C_2 + \ldots + S_{1n}C_n = \bar{x}_1^{(1)} - \bar{x}_1^{(2)}$$
$$S_{21}C_1 + S_{22}C_2 + \ldots + S_{2n}C_n = \bar{x}_2^{(1)} - \bar{x}_2^{(2)}$$
$$\ldots\ldots$$
$$S_{n1}C_1 + S_{n2}C_2 + \ldots + S_{nn}C_n = \bar{x}_n^{(1)} - \bar{x}_n^{(2)}$$

Peking University
Introduction to Computational Biology

2、确定判别值C

判别函数已知，不妨写成：

$$y = C_1x_1 + C_2x_2 + \ldots + C_nx_n$$

将G_1的p个点、 G_2的q个点分别代入判别函数：

$$y_i^{(1)} = C_1x_{i1}^{(1)} + C_2x_{i2}^{(1)} + \ldots + C_nx_{in}^{(1)} \qquad i = 1,\ldots, p$$
$$y_i^{(2)} = C_1x_{i1}^{(2)} + C_2x_{i2}^{(2)} + \ldots + C_nx_{in}^{(2)} \qquad i = 1,\ldots, q$$

$$\bar{y}^{(1)} = \frac{1}{p}\sum_{i=1}^{p} y_i^{(1)} \qquad \bar{y}^{(2)} = \frac{1}{q}\sum_{i=1}^{q} y_i^{(2)}$$

对G_1、G_2的$(p+q)$个点的判别函数值取总体的平均值：

$$\mu = \frac{1}{p+q}\left(\sum_{i=1}^{p} y_i^{(1)} + \sum_{i=1}^{q} y_i^{(2)}\right)$$

$$= \frac{1}{p+q}\left(p\bar{y}^{(1)} + q\bar{y}^{(2)}\right)$$

显然，μ值是两类点的判别函数值的加权平均，处于两类判别函数平均值之间，也等价于两类点的总体几何中心的判别函数值。因此，将判别值C取为μ值：

$$C = \frac{p\bar{y}^{(1)} + q\bar{y}^{(2)}}{p+q}$$

Peking University
Introduction to Computational Biology

R. A. Fisher (1890-1962)
英国统计学家、遗传学家

剑桥大学Caius College宴会厅里的染色玻璃窗，上方的彩绘方格用以纪念拉丁方阵（Latin square），下方的白色文字则是为了纪念R. Fisher。

“几乎独自建立现代统计科学的天才”

“达尔文最伟大的继承者”

3、Fisher线性判别的基本步骤

问　　题

已知数据样本点分为两类：G_1和G_2，G_1有p个点，G_2有q个点。求出判别函数F $(x_1, x_2, \ldots, x_n)$和判别值C。对于任何未知数据点$X(x_1, x_2, \ldots, x_n)$，依据F $(x_1, x_2, \ldots, x_n)$与C值的比较，判别点X属于哪一类。

		属　性　（分量）			
		1	2	…	n
G_1 (i=1, …, p)	1　$X_1^{(1)}$	$x_{11}^{(1)}$	$x_{12}^{(1)}$	…	$x_{1n}^{(1)}$
	…	…	…	…	…
	i　$X_i^{(1)}$	$x_{i1}^{(1)}$	$x_{i2}^{(1)}$	…	$x_{in}^{(1)}$
	…	…	…	…	…
	p　$X_p^{(1)}$	$x_{p1}^{(1)}$	$x_{p2}^{(1)}$	…	$x_{pn}^{(1)}$
G_2 (i=1, …, q)	1　$X_1^{(2)}$	$x_{11}^{(2)}$	$x_{12}^{(2)}$	…	$x_{1n}^{(2)}$
	…	…	…	…	…
	i　$X_i^{(2)}$	$x_{i1}^{(2)}$	$x_{i2}^{(2)}$	…	$x_{in}^{(2)}$
	…	…	…	…	…
	q　$X_q^{(2)}$	$x_{q1}^{(2)}$	$x_{q2}^{(2)}$	…	$x_{qn}^{(2)}$

Peking University
Introduction to Computational Biology

STEP 1

先对样本点数据$X_i^{(1)}(x_{i1}^{(1)}, x_{i2}^{(1)}, \ldots, x_{in}^{(1)})$（$i$=1, …, p）、$X_i^{(2)}(x_{i1}^{(2)}, x_{i2}^{(2)}, \ldots, x_{in}^{(2)})$（$i$=1, …, q）分别计算以下求和以及平均值：

$$\sum_{k=1}^{p} x_{ki}^{(1)} \qquad \sum_{k=1}^{q} x_{ki}^{(2)} \qquad (i = 1,2,\ldots,n)$$

$$\bar{x}_i^{(1)} = \frac{1}{p}\sum_{k=1}^{p} x_{ki}^{(1)} \qquad \bar{x}_i^{(2)} = \frac{1}{q}\sum_{k=1}^{q} x_{ki}^{(2)} \qquad (i = 1,2,\ldots,n)$$

$$\sum_{k=1}^{p} x_{ki}^{(1)} x_{kj}^{(1)} \qquad \sum_{k=1}^{q} x_{ki}^{(2)} x_{kj}^{(2)} \qquad (i, j = 1,2,\ldots,n)$$

STEP 2

计算d_i和S_{ij}，注意对称性$S_{ij}=S_{ji}$：

$$d_i = \bar{x}_i^{(1)} - \bar{x}_i^{(2)} \qquad (i=1,2,...,n)$$

$$S_{ij} = \sum_{k=1}^{p}(x_{ki}^{(1)} - \bar{x}_i^{(1)})(x_{kj}^{(1)} - \bar{x}_j^{(1)}) + \sum_{k=1}^{q}(x_{ki}^{(2)} - \bar{x}_i^{(2)})(x_{kj}^{(2)} - \bar{x}_j^{(2)})$$

$$= \sum_{k=1}^{p} x_{ki}^{(1)} x_{kj}^{(1)} - \frac{1}{p}\left(\sum_{k=1}^{p} x_{ki}^{(1)}\right)\left(\sum_{k=1}^{p} x_{kj}^{(1)}\right)$$

$$+ \sum_{k=1}^{q} x_{ki}^{(2)} x_{kj}^{(2)} - \frac{1}{q}\left(\sum_{k=1}^{q} x_{ki}^{(2)}\right)\left(\sum_{k=1}^{q} x_{kj}^{(2)}\right)$$

$$(i, j = 1,2,...,n)$$

STEP 3

解线性代数方程组：

$$S_{11}C_1 + S_{12}C_2 + ... + S_{1n}C_n = d_1$$
$$S_{21}C_1 + S_{22}C_2 + ... + S_{2n}C_n = d_2$$
$$.....$$
$$S_{n1}C_1 + S_{n2}C_2 + ... + S_{nn}C_n = d_n$$

若方程有解，得到判别函数F：

$$F(x_1, x_2, ..., x_n) = C_1x_1 + C_2x_2 + ... + C_nx_n$$

STEP 4

将平均值代入判别函数，然后计算判别值C：

$$\bar{y}^{(1)} = C_1\bar{x}_1^{(1)} + C_2\bar{x}_2^{(1)} + \ldots + C_n\bar{x}_n^{(1)}$$

$$\bar{y}^{(2)} = C_1\bar{x}_1^{(2)} + C_2\bar{x}_2^{(2)} + \ldots + C_n\bar{x}_n^{(2)}$$

$$C = \frac{p\bar{y}^{(1)} + q\bar{y}^{(2)}}{p+q}$$

Peking University
Introduction to Computational Biology

STEP 5

对未知数据$X(x_1, x_2, \ldots, x_n)$进行判别：将数据$X(x_1, x_2, \ldots, x_n)$代入判别函数F，与判别值进行比较，判别其属于哪一类。

$$y = C_1x_1 + C_2x_2 + \ldots + C_nx_n$$

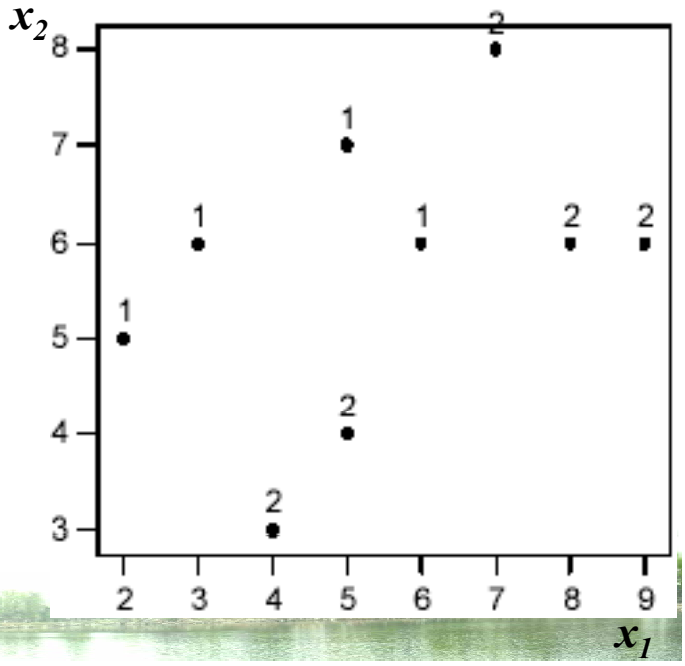
若 $\bar{y}^{(1)} > \bar{y}^{(2)}$,
$y > C$， X属于 G_1
$y \le C$， X属于 G_2

若 $\bar{y}^{(1)} < \bar{y}^{(2)}$,
$y > C$， X属于 G_2
$y \le C$， X属于 G_1

4、Fisher线性判别的应用举例

DNA序列上ORF是否编码蛋白的判别：依据ORF长度和密码子频率两个属性进行打分，得到x_1, x_2，在(x_1 , x_2)平面上进行Fisher线性判别分析。

ORF序号	x_1	x_2	类别
1	0.5	0.7	1(编码)
2	0.4	0.3	2 (非编码)
3	0.7	0.8	2 (非编码)
4	0.8	0.6	2 (非编码)
5	0.3	0.6	1 (编码)
6	0.2	0.5	1 (编码)
7	0.6	0.6	1 (编码)
8	0.9	0.6	2 (非编码)
9	0.5	0.4	2 (非编码)



Peking University
Introduction to Computational Biology

$$\bar{x}^{(1)} = \begin{pmatrix} 4.0 \\ 6.0 \end{pmatrix} \qquad \bar{x}^{(2)} = \begin{pmatrix} 6.6 \\ 5.4 \end{pmatrix}$$

$$d_1 = -2.6 \qquad d_2 = 0.6$$

$$S = \begin{pmatrix} 3.8857 & 2.1143 \\ 2.1143 & 2.4571 \end{pmatrix}$$

$$3.8857C_1 + 2.1143C_2 = -2.6$$
$$2.1143C_1 + 2.4571C_2 = 0.6$$

$$C_1 = -1.5080 \qquad C_2 = 1.5418$$

$$y = -1.5080\ x_1 + 1.5418\ x_2$$

$$\bar{x}^{(1)} = \begin{pmatrix} 4.0 \\ 6.0 \end{pmatrix} \qquad \bar{x}^{(2)} = \begin{pmatrix} 6.6 \\ 5.4 \end{pmatrix}$$

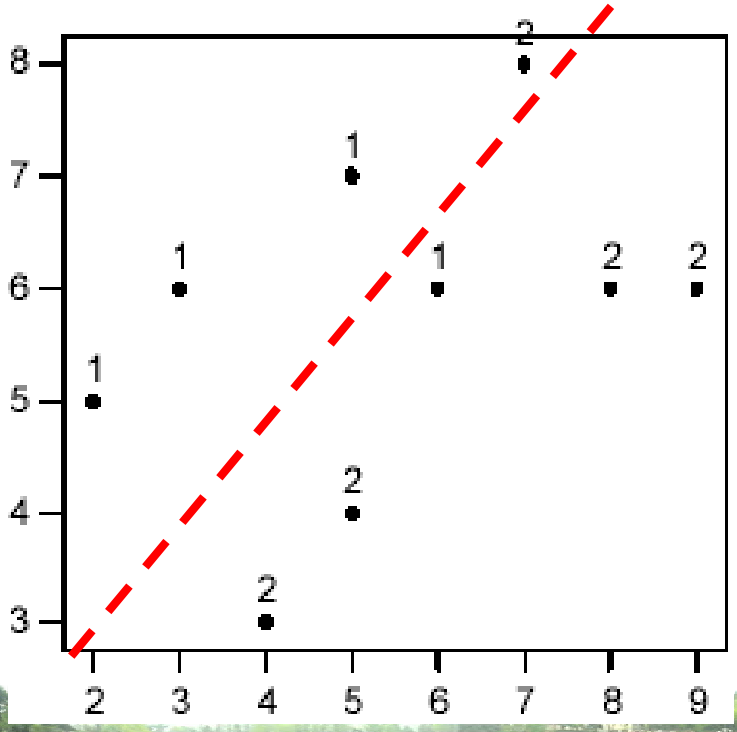
$$\bar{y}^{(1)} = -1.5080\ \bar{x}_1^{(1)} + 1.5418\ \bar{x}_2^{(1)}$$

$$\bar{y}^{(2)} = -1.5080\ \bar{x}_1^{(2)} + 1.5418\ \bar{x}_2^{(2)}$$

$$C = 0.5264$$

Peking University
Introduction to Computational Biology

$$-1.5080\ x_1 + 1.5418\ x_2 = 0.5264$$



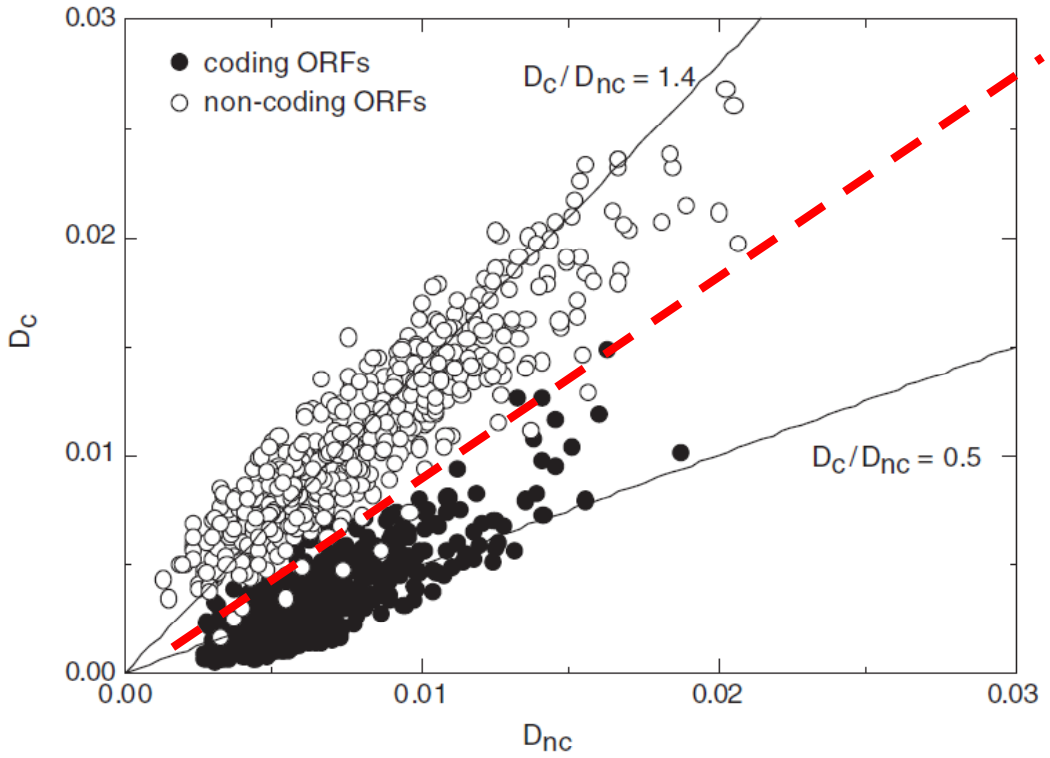
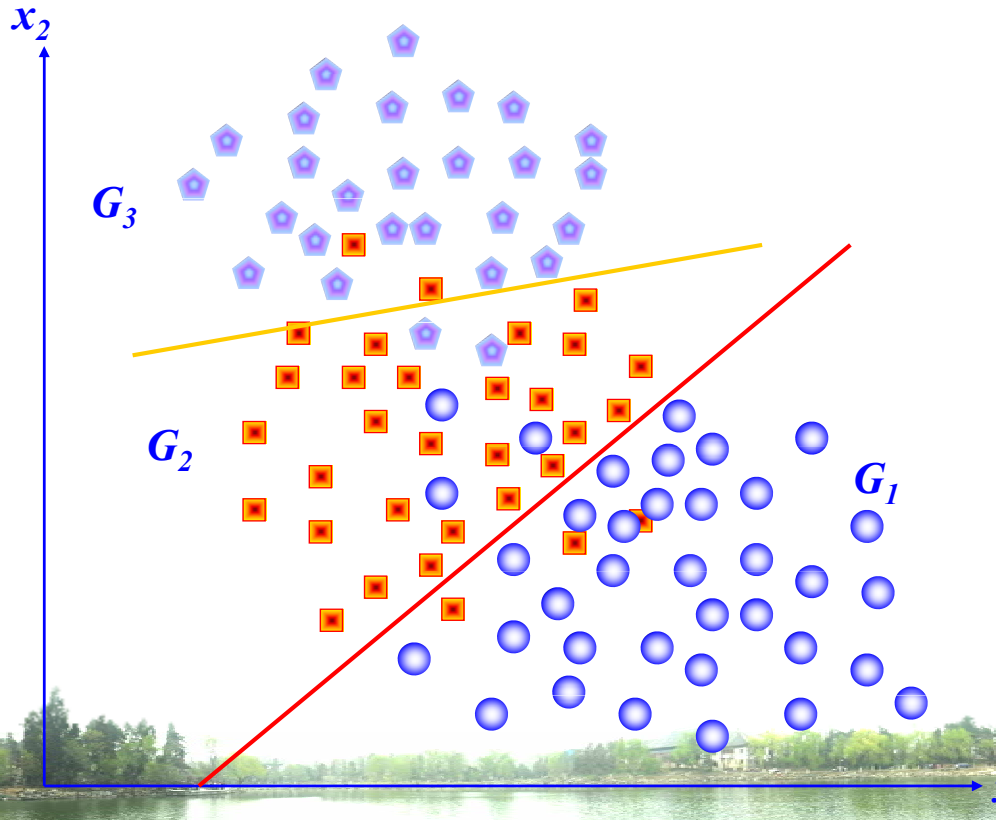


Fig. 2. D_c vs. D_{nc} plots of coding and non-coding ORFs for 500 randomly selected coding ORFs (filled circles) and non-coding ORFs (open circles) in *E. coli* genome.

Peking University
Introduction to Computational Biology

5、逐级Fisher判别法



6、Fisher判别法小结

- 基本思想：投影。使得投影后各组尽可能分开。
- 本质上是基于微分寻优的方法
- 局限

1、可能陷入局部最优的判别结果；

2、对数据属性各变量的要求较为苛刻，如正态性、相互独立性等；

3、对于类别数目太多的判别问题，采用逐级判别比较麻烦、累赘。

Peking University
Introduction to Computational Biology

6.2.2 距离判别法

距离判别的基本思想

样品与哪一类总体的距离最近，就判别它属于哪一类总体。

距离的定义

绝对距离
相对距离

1、马氏（Mahalanobis）距离

定义：Mahalanobis距离

设总体G为n维变量，即含有n个属性指标$(x_1, x_2, \ldots, x_n)$。已知总体G中的t个样品$X_k\,(x_{k1}, x_{k2}, \ldots, x_{kn})$，$k=1, 2, \ldots, t$。总体均值可用样本均值估计：

$$\bar{x}_i = \frac{1}{t}\sum_{k=1}^{t} x_{ki} \qquad i = 1,2,\ldots, n$$

则对于任一点$X(x_1, x_2, \ldots, x_n)$，定义点X与总体G的Mahalanobis距离为：

$$d^2(X, G) = (X - \bar{X})' S^{-1} (X - \bar{X})$$

Peking University
Introduction to Computational Biology

其中，矩阵$S=(s_{ij})_{n\times n}$为：

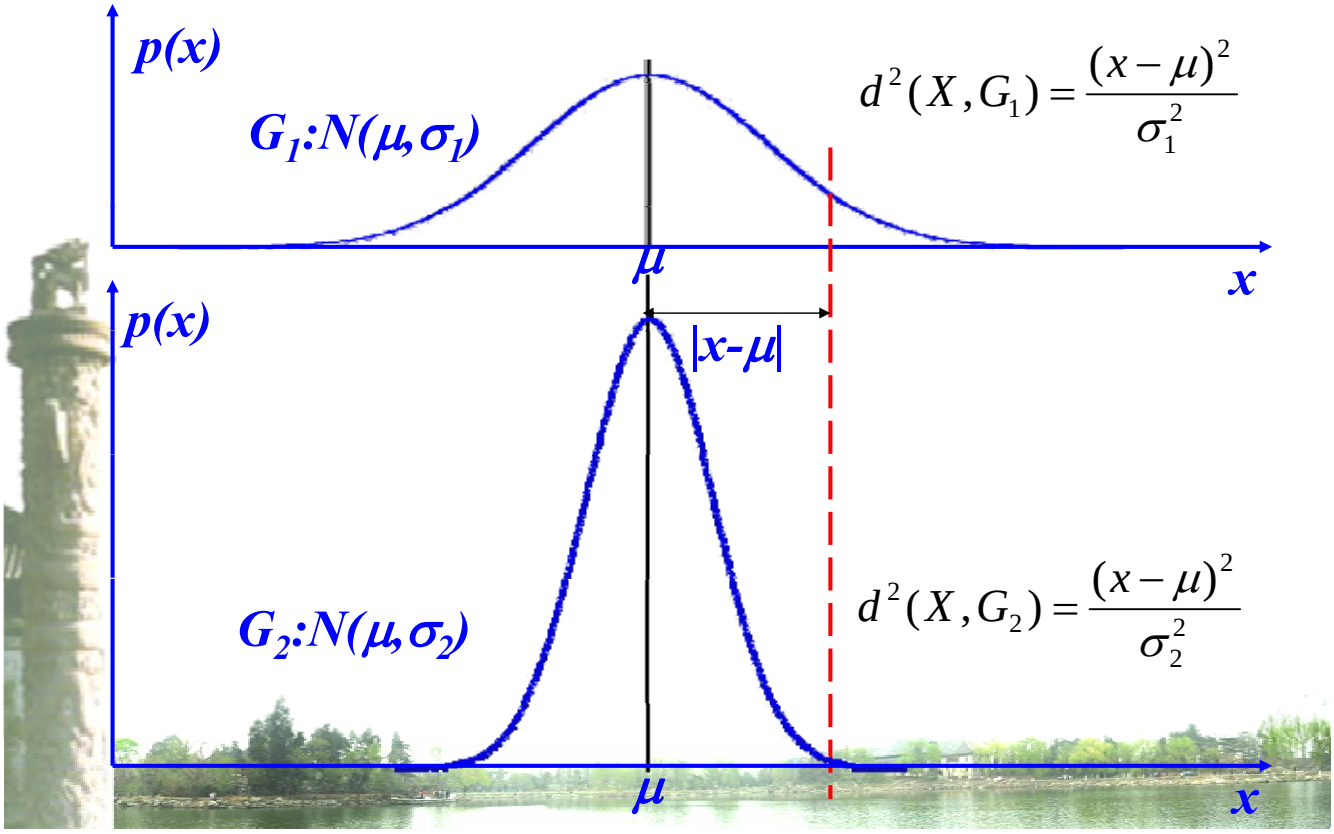
$$s_{ij} = \frac{1}{t-1}\sum_{k=1}^{t} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \qquad i, j = 1,2,\ldots, n$$

矩阵S称为协方差矩阵（covariance matrix），反映总体G的属性指标中第i个分量与第j个分量的相关性。

特别地，当n=1时，假定总体G符合正态分布， Mahalanobis距离为：

$$d^2(X, G) = \frac{(x-\mu)'(x-\mu)}{\sigma^2} = \frac{(x-\mu)^2}{\sigma^2}$$

绝对距离与Mahalanobis距离的比较



Peking University
Introduction to Computational Biology

P. C. Mahalanobis
(1893–1972)

Indian scientist and applied statistician. He is best remembered for the Mahalanobis distance, a statistical measure. He made pioneering studies in anthropometry in India. He founded the *Indian Statistical Institute*, and contributed to the design of large scale sample surveys.

2、两类总体的Mahalanobis距离判别方法

已知：考虑具有n个属性的两类总体G_1、G_2，已知G_1的p个训练样本，G_2的q个训练样本：

$$\begin{bmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1n}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2n}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p1}^{(1)} & x_{p2}^{(1)} & \cdots & x_{pn}^{(1)} \end{bmatrix} \quad \begin{bmatrix} x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1n}^{(2)} \\ x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2n}^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_{q1}^{(2)} & x_{q2}^{(2)} & \cdots & x_{qn}^{(2)} \end{bmatrix}$$

问题：对于未知样本点$X(x_1, x_2, \ldots, x_n)$，判别其类型？

Peking University
Introduction to Computational Biology

G_1、G_2的总体均值根据样本均值估计得到：

$$\bar{x}_i^{(1)} = \frac{1}{p}\sum_{k=1}^{p} x_{ki}^{(1)} \qquad \bar{x}_i^{(2)} = \frac{1}{q}\sum_{k=1}^{q} x_{ki}^{(2)} \qquad i = 1,2,\ldots,n$$

分别求出总体G_1、G_2的协方差矩阵$S^{(1)}$、$S^{(2)}$：

$$s_{ij}^{(1)} = \frac{1}{p-1}\sum_{k=1}^{p}(x_{ki}^{(1)} - \bar{x}_i^{(1)})(x_{kj}^{(1)} - \bar{x}_j^{(1)}) \qquad i, j = 1,2,\ldots,n$$

$$s_{ij}^{(2)} = \frac{1}{q-1}\sum_{k=1}^{q}(x_{ki}^{(2)} - \bar{x}_i^{(2)})(x_{kj}^{(2)} - \bar{x}_j^{(2)}) \qquad i, j = 1,2,\ldots,n$$

对于任一新样本$X(x_1, x_2, \ldots, x_n)$，分别计算它到总体G_1、G_2的Mahalanobis距离：

$$d^2(X, G_1) = (x_1 - \overline{x}_1^{(1)}, x_2 - \overline{x}_2^{(1)}, \ldots, x_n - \overline{x}_n^{(1)}) \cdot \begin{pmatrix} s_{11}^{(1)} & s_{12}^{(1)} & \ldots & s_{1n}^{(1)} \\ s_{21}^{(1)} & s_{22}^{(1)} & \ldots & s_{2n}^{(1)} \\ \ldots & \ldots & \ldots & \ldots \\ s_{n1}^{(1)} & s_{n2}^{(1)} & \ldots & s_{nn}^{(1)} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \overline{x}_1^{(1)} \\ x_2 - \overline{x}_2^{(1)} \\ \ldots \\ x_n - \overline{x}_n^{(1)} \end{pmatrix}$$

$$d^2(X, G_2) = (x_1 - \overline{x}_1^{(2)}, x_2 - \overline{x}_2^{(2)}, \ldots, x_n - \overline{x}_n^{(2)}) \cdot \begin{pmatrix} s_{11}^{(2)} & s_{12}^{(2)} & \ldots & s_{1n}^{(2)} \\ s_{21}^{(2)} & s_{22}^{(2)} & \ldots & s_{2n}^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ s_{n1}^{(2)} & s_{n2}^{(2)} & \ldots & s_{nn}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \overline{x}_1^{(2)} \\ x_2 - \overline{x}_2^{(2)} \\ \ldots \\ x_n - \overline{x}_n^{(2)} \end{pmatrix}$$

Peking University
Introduction to Computational Biology

构造判别函数$W(X)$：

$$W(X) = d^2(X, G_2) - d^2(X, G_1)$$

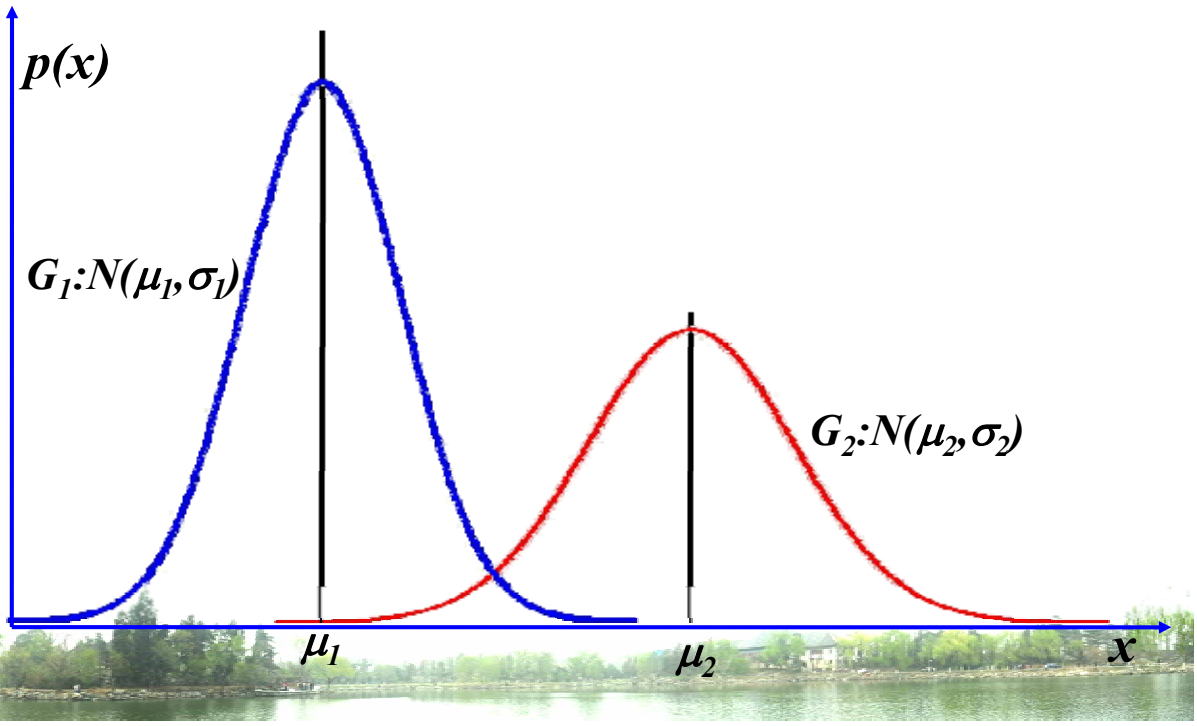
判别准则为：

$$W(X) > 0\text{时}, \quad X \in G_1$$
$$W(X) \leq 0\text{时}, \quad X \in G_2$$

特例：考虑n=1的两类正态总体：

G_1：N(μ_1, σ_1)

G_2：N(μ_2, σ_2)



Peking University
Introduction to Computational Biology

$$d(X,G_1)=\frac{|x-\mu_1|}{\sigma_1} \qquad d(X,G_2)=\frac{|x-\mu_2|}{\sigma_2}$$

不妨设$\mu_2 > \mu_1$， $\sigma_2 > \sigma_1$，且检测值满足$\mu_2 > x > \mu_1$，则：

$$W(x)=\frac{\mu_2-x}{\sigma_2}-\frac{x-\mu_1}{\sigma_1}=\frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}(\mu^*-x)$$

其中

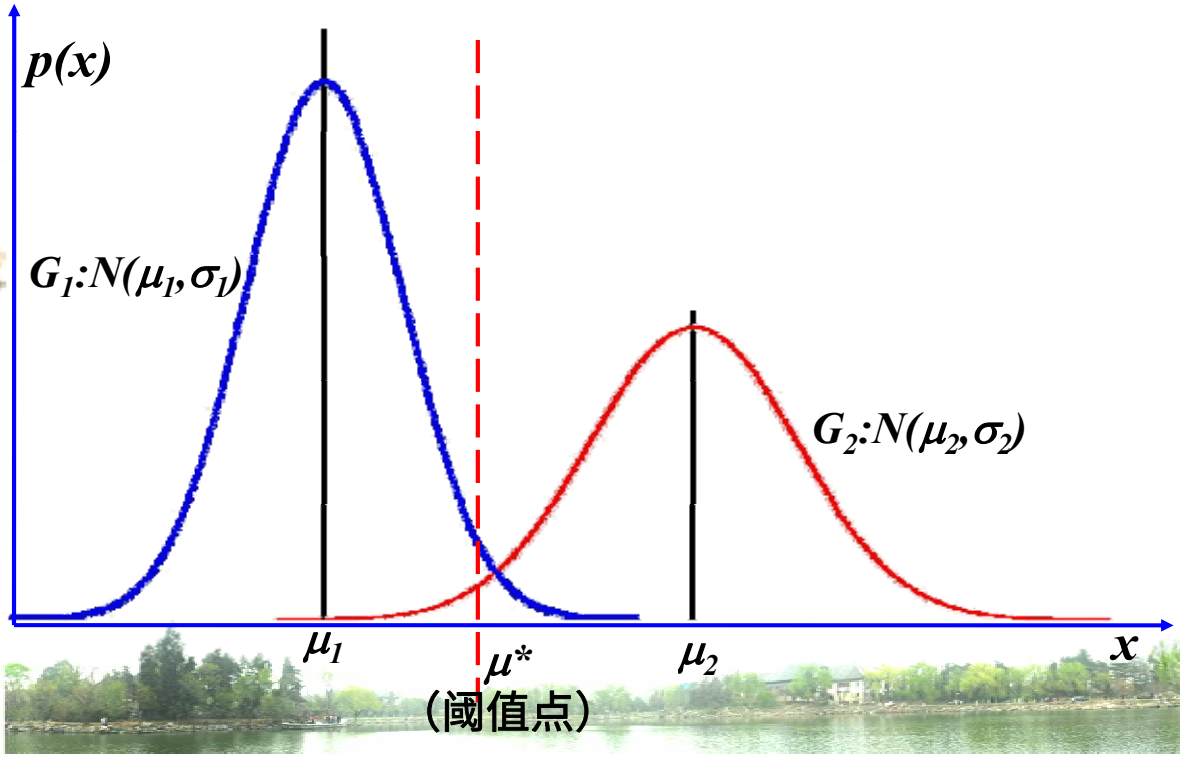
$$\mu^*=\frac{\sigma_2\mu_1+\sigma_1\mu_2}{\sigma_1\sigma_2}$$

于是，判别准则为：

$W(x) > 0$时， $x \in G_1$

$W(x) \le 0$时， $x \in G_2$

$$d(X,G_1)=\frac{|x-\mu_1|}{\sigma_1} \qquad d(X,G_2)=\frac{|x-\mu_2|}{\sigma_2}$$



Peking University
Introduction to Computational Biology

3、多类总体的Mahalanobis距离判别方法

已知：考虑具有n个属性的m类总体G_l（$l=1, 2, \ldots, m$），每类总体已知t_l（$l=1, 2, \ldots, m$)个训练样本：

$$\begin{bmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1n}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2n}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ x_{t_1 1}^{(1)} & x_{t_1 2}^{(1)} & \cdots & x_{t_1 n}^{(1)} \end{bmatrix} \begin{bmatrix} x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1n}^{(2)} \\ x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2n}^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_{t_2 1}^{(2)} & x_{t_2 2}^{(2)} & \cdots & x_{t_2 n}^{(2)} \end{bmatrix} \cdots \begin{bmatrix} x_{11}^{(m)} & x_{12}^{(m)} & \cdots & x_{1n}^{(m)} \\ x_{21}^{(m)} & x_{22}^{(m)} & \cdots & x_{2n}^{(m)} \\ \cdots & \cdots & \cdots & \cdots \\ x_{t_m 1}^{(m)} & x_{t_m 2}^{(m)} & \cdots & x_{t_m n}^{(m)} \end{bmatrix}$$

问题：对于未知样本点$X(x_1, x_2, \ldots, x_n)$，判别其类型？

类似地，分别计算点$X(x_1, x_2, \ldots, x_n)$到每一类G_l的Mahalanobis距离$d^2(X, G_l)$。

$$d^2(X, G_l) = \begin{bmatrix} x_1 - \bar{x}_1^{(l)} & x_2 - \bar{x}_2^{(l)} & \ldots & x_n - \bar{x}_n^{(l)} \end{bmatrix} \cdot \begin{bmatrix} s_{11}^{(l)} & s_{12}^{(l)} & \ldots & s_{1n}^{(l)} \\ s_{21}^{(l)} & s_{22}^{(l)} & \ldots. & s_{2n}^{(l)} \\ \ldots & \ldots & \ldots & \ldots \\ s_{n1}^{(l)} & s_{n2}^{(l)} & \ldots & s_{nn}^{(l)} \end{bmatrix}^{-1} \cdot \begin{bmatrix} x_1 - \bar{x}_1^{(l)} \\ x_2 - \bar{x}_2^{(l)} \\ \ldots \\ x_n - \bar{x}_n^{(l)} \end{bmatrix}$$

其中

$$\bar{x}_i^{(l)} = \frac{1}{t_l} \sum_{k=1}^{t_l} x_{ki}^{(l)} \qquad i = 1,2,\ldots, n$$

$$s_{ij}^{(l)} = \frac{1}{t_l - 1} \sum_{k=1}^{t_l} (x_{ki}^{(l)} - \bar{x}_i^{(l)})(x_{kj}^{(l)} - \bar{x}_j^{(l)}) \qquad i, j = 1,2,\ldots, n$$

Peking University
Introduction to Computational Biology

比较找到其中的最小距离：

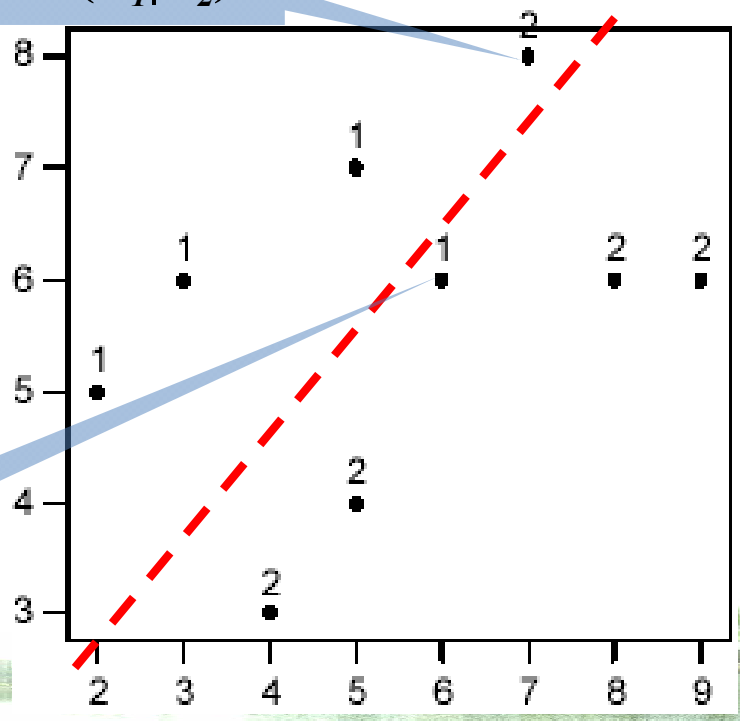
$$d^2(X, G_i) = \min_{l=1,2,\ldots,m} \left\{ d^2(X, G_l) \right\}$$

点$X(x_1, x_2, \ldots, x_n)$到类G_i的距离$d^2(X, G_i)$最小，最后判别点$X(x_1, x_2, \ldots, x_n)$属于第 i 类。

6.2.3 判别效果的评价

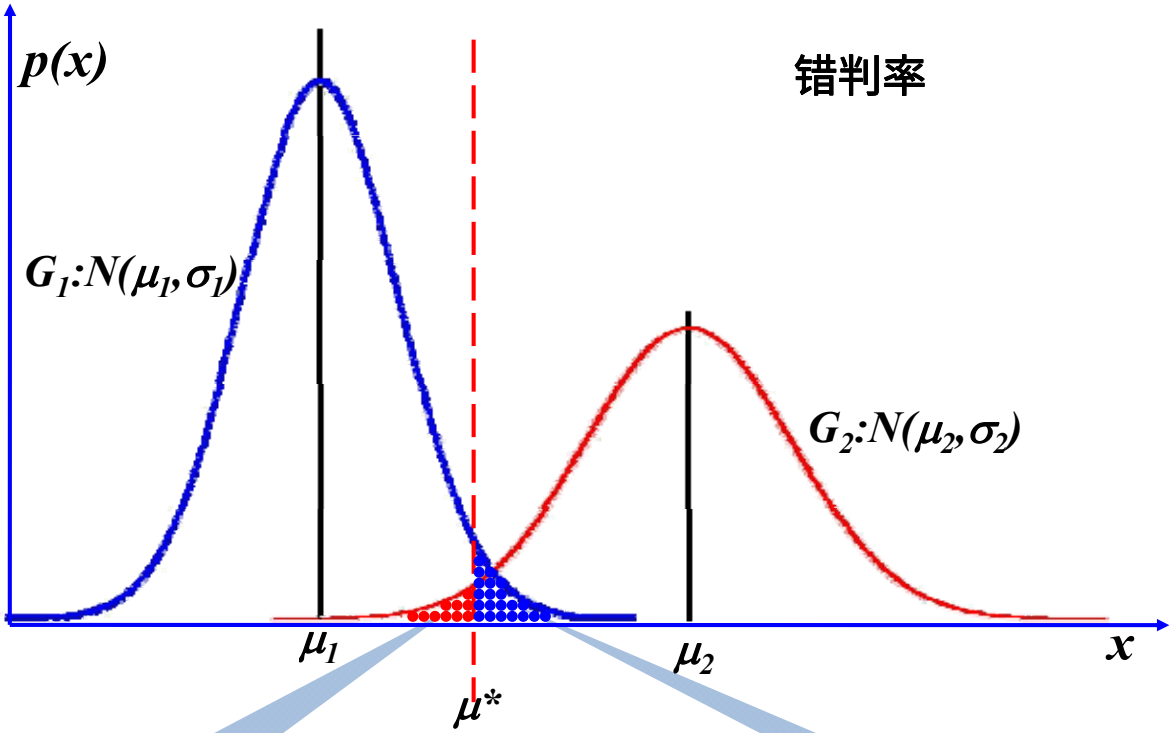
$N(G_1|G_2)$

错判损失



$N(G_2|G_1)$

Peking University
Introduction to Computational Biology



1、检验判别效果的方法

训练集（Learning set）：训练样本集
检测集（Test set）：检测样本集（类别未知）

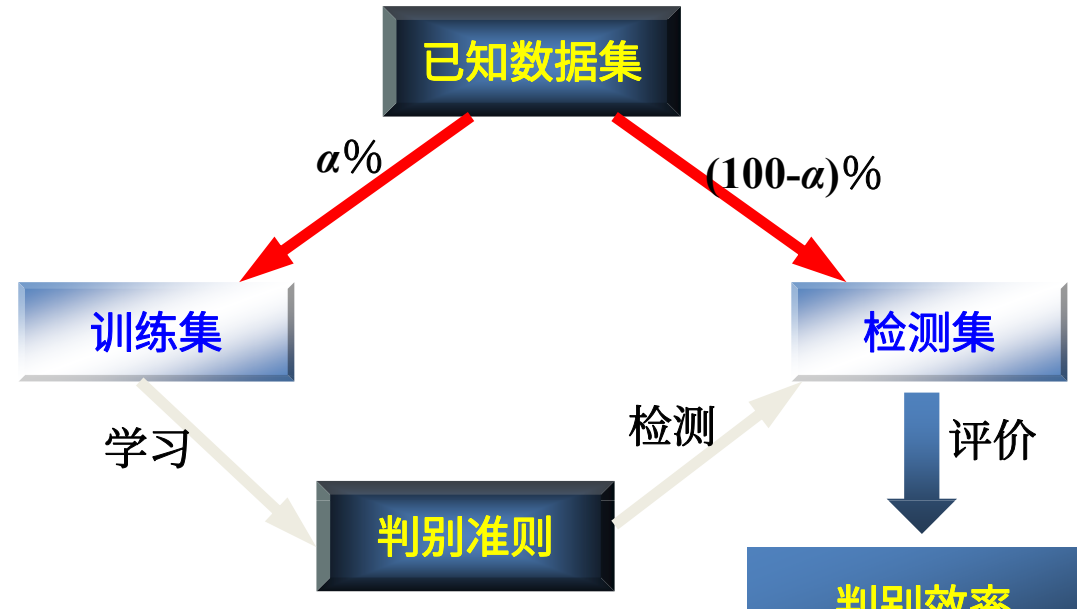
（1）训练集的回判

利用训练集作为检测集：用判别方法对已知类型的样本进行回判，统计判错的个数以及占样本总数的比例，作为错判率的估计。

特点：低估错判率。

Peking University
Introduction to Computational Biology

（2）从训练集中构造检测集



（3）刀切法（Jack-knife Method）

“舍一法（Leaveone-out）”
“Lachenbruch删除法”
“交叉确认法(Cross-validation)”

基本思想：
(1). 每次从训练样本集中剔除1个样本X'；
(2). 利用其余的样本（数量为p＋q－1）作为训练集来训练得到判别准则;
(3). 根据判别准则对样本X'进行判别；
(4). 对训练样本中的每个样本依次重复进行，记录判别对错的个数；
(5). 计算错判率。

Peking University
Introduction to Computational Biology

2、检验判别效果的几个指标

检测结果 / 实际归类	Predicted G_1	Predicted G_2	合计
Real G_1	$N(G_1\|G_1)$	$N(G_2\|G_1)$	N^{real}_1
Real G_2	$N(G_1\|G_2)$	$N(G_2\|G_2)$	N^{real}_2
合计	N^{pred}_1	N^{pred}_2	

<u>错判率</u>（貌似错判率）：

$$\frac{N(G_1|G_2)+N(G_2|G_1)}{N_1^{real}+N_2^{real}}$$

敏感性（sensitivity，Sn）：

$$Sn^{(1)} = \frac{N(G_1|G_1)}{N_1^{real}} \qquad Sn^{(2)} = \frac{N(G_2|G_2)}{N_2^{real}}$$

特异性（specificity，Sp）：

$$Sp^{(1)} = \frac{N(G_1|G_1)}{N_1^{pred}} \qquad Sp^{(2)} = \frac{N(G_2|G_2)}{N_2^{pred}}$$

Peking University
Introduction to Computational Biology

检验判别效果举例：
真核基因内含子剪接位点信号的识别

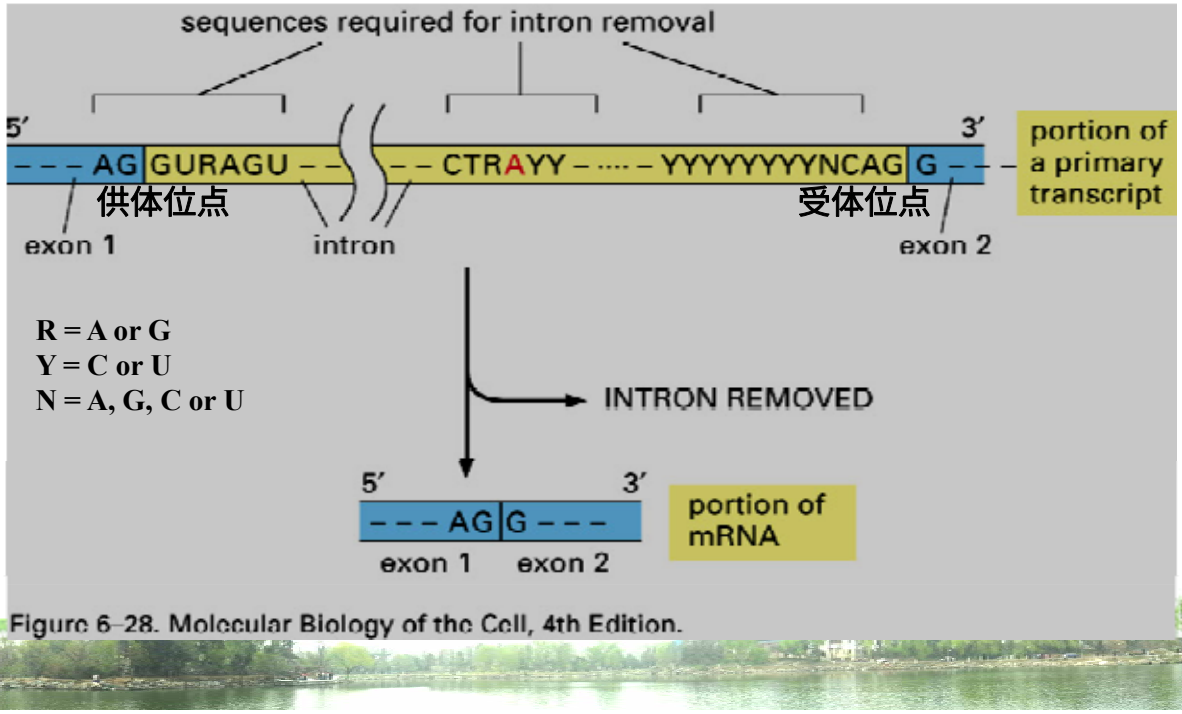


Figure 6–28. Molecular Biology of the Cell, 4th Edition.

供体信号GU识别结果：

	被识别为真信号	被识别为假信号
2832个真信号	2738	94
136422个假信号	12668	123754

检测结果 / 实际归类	Predicted G_1	Predicted G_2	合计
Real G_1	$N(G_1\|G_1)$=2,738	$N(G_2\|G_1)$=94	N^{real}_1 =2,832
Real G_2	$N(G_1\|G_2)$=12,668	$N(G_2\|G_2)$=123,754	N^{real}_2 =136,422
合计	N^{pred}_1=15,406	N^{pred}_2 =123,848	

貌似错判率：9.2%

敏感性Sn ：96.7%　　特异性Sp ：17.8%

受体信号AG识别结果：

	被识别为真信号	被识别为假信号
2832个真信号	2692	140
194409个假信号	22212	172197

貌似错判率：11.3%　　Sn = 95.1%　　Sp = 10.8%

Peking University
Introduction to Computational Biology

§6.3

聚类分析方法
(Clustering method)

分类问题

- 条件：已知研究对象总体的类别数目及其特征（如：分布规律，或各类的训练样本）
- 目的：判断未知类别的样本的归属类别

→ 判别分析 (Discriminant Analysis)

- 条件：研究对象总体的类别数目未知，也不知总体样本的具体分类情况
- 目的：通过分析，选定描述个体相似程度的统计量、确定总体分类数目、建立分类方法；对研究对象给出合理的分类。(“物以类聚”是聚类分析的基本出发点)

→ 聚类分析 (Clustering Analysis)

Peking University
Introduction to Computational Biology

概　　述

- 聚类分析：（群分析）

是实用多元统计分析的一个新分支，正处于发展阶段。理论上尚未完善，但应用十分广泛。

实质上是一种分类问题，目的是建立一种分类方法，将一批数据按照特征的亲疏、相似程度进行分类。

- 定性、经验的分类的局限

分类较粗、数据量小、凭借经验

- 生物信息学中的聚类分析问题：

根据DNA芯片获得的基因表达数据进行基因聚类（数据量庞大）
蛋白质相互作用网络的分类
根据不同物种的大分子序列进行相似性比较并构建系统发育树
……

聚类分析

建立合适的分类方法：
(1). 将一批样本按照亲疏程度进行分类（Q型聚类）
(2). 将样本的多个变量按照相似程度进行分类（R型聚类）

- 系统聚类法 (谱系聚类法)
- 动态聚类法
- 最优分割法 (有序样本聚类法)
- 模糊聚类法
- 图论聚类法

Peking University
Introduction to Computational Biology

6.3.1 分类统计量——距离、相似系数

1、数据的变换方法

数据变换的目的：使不同的量纲、不同取值范围的数据能放在一起比较。

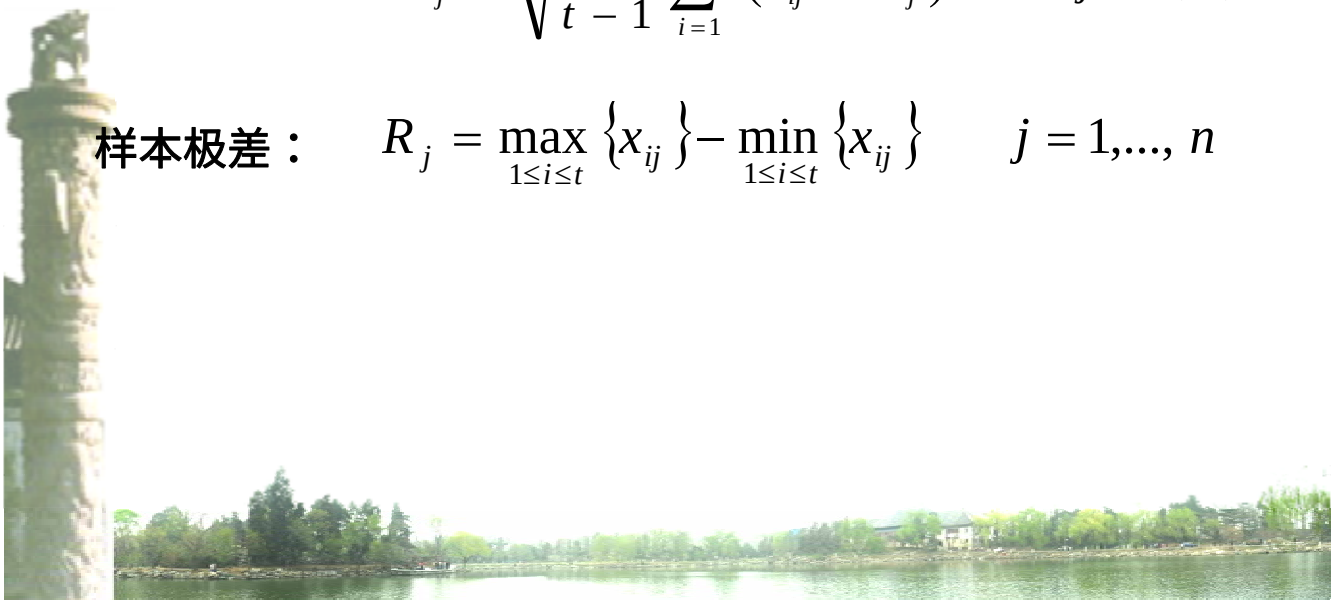
已知样品数目为t，每个样品测得n项属性指标，得到观察数据x_{ij} (i =1, … ,t; j=1,…,n)。

	X_1	…	X_j	…	X_n
$X_{(1)}$	x_{11}	…	x_{1j}	…	x_{1n}
…	…		…		
$X_{(i)}$	x_{i1}	…	x_{ij}	…	x_{in}
…	…		…		…
$X_{(t)}$	x_{t1}	…	x_{tj}	…	x_{tn}

样本均值： $\bar{x}_j = \frac{1}{t}\sum_{i=1}^{t} x_{ij} \qquad j = 1,\ldots, n$

样本标准差： $s_j = \sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\left(x_{ij} - \bar{x}_j\right)^2} \qquad j = 1,\ldots, n$

样本极差： $R_j = \max_{1\le i\le t}\{x_{ij}\} - \min_{1\le i\le t}\{x_{ij}\} \qquad j = 1,\ldots, n$



Peking University
Introduction to Computational Biology

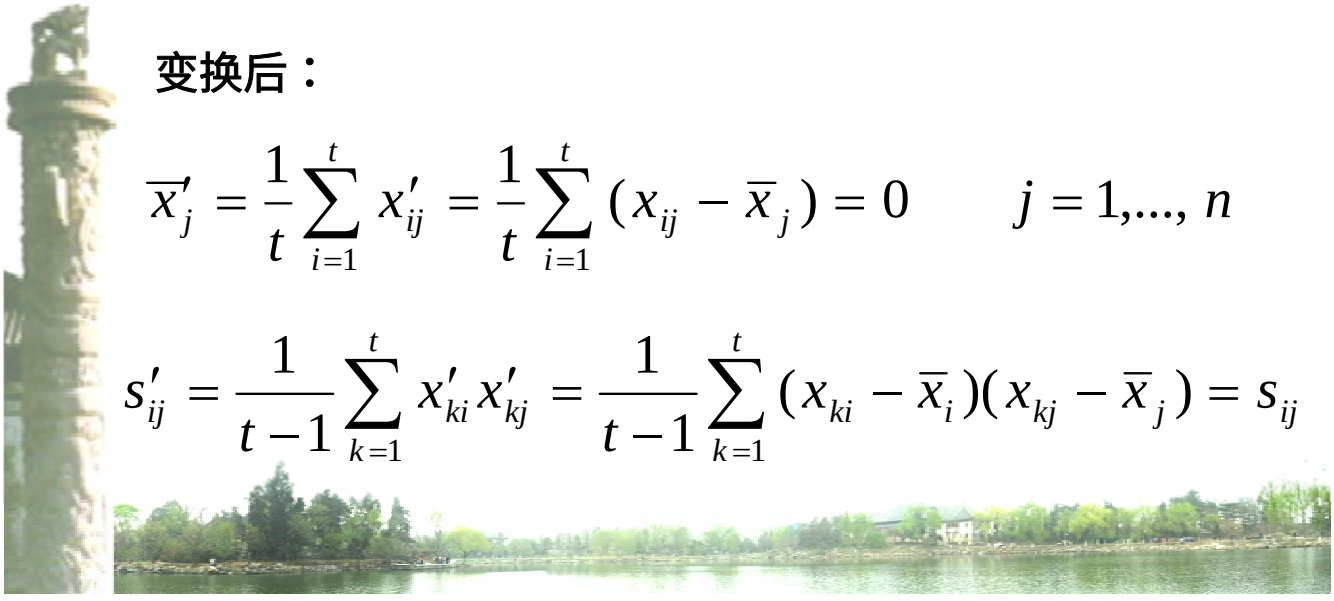
1）、中心化变换

$$x'_{ij} = x_{ij} - \bar{x}_j \qquad (i = 1,2,\ldots, t;\ j = 1,2,\ldots, n)$$

变换后：

$$\bar{x}'_j = \frac{1}{t}\sum_{i=1}^{t} x'_{ij} = \frac{1}{t}\sum_{i=1}^{t}(x_{ij} - \bar{x}_j) = 0 \qquad j = 1,\ldots, n$$

$$s'_{ij} = \frac{1}{t-1}\sum_{k=1}^{t} x'_{ki}x'_{kj} = \frac{1}{t-1}\sum_{k=1}^{t}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) = s_{ij}$$



2）、标准化变换

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad (i = 1,2,..., t;\ j = 1,2,..., n)$$

变换后：

$$\bar{x}'_j = \frac{1}{t}\sum_{i=1}^{t} x'_{ij} = 0 \qquad j = 1,..., n$$

$$s'_j = \sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\left(x'_{ij} - \bar{x}'_j\right)^2} = 1 \qquad j = 1,..., n$$

变换后的数据与变量的量纲无关。

Peking University
Introduction to Computational Biology

3）、极差标准化变换

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{R_j} \qquad (i = 1,2,..., t;\ j = 1,2,..., n)$$

变换后：

$$\bar{x}'_j = \frac{1}{t}\sum_{i=1}^{t} x'_{ij} = 0 \qquad j = 1,..., n$$

$$\left|x'_{ij}\right| < 1$$

$$R'_j = 1 \qquad j = 1,..., n$$

变换后的数据与变量的量纲无关。

4）、极差正规化变换（规格化变换）

$$x'_{ij} = \frac{x_{ij} - \min\limits_{1 \le k \le t} x_{kj}}{R_j} \qquad (i = 1, 2, \ldots, t;\ j = 1, 2, \ldots, n)$$

变换后：

$$0 \le \left| x'_{ij} \right| \le 1$$

$$R'_j = 1 \qquad j = 1, \ldots, n$$

变换后的数据与变量的量纲无关。

Peking University
Introduction to Computational Biology

5）、对数变换

$$x'_{ij} = \log(x_{ij})$$

$$(x_{ij} > 0; i = 1, 2, \ldots, t;\ j = 1, 2, \ldots, n)$$

可将具有指数特征的数据结构化为线性数据结构。

2、样品间的距离

距离的定义

用d_{ij}表示样品$X_{(i)}$与$X_{(j)}$之间的距离，有

- $d_{ij} \geq 0$; 且$d_{ij} = 0 \Leftrightarrow X_{(i)} = X_{(j)}$;
- $d_{ij} = d_{ji}$;
- 三角不等式：$d_{ij} \leq d_{ik} + d_{kj}$

已知样品数目为t，每个样品测得n项属性指标，得到观察数据x_{ij} (i =1, … ,t; j=1,…,n)。

	X_1	…	X_j	…	X_n
$X_{(1)}$	x_{11}	…	x_{1j}	…	x_{1n}
…	…		…		
$X_{(i)}$	x_{i1}	…	x_{ij}	…	x_{in}
…	…		…		…
$X_{(t)}$	x_{t1}	…	x_{tj}	…	x_{tn}

Peking University
Introduction to Computational Biology

1）、Minkowski距离

$$d_{ij}(q) = \left[\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^{q} \right]^{\frac{1}{q}} \qquad (i, j = 1,..., t)$$

绝对值距离

q=1时，得到一阶Minkowski度量：

$$d_{ij}(1) = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right| \qquad (i, j = 1,..., t)$$

称为绝对值距离。

欧氏距离

q=2时，得到二阶Minkowski度量：

$$d_{ij}(2) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \qquad (i, j = 1,..., t)$$

称为欧氏距离。
欧氏距离是聚类分析中使用最为广泛的距离，与各变量的量纲有关。

Chebyshev距离

$$d_{ij}(\infty) = \max_{1 \le k \le n} |x_{ik} - x_{jk}| \qquad (i, j = 1,..., t)$$

Peking University
Introduction to Computational Biology

2）、Lance距离

$$d_{ij}(L) = \frac{1}{n} \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}} \qquad (i, j = 1,..., t; x_{ij} > 0)$$

Lance距离是无量纲的距离

对大的奇异值不敏感，适合处理高度偏倚的数据

没有考虑变量间的相关性

3）、Mahalanobis距离

$$d_{ij}(M) = (X_{(i)} - X_{(j)})' S^{-1} (X_{(i)} - X_{(j)})$$
$$(i, j = 1,..., t)$$

无量纲的距离

考虑变量间的相关性

Peking University
Introduction to Computational Biology

4）、斜交空间距离

$$d_{ij}(\gamma) = \left[\frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} (x_{ik} - x_{jk})(x_{il} - x_{jl}) \gamma_{kl} \right]^{\frac{1}{2}}$$
$$(i, j = 1,..., t)$$

称为斜交空间距离。

其中γ_{kl}是变量X_k与X_l之间的相关系数（即表示变量的夹角）。

$$\gamma_{kl} = \frac{\sum_{i=1}^{t} (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)}{\sqrt{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2} \sqrt{\sum_{i=1}^{n} (x_{il} - \bar{x}_l)^2}} \qquad (k, l = 1,..., n)$$

3、变量间的相似系数

相似系数

用c_{kl}表示变量X_l与X_k之间的相似系数（相关程度的大小），有

- $c_{kl}=\pm 1 \Leftrightarrow X_{(k)}=\alpha X_{(l)}$， $\alpha\neq 0$，常数；
- 对一切k, l都有：$|c_{kl}|\leq 1$；
- 对一切k, l都有： $c_{kl}=c_{lk}$

已知样本数目为t，每个样本测得n项属性指标，得到观察数据x_{ij} (i =1, … ,t; j=1,…,n)。

	X_1	…	X_j	…	X_n
$X_{(1)}$	x_{11}	…	x_{1j}	…	x_{1n}
…	…		…		
$X_{(i)}$	x_{i1}	…	x_{ij}	…	x_{in}
…	…		…		…
$X_{(t)}$	x_{t1}	…	x_{tj}	…	x_{tn}

1）、夹角余弦——相似系数

$$c_{ij}(1)=\cos\alpha_{ij}=\frac{\sum_{k=1}^{t}x_{ki}x_{kj}}{\sqrt{\sum_{k=1}^{n}x_{ki}^2}\sqrt{\sum_{k=1}^{n}x_{kj}^2}}$$

$$(i,\ j=1,...,\ n)$$

用两变量夹角来衡量二者的相似程度。

显然：

当 i = j 时，夹角α_{ij}=0，c_{ij}(1)=1，表明两变量完全相似；

当夹角α_{ij}=π／2， c_{ij}(1)=0，表明两变量正交，不相关。

2）、相关系数

对数据作标准化处理后的夹角余弦：

$$c_{ij}(2) = \gamma_{ij} = \frac{\sum_{k=1}^{t}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}$$

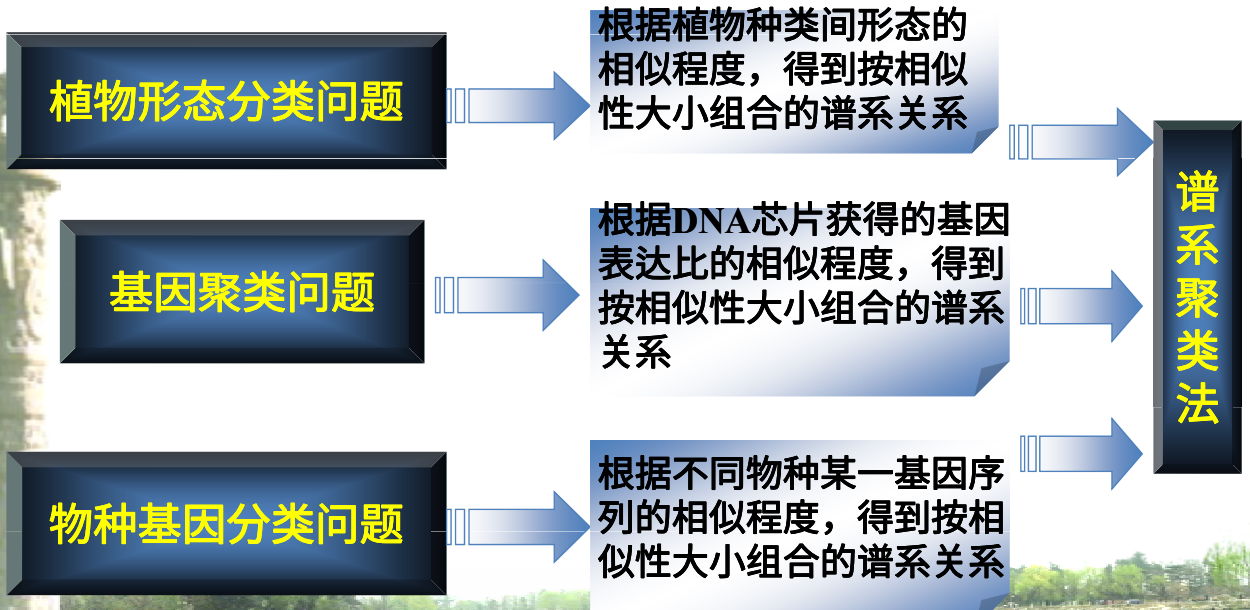
$$(i, j = 1,..., n)$$

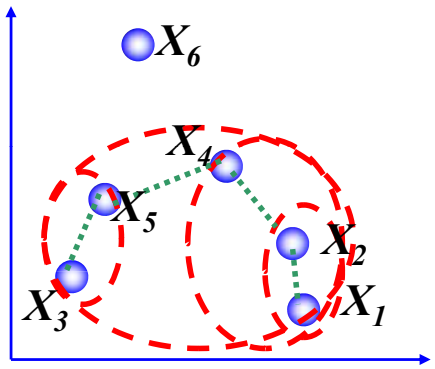
用两变量夹角来衡量二者的相似程度。

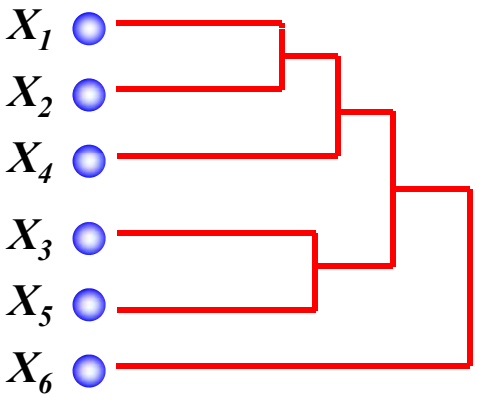
显然：
当 i = j 时，$c_{ij}(2)=1$，表明两变量完全相似；
$|c_{ij}(2)|\leq 1$

Peking University
Introduction to Computational Biology

6.3.2 聚类分析方法之一：谱系聚类法(hierarchical cluster analysis)







谱系关系图

Peking University
Introduction to Computational Biology

数学问题

条件：已知样本数目为t，每个样本测得n项属性指标，得到观察数据x_{ij} (i =1, … ,t; j=1,…,n)。

	X_1	…	X_j	…	X_n
$X_{(1)}$	x_{11}	…	x_{1j}	…	x_{1n}
…	…		…		
$X_{(i)}$	x_{i1}	…	x_{ij}	…	x_{in}
…	…		…		…
$X_{(t)}$	x_{t1}	…	x_{tj}	…	x_{tn}

目的：给出t个样本的谱系聚类关系。

1、谱系聚类法的基本思想和步骤

- 对数据进行变换；
- 定义样品间的距离（如欧氏距离）、类别之间的距离（如最短距离）；
- 首先将t个样品各自视为一类：得到初始的分类$G^{(1)}$（含有t类），计算t个样品两两之间的距离，它们等价于初始的类间距离，得到初始的距离矩阵$D^{(1)}$；
- 将距离最近的两类合并为一新类，得到新的分类$G^{(2)}$（含有t-1类），并计算新类与其它类的类间距离，得到新的类间距离矩阵$D^{(2)}$，再按照最小距离准则并类，得到$G^{(3)}$（含有t-2类）、$D^{(3)}$,…。直到所有样品都并成一类；
- 画出谱系聚类图，决定分类的个数及各类的成员。

Peking University
Introduction to Computational Biology

谱系聚类法举例

已知：根据5种灵长类动物肌粒蛋白的氨基酸序列比较，得到它们之间的距离矩阵（已经过数据变换处理）。

$X_{(1)}$：Gibbon（长臂猿）；
$X_{(2)}$：Symphalangus（猩猩）；
$X_{(3)}$：Human（人）；
$X_{(4)}$：Gorilla（大猩猩）；
$X_{(5)}$：Chimpanzee（黑猩猩）

	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
$X_{(1)}$	0	1	3.5	5	7
$X_{(2)}$		0	2.5	4	6
$X_{(3)}$			0	1.5	3.5
$X_{(4)}$				0	2
$X_{(5)}$					0

构造：
样品间距离——欧氏距离；
类间距离——最短距离；

Step 1

5个物种各自构成1类，得到5类，有：

初始分类$G^{(1)}=\{X_{(i)}\}(i=1, 2, 3, 4, 5)$
初始类别数目$m=5$
初始类间距离矩阵$D^{(1)}$

$D^{(1)}$

	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
$X_{(1)}$	0	1	3.5	5	7
$X_{(2)}$		0	2.5	4	6
$X_{(3)}$			0	1.5	3.5
$X_{(4)}$				0	2
$X_{(5)}$					0

$C_{(4)}$

Step 2

由$D^{(1)}$知，合并$X_{(1)}$和$X_{(2)}$为一新类$C_{(4)}=\{X_{(1)}, X_{(2)}\}$，有：

新的$G^{(2)}=\{X_{(3)}, X_{(4)}, X_{(5)}, C_{(4)}\}$
新的类别数目$m=4$
新的类间距离矩阵$D^{(2)}$

$D^{(2)}$

	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$C_{(4)}$
$X_{(3)}$	0	1.5	3.5	2.5
$X_{(4)}$		0	2	4
$X_{(5)}$			0	6
$C_{(4)}$				0

$C_{(3)}$

Step 3

由$D^{(2)}$知，合并$X_{(3)}$和$X_{(4)}$为一新类$C_{(3)}=\{X_{(3)}, X_{(4)}\}$，有：

新的$G^{(3)}=\{X_{(5)}, C_{(4)}, C_{(3)}\}$
新的类别数目$m=3$
新的类间距离矩阵$D^{(3)}$

$D^{(3)}$

	$X_{(5)}$	$C_{(4)}$	$C_{(3)}$
$X_{(5)}$	0	6	2
$C_{(4)}$		0	2.5
$C_{(3)}$			0

$C_{(2)}$

Peking University
Introduction to Computational Biology

Step 4

由$D^{(3)}$知，合并$X_{(5)}$和$C_{(3)}$为一新类$C_{(2)}=\{X_{(5)}, C_{(3)}\}$，有：

新的$G^{(4)}=\{C_{(4)}, C_{(2)}\}$
新的类别数目$m=2$
新的类间距离矩阵$D^{(4)}$

$D^{(4)}$

	$C_{(4)}$	$C_{(2)}$
$C_{(4)}$	0	2.5
$C_{(2)}$		0

$C_{(1)}$

Step 5

由$D^{(4)}$知，最后合并$C_{(4)}$和$C_{(2)}$为一新类$C_{(1)}=\{C_{(4)}, C_{(2)}\}$，有：

新的$G^{(5)}=\{C_{(4)}, C_{(2)}\}$
新的类别数目$m=1$
新的类间距离矩阵$D^{(5)}$

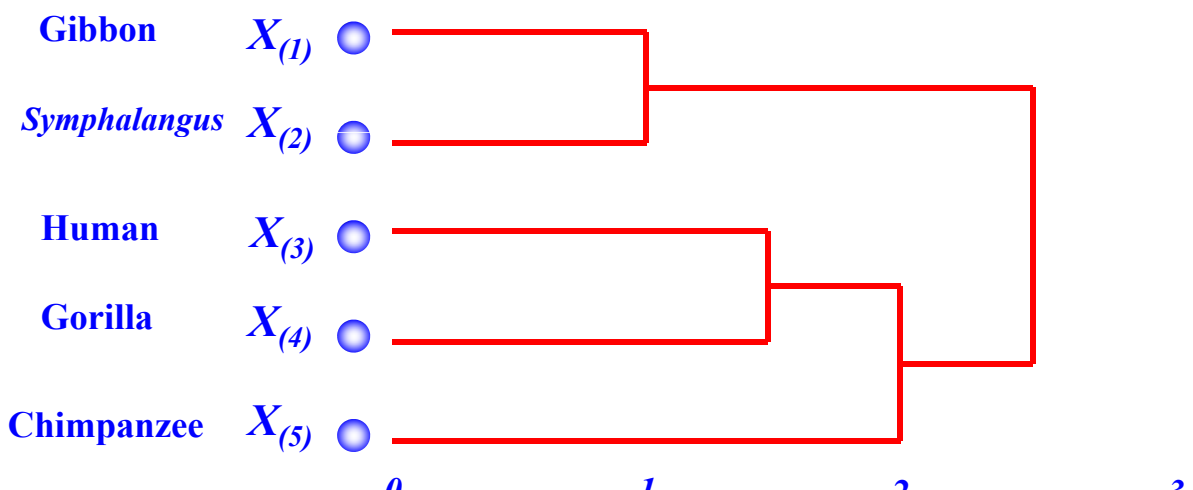
$D^{(5)}$

	$C_{(1)}$
$C_{(1)}$	0

Peking University
Introduction to Computational Biology

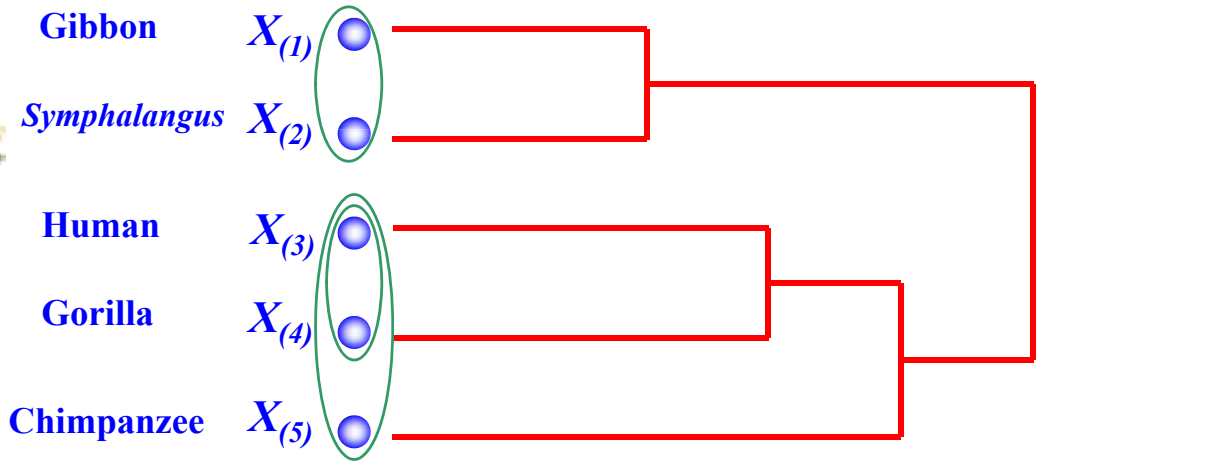
Step 6

画谱系聚类图



Step 7

确定类别的数目以及各类的成员。



Peking University
Introduction to Computational Biology

2、类间距离的定义

影响聚类结果的主要因素

- 样品间距离的定义d_{ij}
- 类间距离的定义D_{ij}

用G_p和G_q表示两个类，它们所包含的样品数目分别为t_p和t_q，类G_p和G_q之间的距离用D_{pq}表示。

1）、最短距离

定义： G_p和G_q中最邻近的两个样品的距离为这两个类之间的距离。

$$D_{pq} = \min\{ d_{ij} \mid i \in G_p, j \in G_q \}$$

讨论（递推公式）：设G_r是由G_p和G_q合并得到的新类，考虑G_r与G_s（$s \neq p, q$）的类间距离（最短距离） D_{rs}，有：

$$\begin{aligned} D_{rs} &= \min\{d_{ij} \mid i \in G_r, j \in G_s\} \\ &= \min\{\min\{d_{ij} \mid i \in G_p, j \in G_s\}, \min\{d_{ij} \mid i \in G_q, j \in G_s\}\} \\ &= \min\{D_{ps}, D_{qs}\} \end{aligned}$$

Peking University
Introduction to Computational Biology

2）、最长距离

定义： G_p和G_q中相距最远的两个样品的距离为这两个类之间的距离。

$$D_{pq} = \max\{ d_{ij} \mid i \in G_p, j \in G_q \}$$

讨论（递推公式）：设G_r是由G_p和G_q合并得到的新类，考虑G_r与G_s（$s \neq p, q$）的类间距离（最长距离） D_{rs}，有：

$$\begin{aligned} D_{rs} &= \max\{d_{ij} \mid i \in G_r, j \in G_s\} \\ &= \max\{\max\{d_{ij} \mid i \in G_p, j \in G_s\}, \max\{d_{ij} \mid i \in G_q, j \in G_s\}\} \\ &= \max\{D_{ps}, D_{qs}\} \end{aligned}$$

3）、类平均距离

定义：用G_p和G_q中每两两样品间平方距离的平均值作为两个类之间的距离。

$$D_{pq}^{\ 2} = \frac{1}{t_p t_q}\sum_{i=1}^{t_p}\sum_{j=1}^{t_q} d_{ij}^{2} \qquad i \in G_p,\ j \in G_q$$

讨论（递推公式）：设G_r是由G_p和G_q合并得到的新类，考虑G_r与G_s（$s \neq p, q$）的类间距离（类平均距离） D_{rs}，有：

$$D_{rs}^{\ 2} = \frac{1}{t_r t_s}\sum_{i=1}^{t_r}\sum_{j=1}^{t_s} d_{ij}^{2} \qquad i \in G_r,\ j \in G_s$$

$$= \frac{1}{t_r t_s}\sum_{i=1}^{t_s}\left(\sum_{j=1}^{t_p} d_{ij}^{2} + \sum_{j=1}^{t_q} d_{ij}^{2}\right) = \frac{1}{t_r t_s}\left[t_s t_p D_{sp}^{\ 2} + t_s t_q D_{sq}^{\ 2}\right]$$

$$= \frac{t_p}{t_r} D_{sp}^{\ 2} + \frac{t_q}{t_r} D_{sq}^{\ 2}$$

Peking University
Introduction to Computational Biology

4）、几何中心距离

定义：用G_p和G_q两类几何中心的距离为两个类之间的距离。

$$D_{pq} = d\left\{\overline{X}^{(p)}, \overline{X}^{(q)}\right\}$$

$$\overline{X}^{(p)} = \frac{1}{t_p}\sum_{i=1}^{t_p} X_i^{(p)} \qquad \overline{X}^{(q)} = \frac{1}{t_q}\sum_{i=1}^{t_q} X_i^{(q)}$$

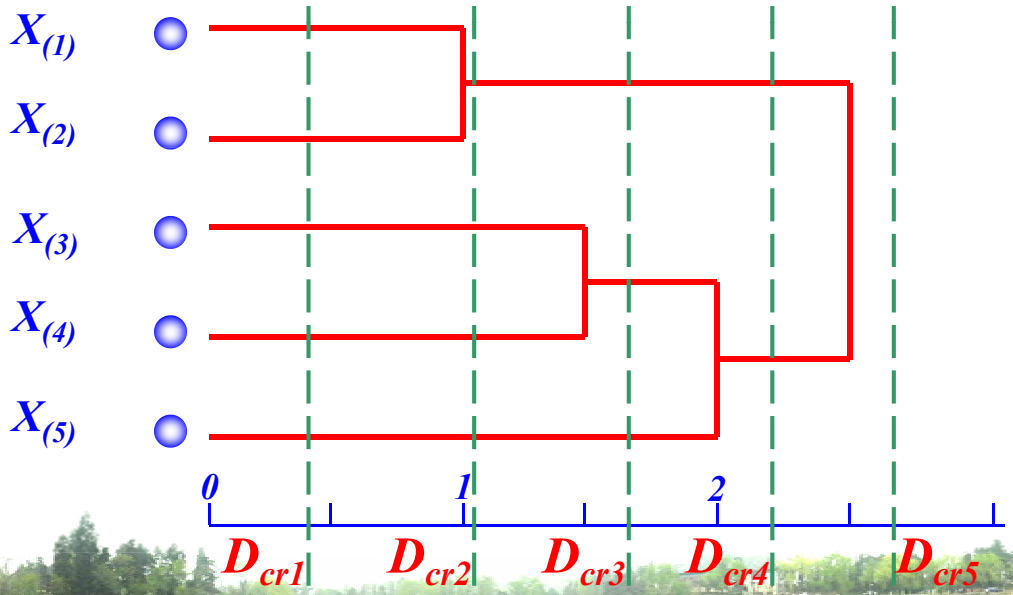
讨论（递推公式）：设G_r是由G_p和G_q合并得到的新类，考虑G_r与G_s（$s \neq p, q$）的类间距离（几何中心距离） D_{rs}，有：

$$D_{rs} = d\left\{\overline{X}^{(r)}, \overline{X}^{(s)}\right\}$$

3、类别数目的确定

问题：谱系聚类图仅仅反映样品间的亲疏、远近关系，本身并没有给出分类关系。

1）、由临界值确定



Peking University
Introduction to Computational Biology

2）、由数据散点图直观确定

二维散点图
三维散点图
高维散点图

3）、由统计量确定

（略）

4）、确定类别数目的基本原则

1、各类几何中心之间的距离应该尽可能地大；
2、确定的类中，各类所包含的元素不宜太多；
3、类别数目应该符合实用目的；
4、若采用几种不同的聚类方法，在不同的谱系聚类图中应该发现相同的类。

Peking University
Introduction to Computational Biology

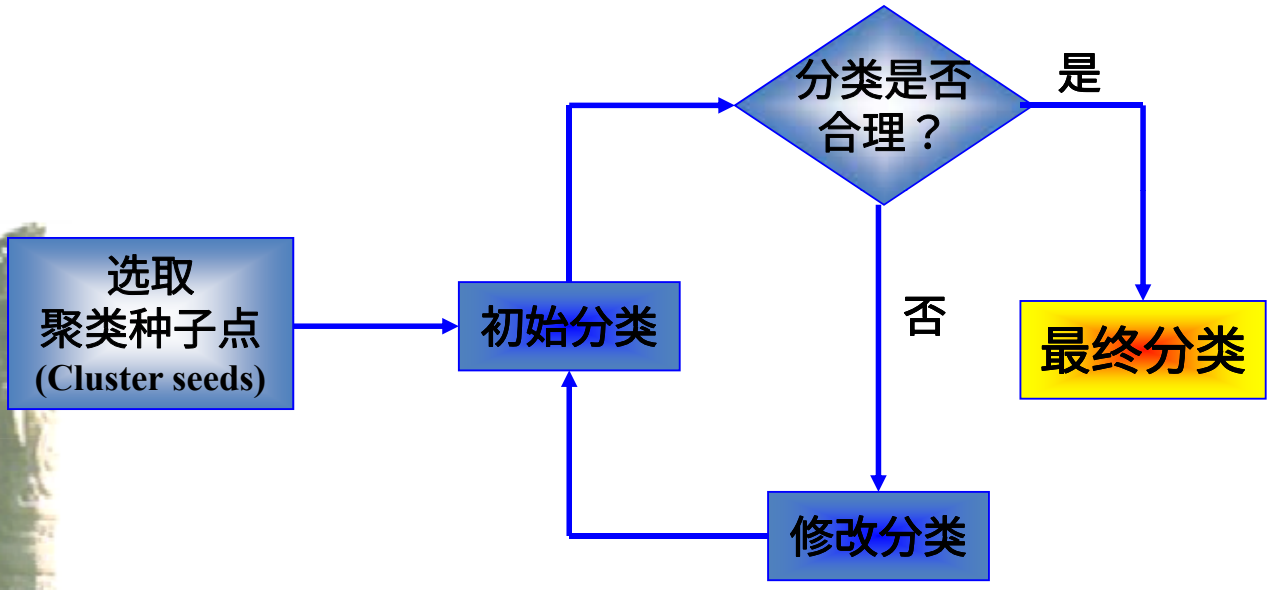
6.3.3 聚类分析方法之二：动态聚类法

谱系聚类法 → 静态：一次分类
计算量大
不适合处理大样本问题

动态聚类法 → 动态：逐步分类
计算量较小
适合处理大样本问题

基本思想

首先粗略分类，然后按照某种最优原则修改不合理的分类，直至分类合理。



动态聚类法的基本过程

Peking University
Introduction to Computational Biology

1、选取聚类种子点

聚类种子点（Cluster seeds）：准备形成类的中心，是一批有代表性的点。

聚类种子点选取的重要性：直接决定初始分类

初始分类的重要性：影响最终分类结果

（1）人为选择

条件：对所分类问题有一定的了解

根据经验，预先确定分类的数目、初始分类，并从每类中选择有代表性的一个点作为种子点。

（2）人为分类、选取几何中心

条件：对所分类问题有一定的了解

根据经验，预先将数据人为地分为k类，计算每一类的几何中心，选取这些中心作为聚类种子点。

Peking University
Introduction to Computational Biology

（3）密度法

以d (d>0)为半径，以某个样品X为球心，落在小球内的全部样品数就是样本X的密度。

- 计算所有样品点的密度，首先选取密度最大的样品点作为第一种子点；
- 在所有与第一种子点距离大于D (一般D=2d)的样品点中，选取密度最大的样品点作为第二种子点；
- 在所有与第一、第二种子点距离大于D的样品点中，选取密度最大的样品点作为第三种子点；
- 依次考察全部样品点，得到全部聚类种子点。

半径d的选择要合理

（4）选取总体几何中心

首先以所有样品的几何中心为第一种子点。
然后依次考察每个样品点，若某一点与已有种子点距离均大于d值，则选取该点为新的种子点。

（5）随机选取

随机选取聚类种子点。

假设分为k类，则用前k个样品作为聚类种子点。

Peking University
Introduction to Computational Biology

2、确定初始分类

（1）人为分类

条件：对所分类问题有一定的了解

根据经验，预先确定分类的数目、初始分类。

（2）最近距离归类

条件：已经选定聚类种子点

选定聚类种子点后，每个样品点按照与其距离最近的种子点分类。

（3）动态调整种子点

选定初始聚类种子点后，依次将每个样品点归入与其距离最近的种子点所在的类，并重新计算该类的几何中心，以中心代替该类的种子点。直到考察完所有的样品点。

（4）部分抽样分类

样本量太大时：抽取部分样本，按照前面几种方法得到初始分类。

Peking University
Introduction to Computational Biology

3、按批修改动态聚类法

基本步骤

(1). 选择聚类种子点，选定距离的定义；

(2). 将所有样品点按照最近距离原则归入种子点所在的类；

(3). 计算每一类的几何中心，将几何中心点作为新的聚类种子点；

(4). 对所有样品点按照最近距离原则重新归类；

(5). 转到(3)，若聚类种子点与前一次聚类种子点重合，或者满足给定的收敛条件，迭代过程结束。

分类函数与修改原则

假设全部t个样品点为$X_{(i)}$ (i=1, 2, …, t)，初始分为k类：$G^{(1)}$, $G^{(2)}$, …, $G^{(k)}$，每类样品点数为t_i (i=1, 2, …, k)。

用$n(i)$表示样品点$X_{(i)}$到所属类的标号，则分类函数定义为：

$$\ell\left(G^{(1)}, G^{(2)}, ..., G^{(k)}\right) = \sum_{i=1}^{t} D^2\left(X_{(i)}, G^{(n(i))}\right) \qquad n(i) \in (1, ..., k)$$

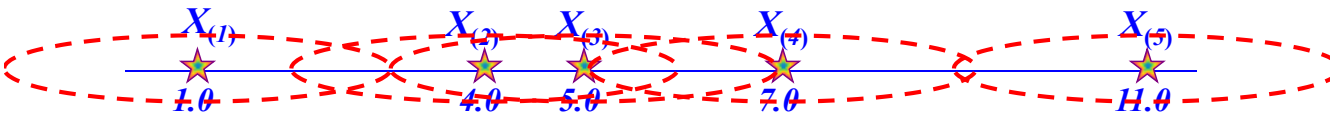
$$D^2\left(X_{(i)}, G^{(n(i))}\right) = \left(X_{(i)} - \overline{X}^{(j)}\right)' \left(X_{(i)} - \overline{X}^{(j)}\right)$$

分类函数实际上就是离差平方和。

修改原则：使分类函数的值达到最小。

Peking University
Introduction to Computational Biology

例



（1）、用密度法选取聚类种子点：

取d=2, D=2d=4；

采用欧氏距离

	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
密度	1	2	3	2	1

得到　第一种子点：$X_{(3)}$

第二种子点：$X_{(1)}$

第三种子点：$X_{(5)}$

（2）、初始分类：

按照最小距离原则将所有样品点归类。结果是

$$G^{(1)} = \{X_{(3)}, X_{(2)}, X_{(4)}\}$$
$$G^{(2)} = \{X_{(1)}\}$$
$$G^{(3)} = \{X_{(5)}\}$$

（3）、修改分类：

首先计算各类的几何中心：5.333，1.0，11.0

以它们作新的聚类种子点，按照最小距离原则重新归类：

$$G^{(1)} = \{X_{(3)}, X_{(2)}, X_{(4)}\}$$
$$G^{(2)} = \{X_{(1)}\}$$
$$G^{(3)} = \{X_{(5)}\}$$

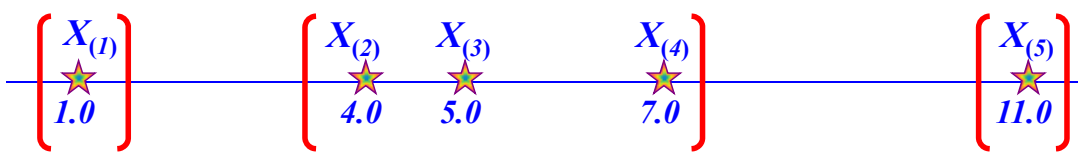
再次计算各类的几何中心：5.333，1.0，11.0

与前一次重合，迭代过程终止。

Peking University
Introduction to Computational Biology

（3）、最终分类：

$$G^{(1)} = \{X_{(3)}, X_{(2)}, X_{(4)}\}$$
$$G^{(2)} = \{X_{(1)}\}$$
$$G^{(3)} = \{X_{(5)}\}$$



4、*k*-means法
（逐点修改动态聚类法）

基本步骤

(1). 给定3个参数：

K ——分类的数目（初步估计）

D_{min} ——类间点距离的最小值

d_{max} ——类内点距离的最大值

(2). 选择聚类种子点

通常选择前K个样品点作为聚类种子点

或者选取有代表性的K个样品点作为聚类种子点

Peking University
Introduction to Computational Biology

(3). 调整聚类种子点

计算这K个种子点两两之间的距离：

将距离小于D_{min}的两个种子点合并，以它们的中心点作为新的种子点；

保证所有的种子点两两之间距离大于或等于D_{min}。

(4). 逐点调整聚类种子点

将剩下的$t-K$个样品点逐个归类：

若某样品点与所有种子点的距离均大于d_{max}，则将该样品点视为新的聚类种子点添加进来；

否则，归为与之距离最近的种子点所在类别，同时计算该类的几何中心，以中心点作为新的聚类种子点；

返回（3），调整聚类种子点，保证所有的种子点两两之间距离大于或等于D_{min}。

考虑下一个样品点。

(5). 对所有样品点重新归类

将所有样品点重新逐个归类：

若某样品的分类与原来不同，则要重新计算它所涉及的两类的几何中心，并调整它们的聚类种子点。然后调整所有的聚类种子点，保证所有的种子点两两之间距离大于或等于 D_{min}。

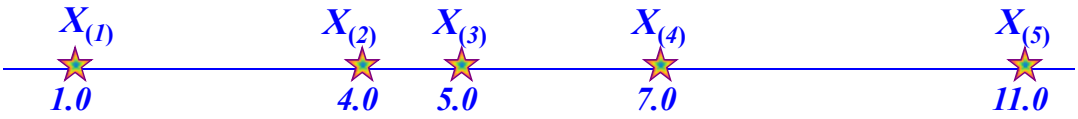
(6). 分类迭代终止

若所有样品的分类与上一次相同，分类迭代终止。

得到最终的分类。

Peking University
Introduction to Computational Biology

例



（1）、选择参数：

K ——3

D_{min} ——2

d_{max} ——3

采用欧氏距离

（2）、选择聚类种子点：

$X_{(1)}$、$X_{(2)}$、$X_{(3)}$。

（3）、调整聚类种子点：（$D_{min}=2$）

$d(X_{(1)}, X_{(2)})=3$

$d(X_{(1)}, X_{(3)})=4$

$d(X_{(2)}, X_{(3)})=1$

将$X_{(2)}, X_{(3)}$合并，用它们的几何中心4.5作为新的聚类种子点

得到2个符合要求的聚类种子点：1.0　　4.5

（4）、逐点考察其它样本点　（$d_{max}=3$）

$X_{(4)}, X_{(5)}$，得到3类：

$$G^{(1)} = \{X_{(3)}, X_{(2)}, X_{(4)}\}$$

$$G^{(2)} = \{X_{(1)}\}$$

$$G^{(3)} = \{X_{(5)}\}$$

Peking University
Introduction to Computational Biology

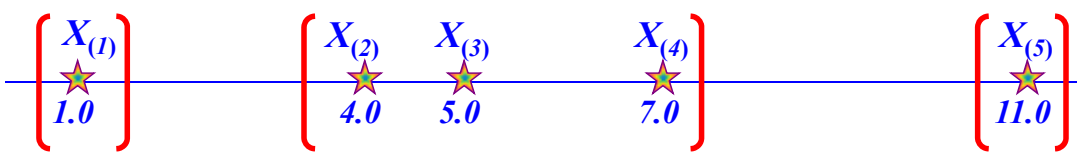
（5）、对所有样本点重新归类、调整：

收敛，迭代终止。

$$G^{(1)} = \{X_{(3)}, X_{(2)}, X_{(4)}\}$$

$$G^{(2)} = \{X_{(1)}\}$$

$$G^{(3)} = \{X_{(5)}\}$$



6.3.4 聚类分析方法之三：试探法

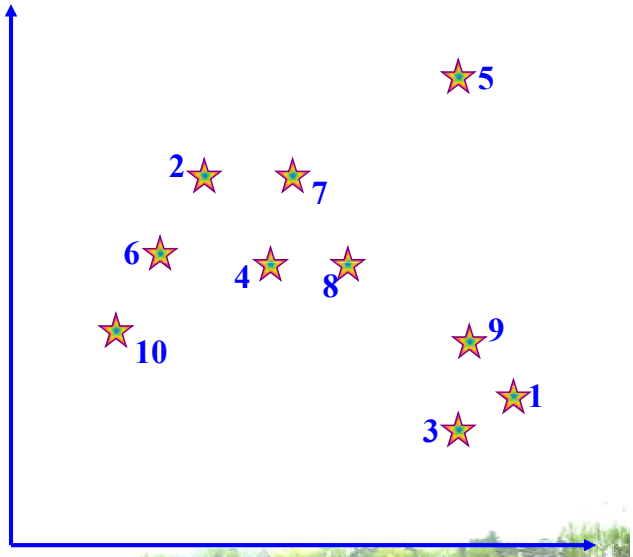
1、基于最邻近规则的试探聚类法

问题

已知全部t个样品点为$X_{(i)}$ (i=1, 2, …, t)，找出合理的聚类。

定义欧氏距离为样本间的距离。

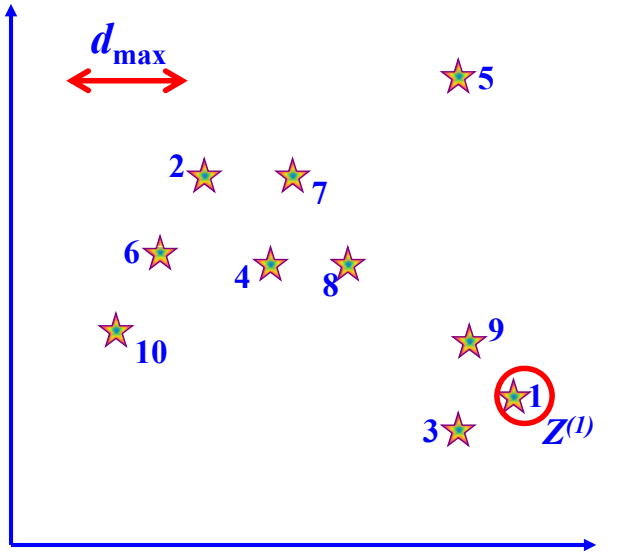
类间距离为最短距离。



Peking University
Introduction to Computational Biology

(1). 给定参数：

d_{max}——类内样品点距离的最大值（非负值）



(2). 选聚类中心$Z^{(1)}$

任选一样品点为聚类中心$Z^{(1)}$。

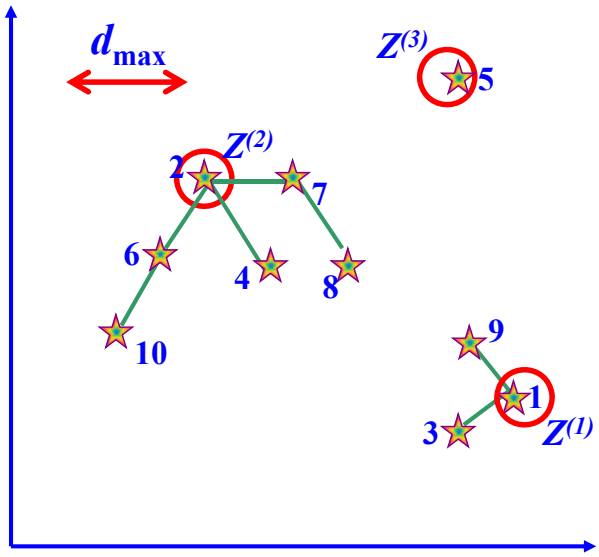
一般选$X_{(1)}$为$Z^{(1)}$。

(3). 逐点确定：

考察点$X_{(2)}$，当$d_{12} > d_{max}$时，选定$X_{(2)}$为新的聚类中心$Z^{(2)}$，否则，$X_{(2)}$属于$Z^{(1)}$类；

考察点$X_{(3)}$，当$d_{31} > d_{max}$且$d_{32} > d_{max}$时，选定$X_{(3)}$为新的聚类中心$Z^{(3)}$，否则，$X_{(3)}$属于距离最近的一类；

逐点考察所有的t个样品点$X_{(i)}$ (i=1, 2, …, t)，得到最后的聚类。



Peking University
Introduction to Computational Biology

特点：

（1）聚类速度快，计算量是样品点数的线性关系；
（2）简单方便。

聚类的效果受下列因素影响：

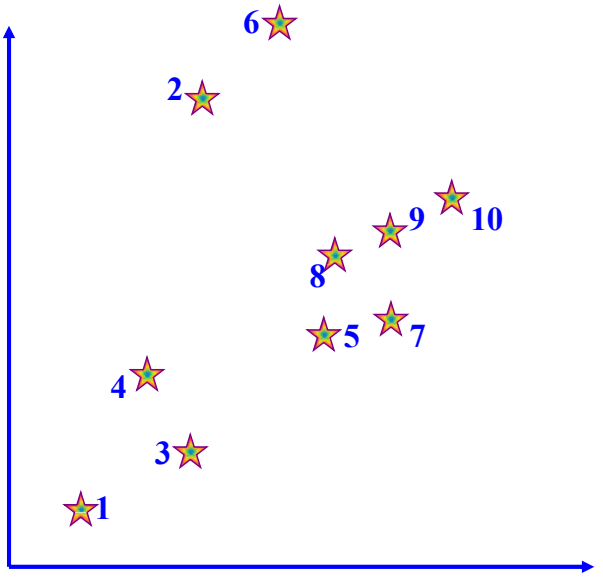
（1）第一个聚类中心的选取
（2）样品点的排序
（3）参数d_{max}的选取
（4）样品点的分布

2、最大最小距离算法

问题

已知全部t个样品点为$X_{(i)}$ (i=1, 2, …, t)，找出合理的聚类。

定义欧氏距离为样品间的距离。

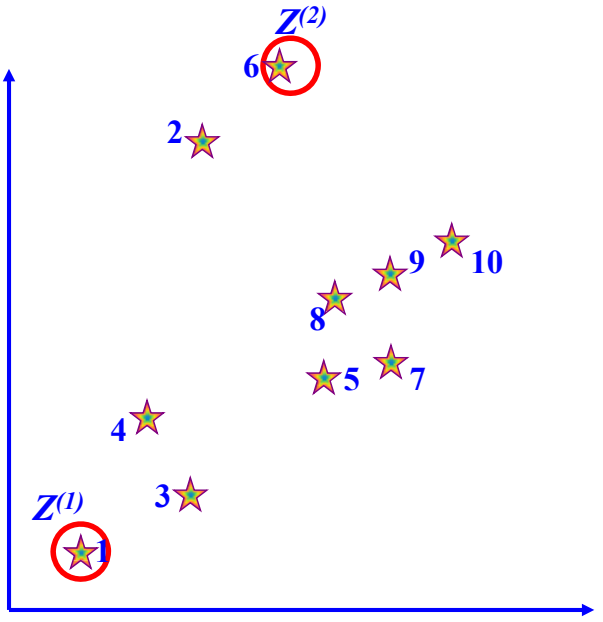


Peking University
Introduction to Computational Biology

(1). 确定最初的两个聚类中心：

计算两两之间距离，以最远的两个样品点作为两个聚类中心。

$d(X_{(1)}, X_{(6)})$最大，故选取$X_{(1)}$和$X_{(6)}$为两个聚类中心，记为$Z^{(1)}$和$Z^{(2)}$。



(2). 确定其它新的聚类中心：

逐点计算所有样品点$X_{(i)}$ (i=1, 2, …, t) 与$Z^{(1)}$和$Z^{(2)}$之间的距离：$d(X_{(i)}, Z^{(1)})$、 $d(X_{(i)}, Z^{(2)})$

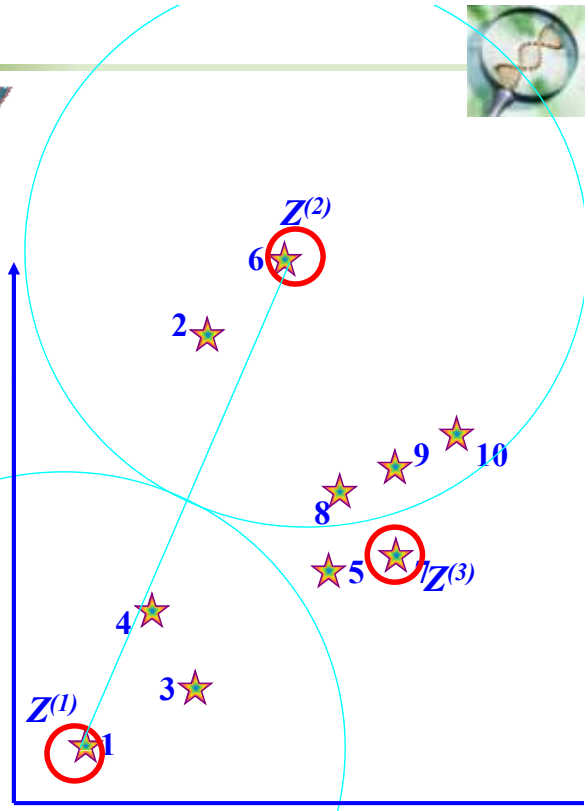
若有

$$\max\left\{\min\{d(X_{(i)}, Z^{(1)}), d(X_{(i)}, Z^{(2)})\}\right\} > \frac{1}{2}\left|d(Z^{(1)}, Z^{(2)})\right|$$

则令$X_{(i)}$为新添加的聚类中心点$Z^{(3)}$。

否则，转到最后一步。

选择$X_{(7)}$为$Z^{(3)}$。



Peking University
Introduction to Computational Biology

逐点计算所有样品点$X_{(i)}$ (i=1, 2, …, t) 与$Z^{(1)}$ 、$Z^{(2)}$和$Z^{(3)}$之间的距离：$d(X_{(i)}, Z^{(1)})$、$d(X_{(i)}, Z^{(2)})$、 $d(X_{(i)}, Z^{(3)})$

若有

$$\max\left\{\min\{d(X_{(i)}, Z^{(1)}), d(X_{(i)}, Z^{(2)}), d(X_{(i)}, Z^{(3)})\}\right\} > \frac{1}{2}\left|d(Z^{(1)}, Z^{(2)})\right|$$

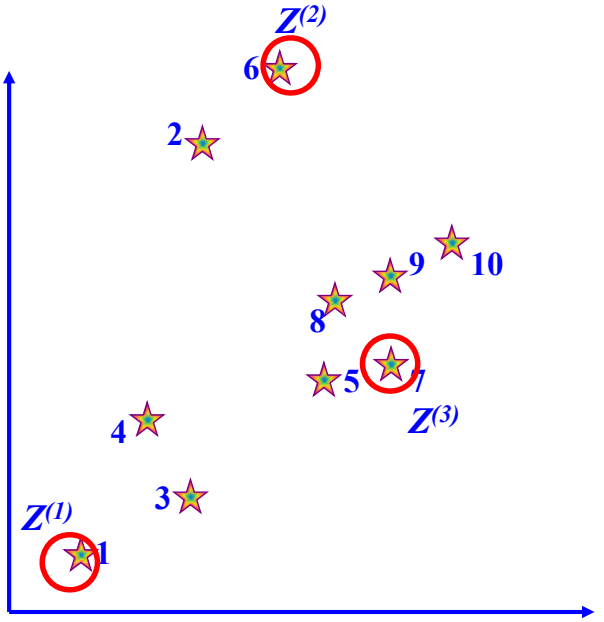
则令$X_{(i)}$为新添加的聚类中心点$Z^{(4)}$。

否则，转到最后一步。

……

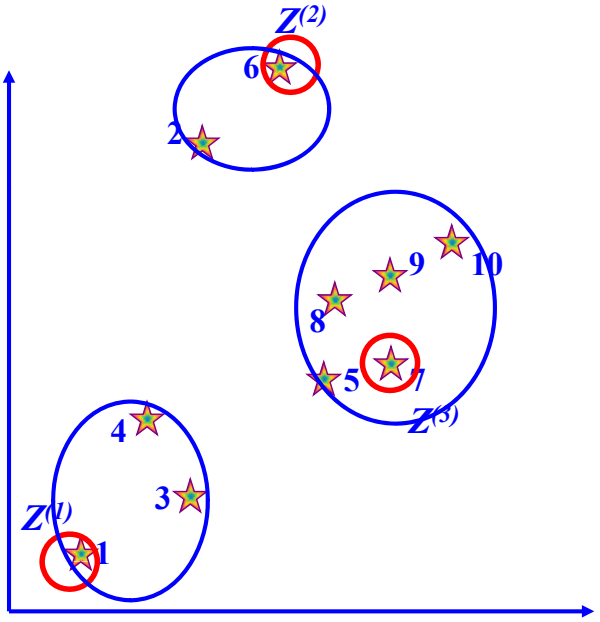
得到所有的k个聚类中心点：

$Z^{(1)}, Z^{(2)}, \ldots, Z^{(k)}$



(3). 按照聚类中心进行归类：
将所有样品点$X_{(i)}$ (i=1, 2, …, t) 按照与$Z^{(1)}$, $Z^{(2)}$,…,$Z^{(k)}$的最近距离进行归类。

$\{X_{(1)}, X_{(3)}, X_{(4)}\}$
$\{X_{(2)}, X_{(6)}\}$
$\{X_{(5)}, X_{(7)}, X_{(8)}, X_{(9)}, X_{(10)}\}$



Peking University
Introduction to Computational Biology

多元统计方法的经典参考书

1、J. M. Lattin等著，“多元数据分析（英文版）”，机械工业出版社，2003年

作者J. M. Lattin是Stanford大学商学院教授。该书介绍多元数据分析的现代方法，主要讲解多元统计学中的最新方法及其应用。书中大量习题和示例采用了来源于心理学、社会学、营销学等领域的真实数据。

2、高惠璇著，“应用多元统计分析”，北京大学出版社，2005年

北大数学教学系列教材，主要介绍一些实用的多元统计分析方法的理论及应用，结合SAS系统介绍应用实例。