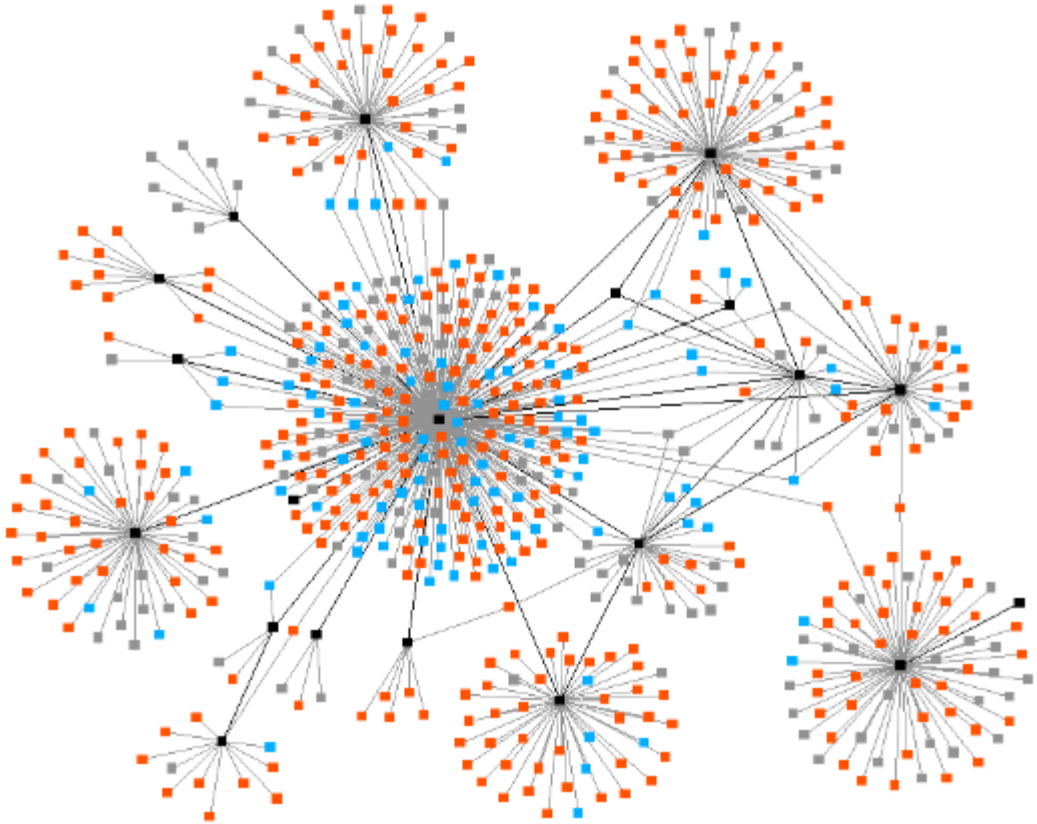
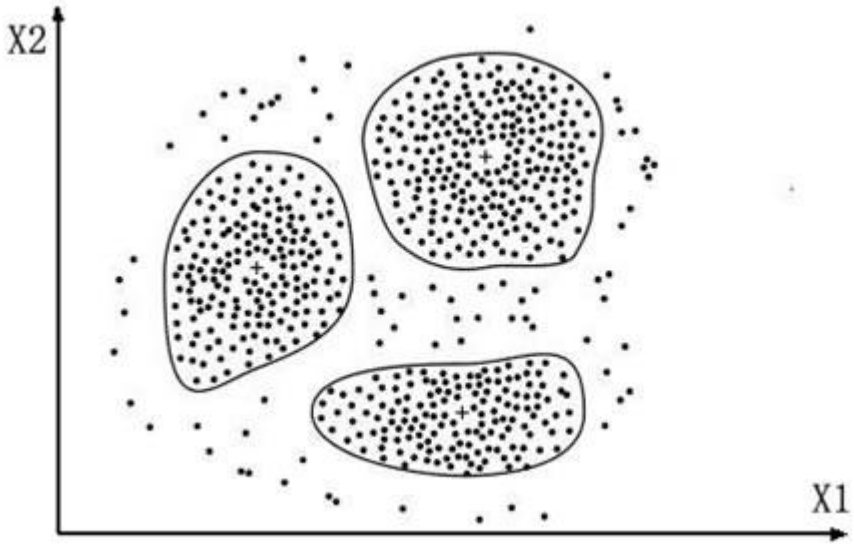


数据分析与R语言 第10周

2012.7.20

聚类

聚类和分类判别有什么区别？



关键度量指标：距离

- 距离的定义
- 常用距离（薛毅书P469）

绝对值距离

欧氏距离

闵可夫斯基距离

切比雪夫距离

马氏距离

Lance和Williams距离

离散变量的距离计算

dist( )函数

x1=c(1,2,3,4,5)

x2=c(3,2,1,4,6)

x3=c(5,3,5,6,2)

x=data.frame(x1,x2,x3)

```
> dist(x,method="euclidean")
         1        2        3        4
2 2.449490
3 2.828427 2.449490
4 3.316625 4.123106 3.316625
5 5.830952 5.099020 6.164414 4.582576

> dist(x,method="minkowski")
         1        2        3        4
2 2.449490
3 2.828427 2.449490
4 3.316625 4.123106 3.316625
5 5.830952 5.099020 6.164414 4.582576
> dist(x,method="minkowski",p=5)
         1        2        3        4
2 2.024397
3 2.297397 2.024397
4 3.004922 3.143603 3.004922
5 4.323101 4.174686 5.085057 4.025455
```

各种类与类之间距离计算的方法

- 薛毅书P476
- 最短距离法
- 最长距离法
- 中间距离法
- 类平均法
- 重心法
- 离差平方和法

动态聚类：K-means方法

- 算法：

1 选择K个点作为初始质心

2 将每个点指派到最近的质心，形成K个簇（聚类）

3 重新计算每个簇的质心

4 重复2-3直至质心不发生变化

kmeans( )函数

```
> X=iris[,1:4]
> km=kmeans(X,3)
>
>
> km
K-means clustering with 3 clusters of sizes 62, 50, 38

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.901613    2.748387     4.393548    1.433871
2     5.006000    3.428000     1.462000    0.246000
3     6.850000    3.073684     5.742105    2.071053

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [37] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [73] 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 3 3 3 3 1 3
[109] 3 3 3 3 3 1 1 3 3 3 3 1 3 1 3 1 3 3 1 1 3 3 3 3 3 1 3 3 3 1 3 3 3 1 3
[145] 3 3 1 3 3 1
```

K-means算法的优缺点

- 有效率，而且不容易受初始值选择的影响
- 不能处理非球形的簇
- 不能处理不同尺寸，不同密度的簇
- 离群值可能有较大干扰（因此要先剔除）

基于有代表性的点的技术：K中心聚类法

- 算法步骤

1 随机选择k个点作为“中心点”

2 计算剩余的点到这k个中心点的距离，每个点被分配到最近的中心点组成聚簇

3 随机选择一个非中心点O_r，用它代替某个现有的中心点O_j，计算这个代换的**总代价S**

4 如果S<0，则用O_r代替O_j，形成新的k个中心点集合

5 重复2，直至中心点集合不发生变化

K中心法的实现：PAM

- PAM使用离差平方和来计算成本S（类似于ward距离的计算）
- R语言的cluster包实现了PAM
- K中心法的优点：对于“噪音较大和存在离群值的情况，K中心法更加健壮，不像Kmeans那样容易受到极端数据影响
- K中心法的缺点：执行代价更高

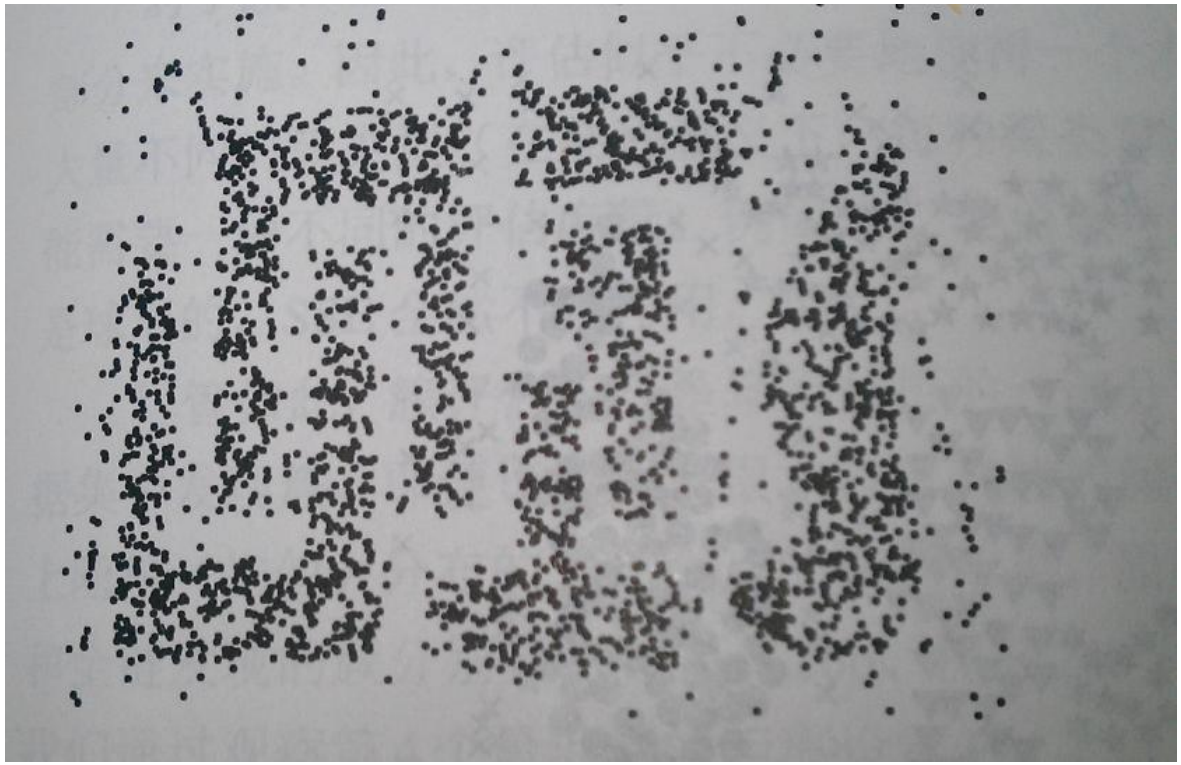
cluster包的pam( )函数

```
> x=iris[,1:4]
> kc=pam(x,3)
> kc
Medoids:
      ID Sepal.Length Sepal.Width Petal.Length Petal.Width
[1,]   8          5.0         3.4          1.5         0.2
[2,]  79          6.0         2.9          4.5         1.5
[3,] 113          6.8         3.0          5.5         2.1
Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [37] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [73] 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3
[109] 3 3 3 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3 3 3 2 3
[145] 3 3 2 3 3 2
Objective function:
    build      swap
0.6709391 0.6542077

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
 [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

基于密度的方法: DBSCAN

- DBSCAN = Density-Based Spatial Clustering of Applications with Noise
- 本算法将具有**足够高密度**的区域划分为簇，并可以发现**任何形状**的聚类



若干概念

r-邻域：给定点半径**r**内的区域

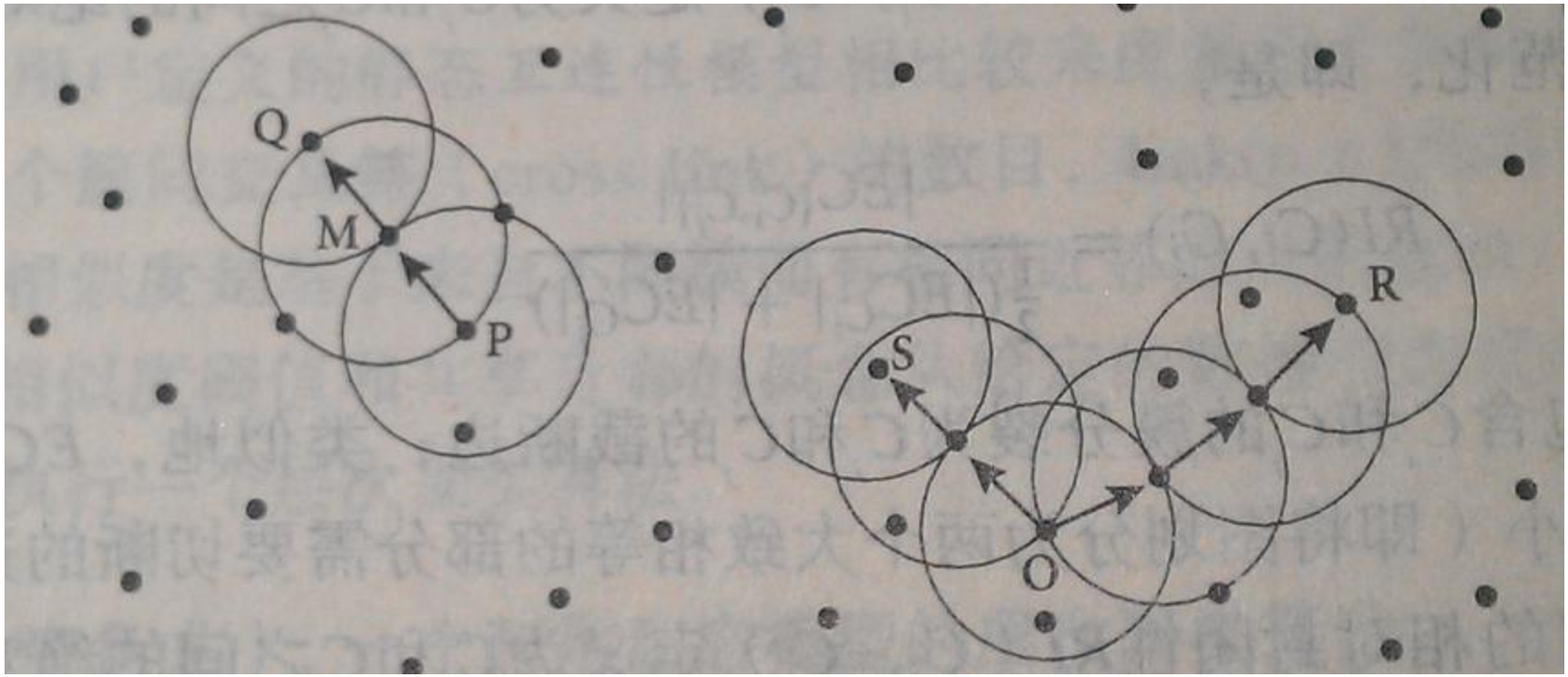
核心点：如果一个点的r-邻域至少包含最少数目**M**个点，则称该点为核心点

直接密度可达：如果点p在核心点q的r-邻域内，则称p是从q出发可以直接密度可达

如果存在点链$p_1,p_2, \ldots, p_n$，$p_1=q$，$p_n=p$，p_{i+1}是从p_i关于r和M直接密度可达，则称点p是从q关于r和M**密度可达**的

如果样本集D中存在点o，使得点p、q是从o关于 r和M密度可达的，那么点p、q是关于r和M**密度相连**的

DBSCAN



DBSCAN

- 算法基本 思想

1 指定合适的 r 和 M

2 计算所有的样本点，如果点p的r邻域里有超过M个点，则创建一个以p为核心点的新簇

3 反复寻找这些核心点直接密度可达（之后可能是密度可达）的点，将其加入到相应的簇，对于核心点发生“密度相连”状况的簇，给予合并

4 当没有新的点可以被添加到任何簇时，算法结束

DBSCAN算法描述

输入: 包含n个对象的数据库，半径e，最少数目MinPts;

输出:所有生成的簇，达到密度要求。

(1)Repeat

(2)从数据库中抽出一个未处理的点；

(3)IF抽出的点是核心点 THEN 找出所有从该点密度可达的对象，形成一个簇；

(4)ELSE 抽出的点是边缘点(非核心对象)，跳出本次循环，寻找下一个点；

(5)UNTIL 所有的点都被处理。

DBSCAN对用户定义的参数很敏感，细微的不同都可能导致差别很大的结果，而参数的选择无规律可循，只能靠经验确定。

孤立点检测

- 又称为异常检测，离群值检测等
- 什么是孤立点？**孤立点是一个观测值，它与其它观测值的差别如此之大，以至于怀疑它是由不同的机制产生的**
- 孤立点的一些场景

1 网站日志中的孤立点，试图入侵者

2 一群学生中的孤立点，天才 or 白痴？

3 天气数据，灾害，极端天气

4 信用卡行为，试图欺诈者

5 低概率事件，接种疫苗后却发病的

6 实验误差或仪器和操作问题造成的错误数据

等等

统计方法

- 检测一元正态分布中的离群点，指出离均值标准差数

表 10-1　均值为 0，标准差为 1 的高斯分布的样本对(c, α)，$\alpha = prob(|x| \geqslant c)$

c	$N(0,1)$的α
1.00	0.3173
1.50	0.1336
2.00	0.0455
2.50	0.0124
3.00	0.0027
3.50	0.0005
4.00	0.0001

多元正态分布的离群值

- 判断点到分布中心的距离（马氏距离，why？）

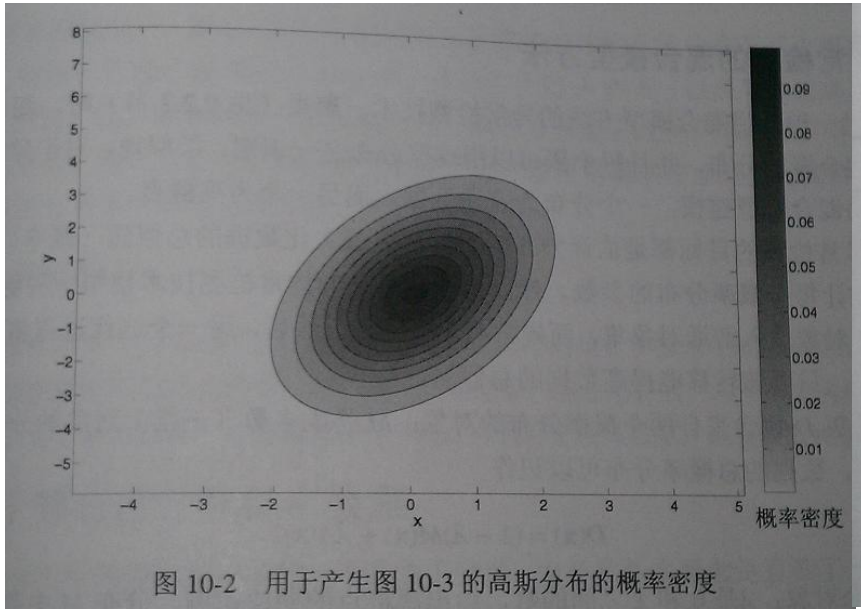
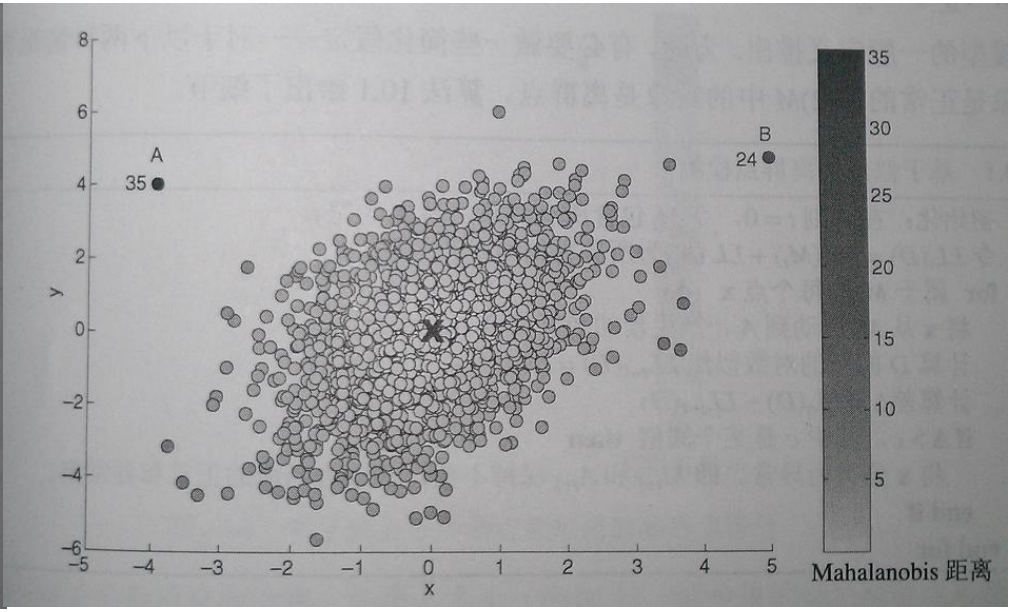


图 10-2　用于产生图 10-3 的高斯分布的概率密度



基于邻近度的孤立点检测

- 选取合适的正整数k
- 计算每个点和前k个最近邻的平均距离，得到孤立度指标
- 如果孤立度超过预定阈值，则找到孤立点

基于聚类的孤立点检测

- 首先聚类所有的点
- 对某个待测点评估它属于某一簇的程度。方法是设定一目标函数（例如kmeans法时的簇的误差平方和），如果删去此点能显著地改善此项目标函数，则可以将该点定位为孤立点

Thanks

FAQ时间