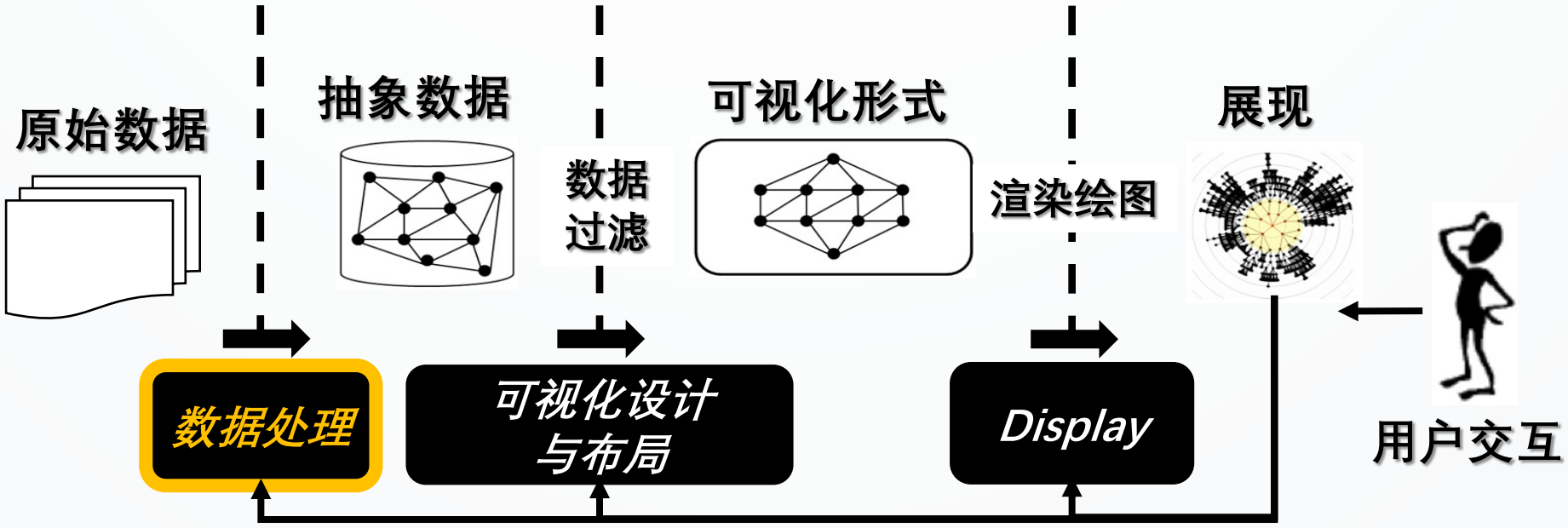
第二讲：数据及数据分析基础

信息可视化

曹楠（教授）
https://idvxlab.com
同济大学

智能大数据可视化实验室

怎样对数据进行可视化？

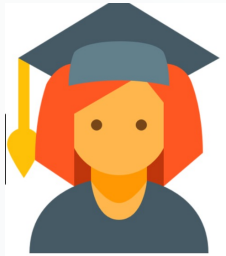


信息可视化参考模型

课程大纲

- 数据元素及元素的属性
- 数据元素之间的差异与聚类分析
- 数据属性之间的关联

数据的维度



Student

Machine Learning

Data Mining Visualization

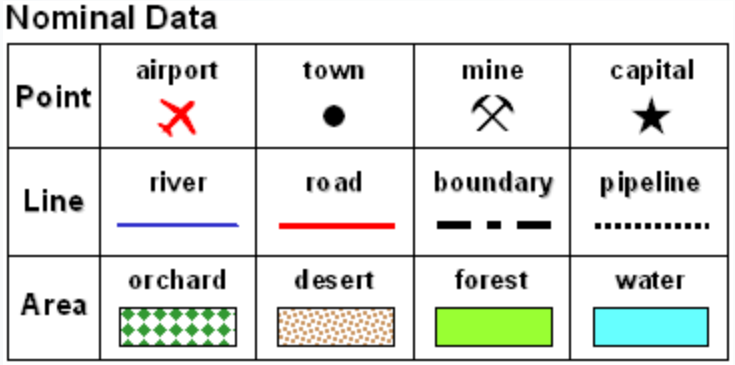
[22, Male, 3000, 20, 30, 5]

Age Sex scholarship Skills

- 数据的维度，是数据中用于描述数据元素的各种属性。例如，在学生数据集中，一个学生可以通过她的年龄、性别、助学金的额度、以及学生在相关课程上的技能，通过花费在课程上的时间来衡量，等
- 真实世界中的数据一般都是多维度的

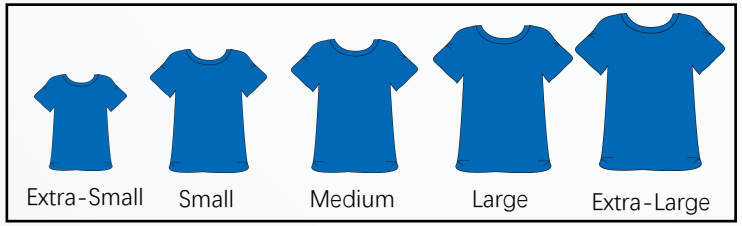
数据维度的类型

Nominal Data



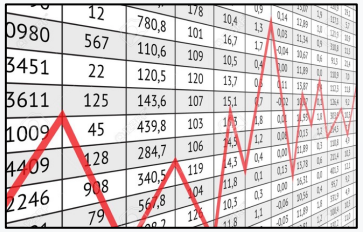
- 分类属性（Nominal ，Categorical)：该类型属性的取值代表了数据的类别，且相互间是无法排序的

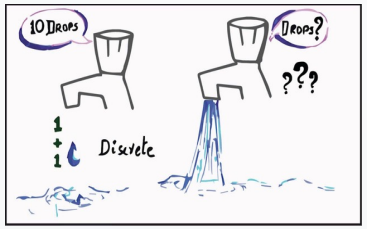
Ordinal



- 有序属性（Ordinal）：该类型属性的取值是有顺序的

Numerical





- 数值属性（Ordinal）：该类型属性的取值是数字。根据其取值的特点可以进一步分为 离散数值属性及连续的数值属性

一维数值属性的统计特征

- 均值（Mean）
- 中位数（Median）
- 方差（Variance）
- 标准差（Standard Deviation）

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

$$Q_{\frac{1}{2}}(x) = \begin{cases} x'_{\frac{n+1}{2}}, & \text{if } n \text{ is odd.} \\ \frac{1}{2}\left(x'_{\frac{n}{2}} + x'_{\frac{n}{2}+1}\right), & \text{if } n \text{ is even.} \end{cases}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2,$$

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

课程大纲

- 数据元素及元素的属性
- 数据元素之间的差异与聚类分析
- 数据属性之间的关联

数据元素之间的差异性

- 数据元素的差异性是通过数据元素在各个维度上的取值的差异性来加以度量的

- 整个数据集所有元素两两之间的差异性可以用差异矩阵来表示：
  - 差异矩阵是一个对称矩阵
  - 差异矩阵的每一行及每一列代表数据集中的一个数据元素
  - 差异矩阵中的每一个取值，代表对应元素之间的数据距离

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Dissimilarity Matrix

- 不同类型的数据维度之间有不同的距离计算方法

数据元素之间的差异性

- **分类属性** 之间的距离 通过“不匹配率”进行衡量
  - P 在某一分类属性上所有可能的取值的总数
  - m 代表数据元素在该属性上的取值相同情况的数目

$$d(i, j) = \frac{p - m}{p}$$

- 当分类属性只有两种可能的取值时（即为 binary），上述距离可以通过 “Jaccard coefficient”计算：

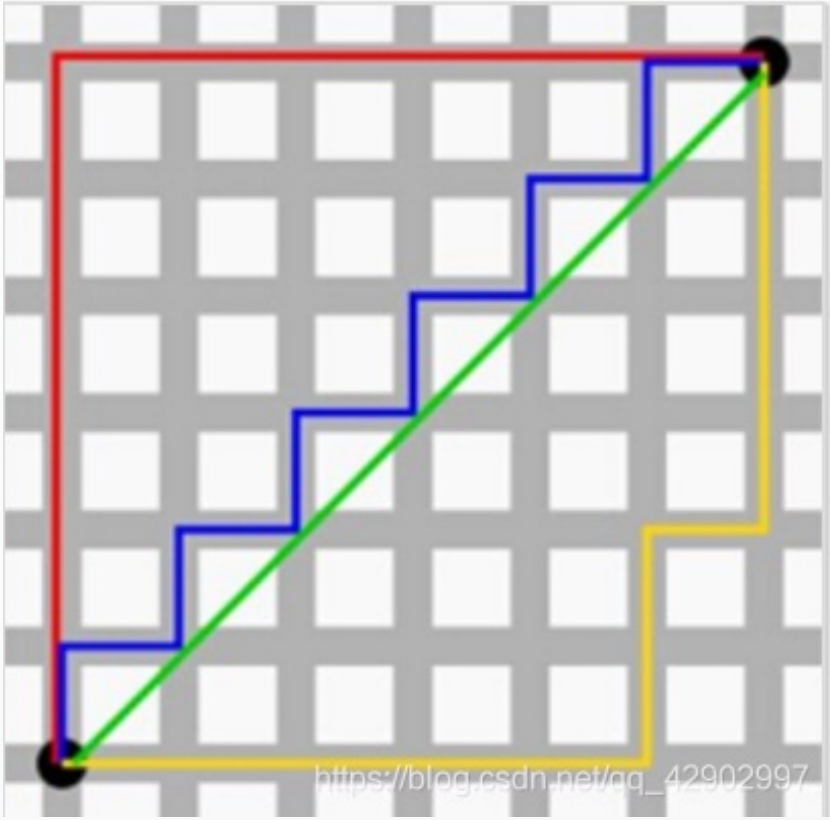
$$d(i, j) = \frac{r + s}{q + r + s}$$

q: 两个二进制字符串中相互匹配的总位数
r: 第一个字符串中取值为1，但是在第二个字符串中取值为0的位数
s: 第一个字符串中取值为0，但是在第二个字符串中取值为0的位数

String1: 0100010
String2: 1000010
q = 4, r = 1, s = 1
d(1, 2) = (1 + 1) / (4 + 1 + 1) = 1/3 = 0.333

数据元素之间的差异性

- **数值属性** 之间的距离 通过 **明科夫斯基距离** 进行计算
  - P = （xl ，x2, …, xn ）, Q = (yl, y2, …, yn)
  - 明科夫斯基距离（Lq ）- $\left[\sum_{i=1}^{n} |x_i - y_i|^q\right]^{\frac{1}{q}}$
  -
  - 曼哈顿距离（Ll ） - q = l
  - 欧几里得距离（L2 ） - q = 2



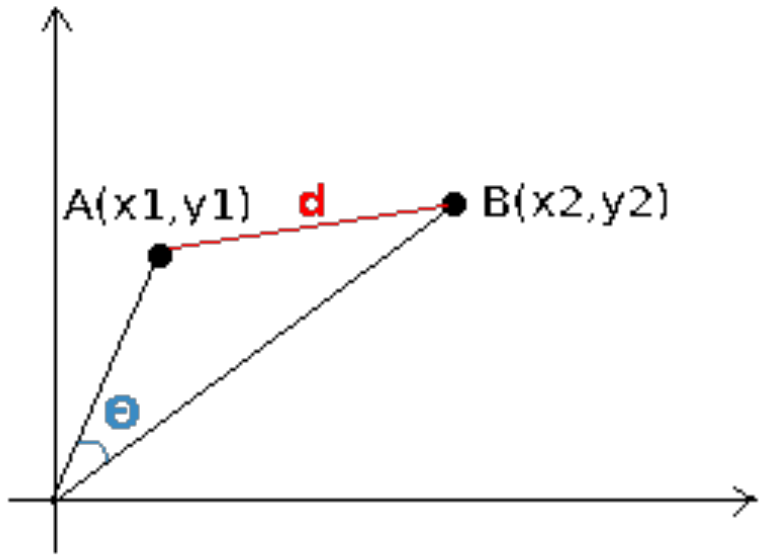
数据元素之间的差异性

- 数据维度取值量的大小会之间影响 欧式距离 的大小

- 有的时候并不希望计入这样的差异，需要用到 “角距离” （Cosine Distance）：

  例如，做文本分析时，为了计算文档与文档之间的差异性，文档的特征是通过文档中特定关键词出现的频次加以度量的。在这里，每一文档是一个数据元素，每一个关键词是数据元素的维度，而关键词在文档中出现的次数代表数据元素在该维度上的取值

- 如图所示，角距离度量的是两个向量之间夹角的大小（用 θ 表示），而不是空间中 点A 与 点B 的距离远近（用 d 表示），后者会因为向量的模的大小而改变。



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Example:

	New,	Vis,	Semester,	Course,	Data,	Score	
A =	[10,	3,	5,	20,	12,	11]	|A| = 28.26
B =	[5,	3,	6,	5,	3,	3]	|B| = 10.64

聚类分析

- 一组相互 关联/相似 的数据构成的集合称为簇
- 通过计算发现簇的过程称为聚类分析
- 聚类分析被应用于很多方面：在商业上，被用来进行用户画像，即用于发现不同的客户群，并且通过购买模式刻画不同的客户群的特征；在生物上，聚类分析被用来进基因分类及物种划分，获取对种群固有结构的认识；在互联网上，聚类分析被用于文档归类，主题提取与分析等

聚类分析

- 聚类分析是一种无监督学习算法，即对数据分类不依赖于任何先验知识及数据标签，仅仅取决于数据的内在属性

- 好的聚类分析算法应该能够分析得到高质量的聚类结果
  - 属于同一个类中的数据元素之间应该相互相似
  - 属于不同类中的数据元素之间应该有显著的差异

- 聚类分析的好坏与算法本身息息相关 也与算法所采用的数据度量方法相关

常见的聚类分析策略

- 基于数据划分的聚类分析
- 层次化的聚类分析
- 基于密度的聚类分析

基于数据划分的聚类分析

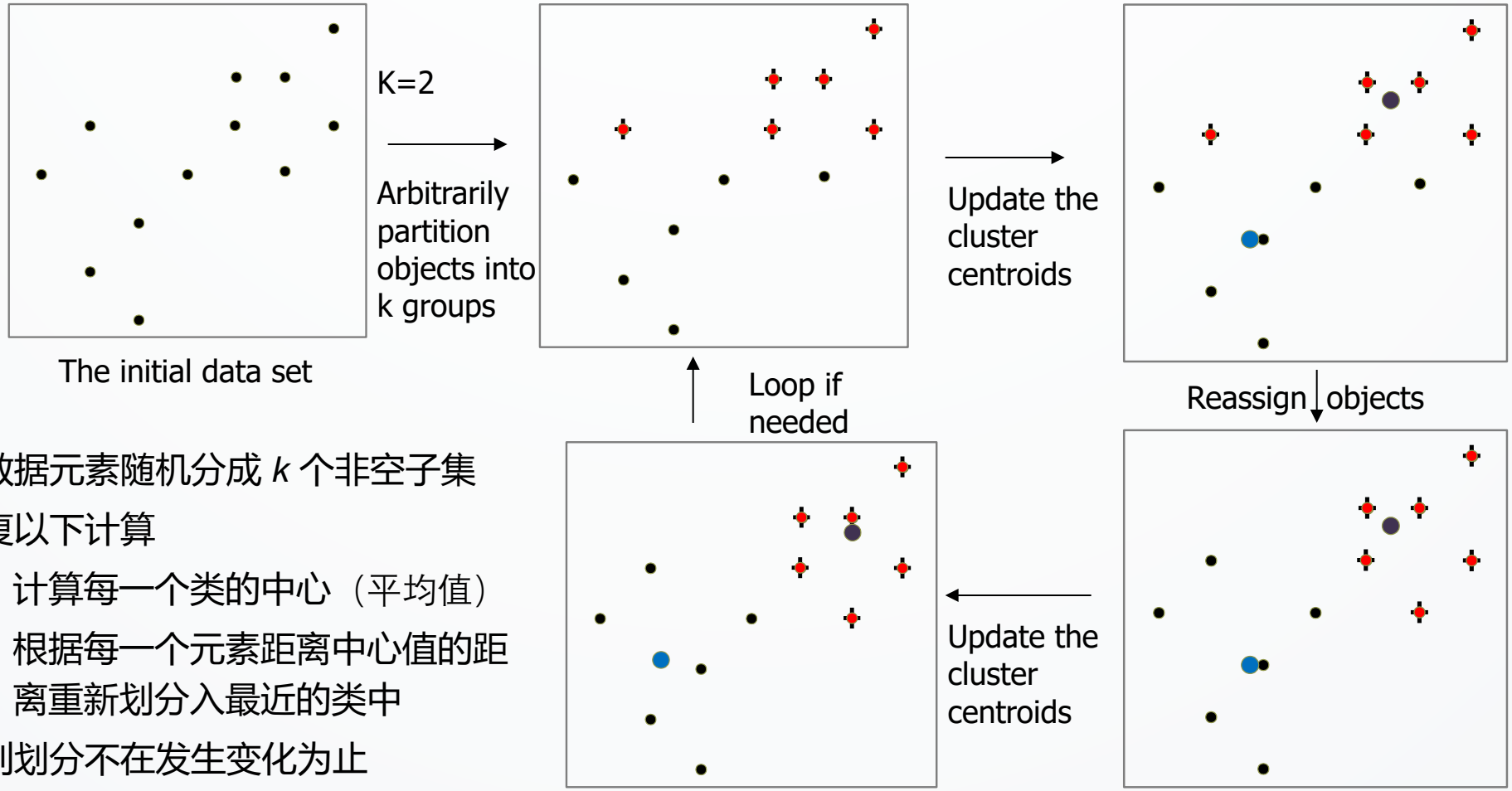
- 基于数据划分的聚类分析
- 讲一个数据集 D 中的数据元素根据其相似度划分为 K 份，并使得一下目标函数最小

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

类中的数据元素　类的中心点

- 求解最优结果需要穷举所有一切可能的划分方法
- 近似解法同样能够很好的解决问题

K-Means 聚类算法



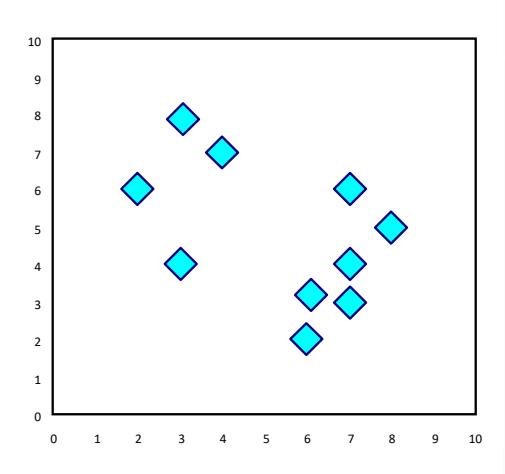
- 将数据元素随机分成 *k* 个非空子集
- 重复以下计算
  - 计算每一个类的中心（平均值）
  - 根据每一个元素距离中心值的距离重新划分入最近的类中
- 直到划分不在发生变化为止

K-Means 的优缺点

- 非常高效O（tkn），t 是迭代次数、k是簇的数量，n是数据元素的数量，正常情况下 $k, t << n$.

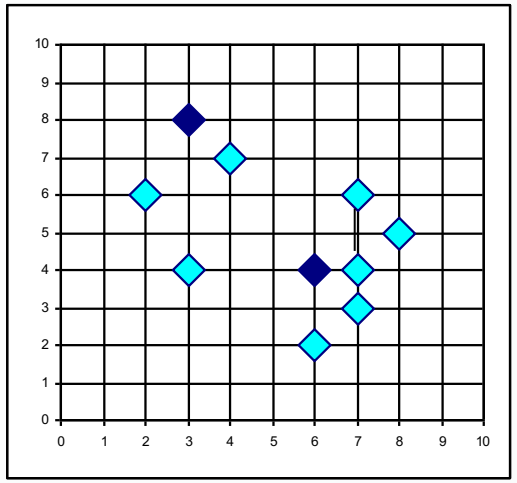
- 近似算法，最终结果往往不是最优解
- 需要输入 K
- 因为需要计算 means, 所以需要数据用于聚类的属性必须是数值型属性
- 对离群点（数据噪声、异常）比较敏感
- 只适用于凸包的情况

K-Centroid 聚类算法



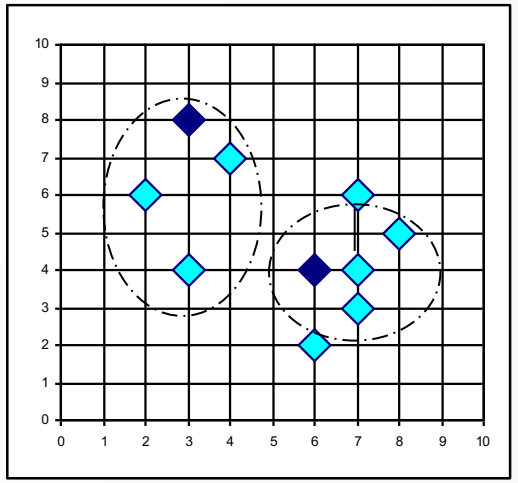
K=2

Arbitrary choose k object as initial medoids

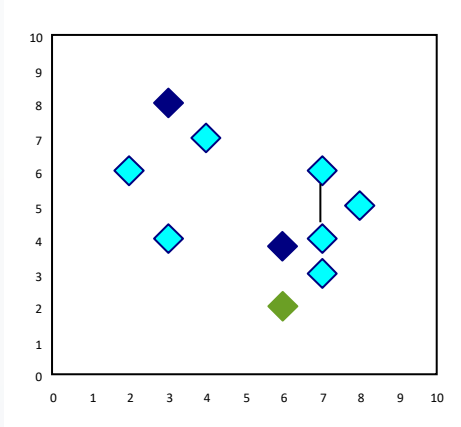


Assign each remaining object to nearest medoids

Total Cost = 20

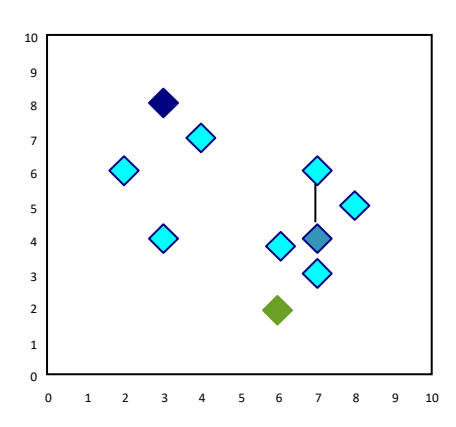


Randomly select a nonmedoid object,O_{ramdom}



Compute total cost of swapping

Total Cost = 26

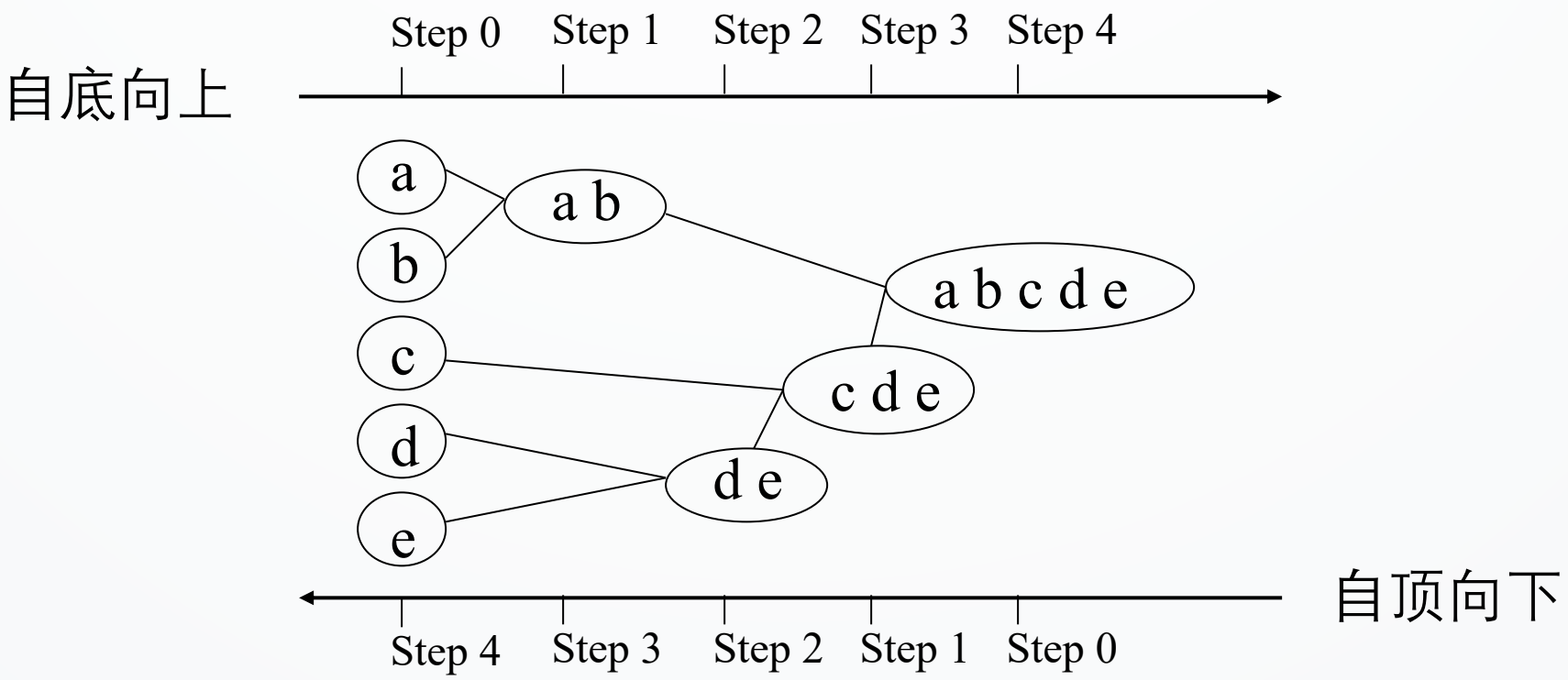


Swapping O and O_{ramdom}

If quality is improved.

- 重复以下计算
  - 查找每一个类的中心点
  - 根据每一个元素距离中心值的距离重新划分入最近的类中
- 直到划分不在发生变化为止

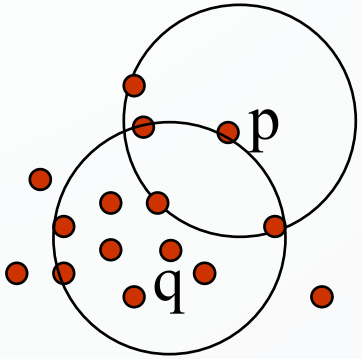
层次化的聚类分析


自底向上
Step 0
Step 1
Step 2
Step 3
Step 4
a
b
c
d
e
a b
a b c d e
c d e
d e
自顶向下
Step 4
Step 3
Step 2
Step 1
Step 0

基于数据密度的聚类分析方法

- 两个重要参数：
  - 最大检测半径 Eps
  - 最小相邻数据元素数 MinPts

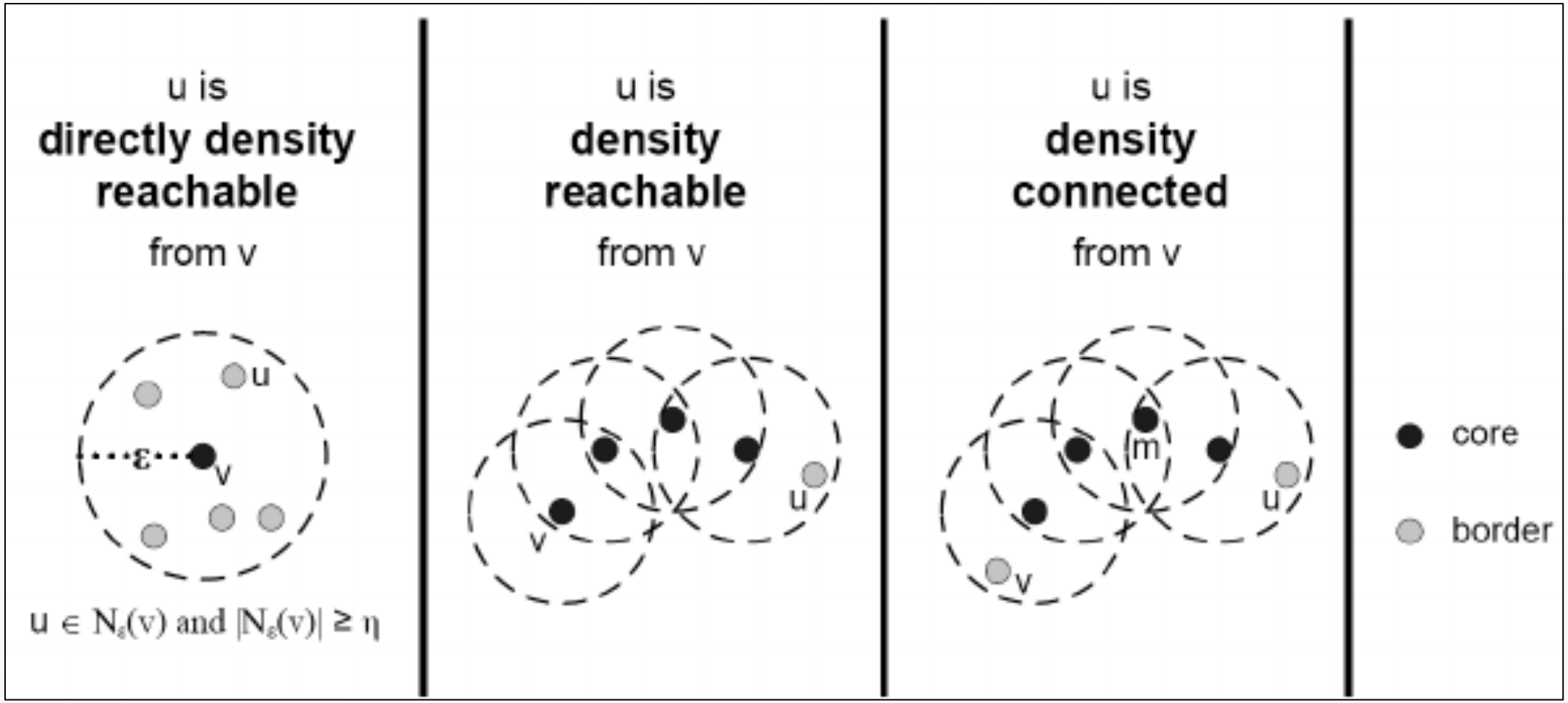
- 点 p 的邻居 N(p) 可以定义为以 p 为中心，以 Eps 为半径的圆所涵盖的所有数据元素



MinPts = 5

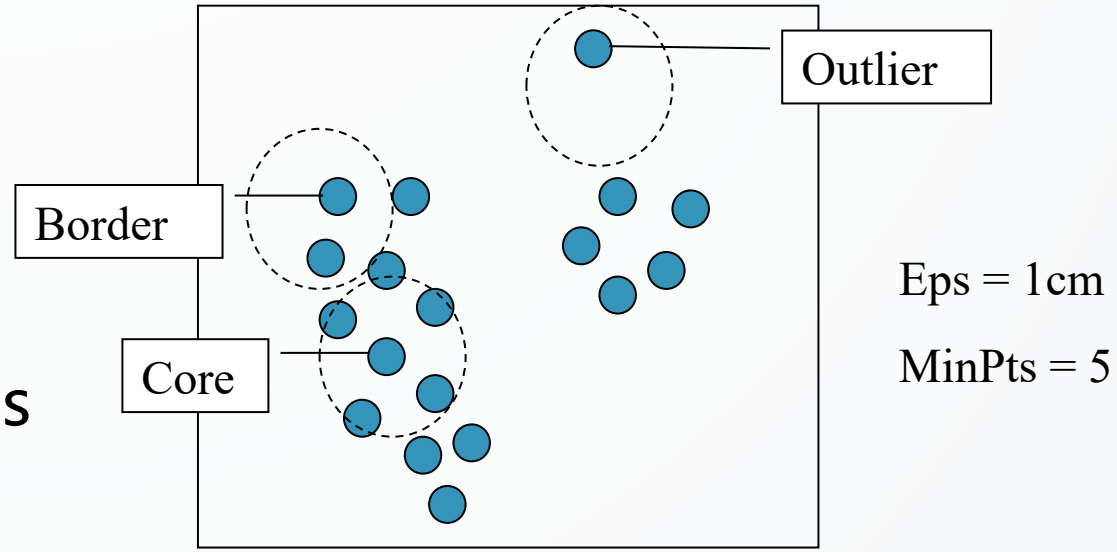
Eps = 1 cm

基于数据密度的聚类分析方法



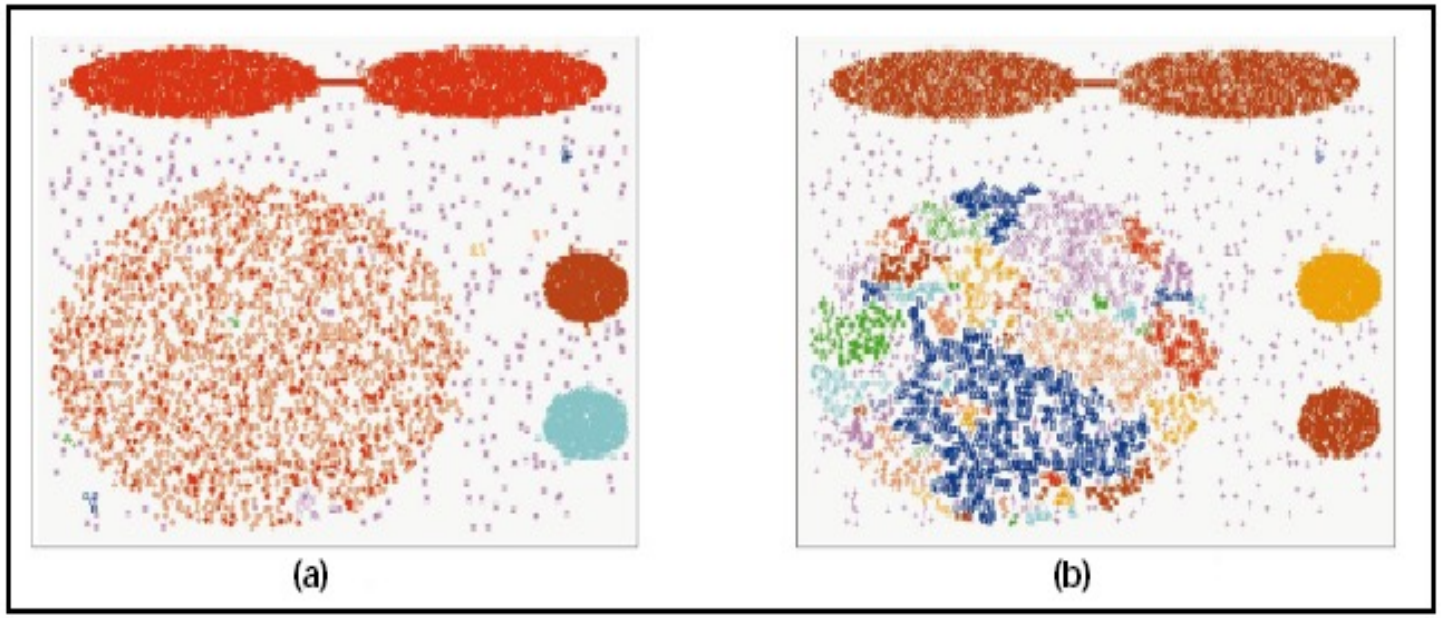
基于数据密度的聚类分析方法

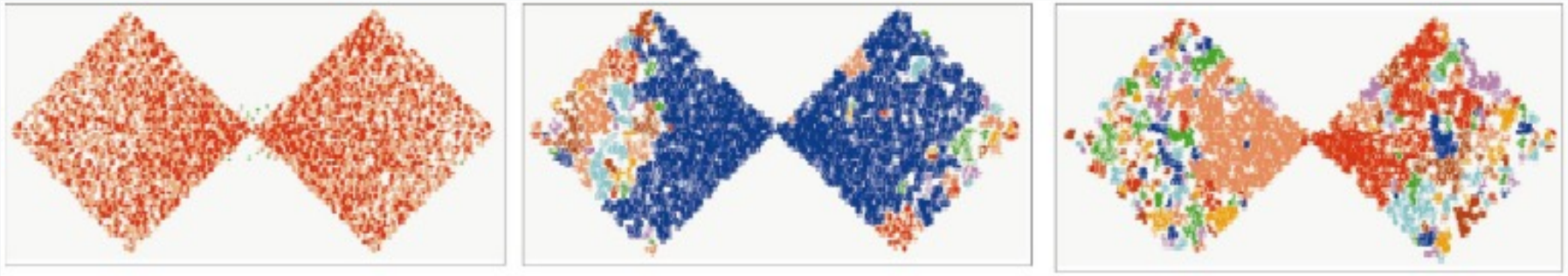
- 根据数据分布进行扫描
- 可以检测到任何形状的簇
- 不会收到数据噪声的影响
- 不用输入类的数目但是算法结果受到Eps以及 MinPts 的影响



基于数据密度的聚类分析方法

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.





课程大纲

- 数据元素及元素的属性
- 数据元素之间的差异与聚类分析
- 数据属性之间的关联

数据属性之间的关联

- 相关性分析（Correlation Analysis）：X 与 Y 之间是否有关联？ 用户判断属性之间是否独立
- 回归分析（Regression Analysis）：如果有关联，能否根据 X 的取值预测 Y 得取值？当时数据中有缺失值时能不能通过其他属性评估缺失值的大小？

	X	Y	Z
数据元素 1	0.1	0.1	0.1
	0.2	0.2	0.2
	0.5	0.5	0.5
	0.3	0.3	0.3
	0.7	0.7	0.7
	0.4	0.4	0.4
	0.2	0.2	0.2
数据元素 N	0.5	0.5	0.5

相关性分析

- 相关性分析用于解释数据属性之间关联度的强弱
- 相关性分析揭示了 一个属性的取值 是如何 跟随另外一个属性的取值 发生变化的内在规律
- 相关性分析不同于因果关系分析

相关性分析

- 方差（variance）、标准差（standard deviation）及 协方差（covariance）

  - **方差及标准差：**方差及标准差是用来衡量 单一变量（属性）取值变化程度的统计量

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

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

  - **协方差：**当一个变量Y 的取值 (y1, y2, ..., yn) 随另外一个变量 X 的取值 (x1, x2, ..., xn) 变化而变化时，协方差是用来衡量相互间变化程度的统计量

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

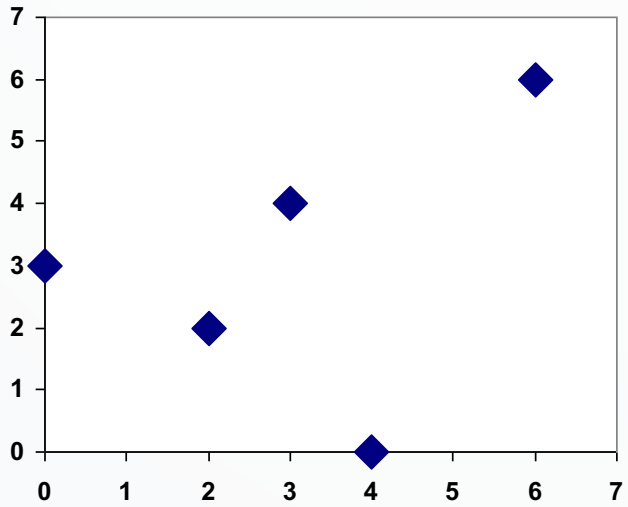
  - **自由度：**上边的公式中，n 是 所有取值样本的数量，n-1 被称为自由度，反映了所有能够自由改变的样本数量，当均值固定时，n 个样本中只有 n-1 个样本可以自由取值，第n 个样本的取值可以通过 均值及 前n-1 个样本计算出来

相关性分析 – 协方差

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- When X↑ and Y↑ : cov (x,y) = pos.
- When X↓ and Y↑ : cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

相关性分析 - 协方差



x	y	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
0	3	3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
$\bar{x}=3$	$\bar{y}=3$			$\sum=7$

$$\mathrm{cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}))}{n-1}=\frac{7}{4}=1.75$$

这个数值告诉了我们什么?

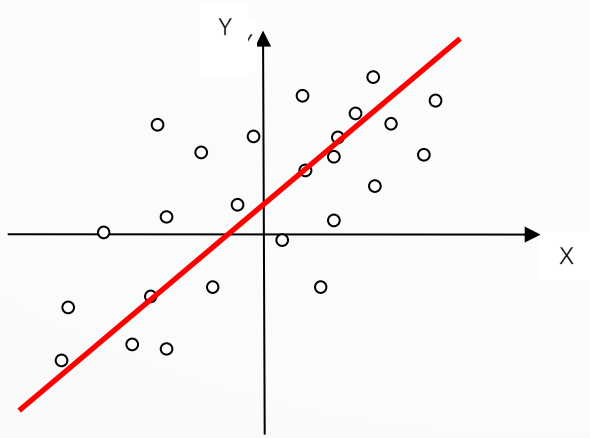
相关性分析 – 协方差

	High variance data				Low variance data		
Subject	**x**	**y**	**x error * y error**		**x**	**y**	**x error * y error**
1	101	100	2500		54	53	9
2	81	80	900		53	52	4
3	61	60	100		52	51	1
4	51	50	0		51	50	0
5	41	40	100		50	49	1
6	21	20	900		49	48	4
7	1	0	2500		48	47	9
Mean	51	50			51	50	
Sum of x error * y error :			7000		**Sum of x error * y error :**		28
Covariance:			**1166.67**		**Covariance:**		**4.67**

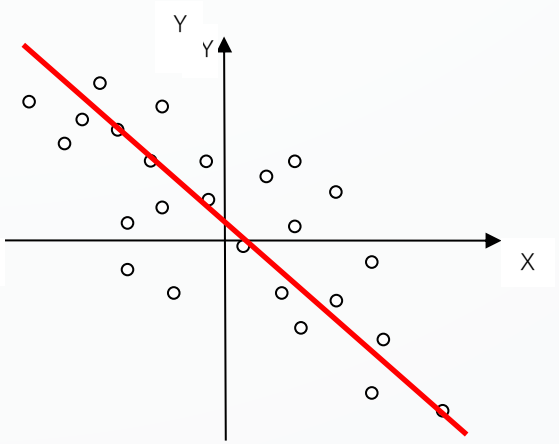
相关性分析 – 皮尔森系数

- 协方差并不能真正告诉我们有用的信息
- 皮尔森系数就是使用标准差对协方差进行正则化处理，让计算数值变得可以相互比较

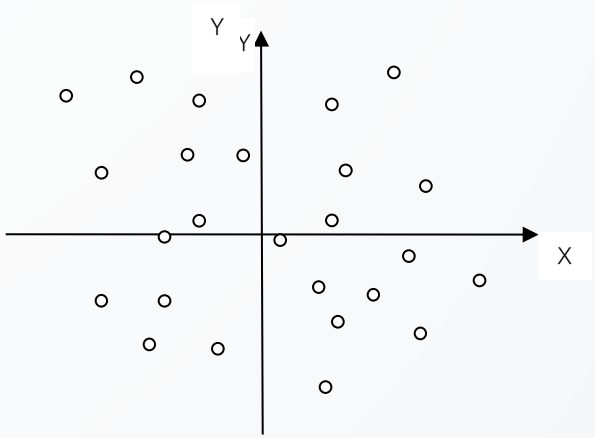
$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$



Positive correlation



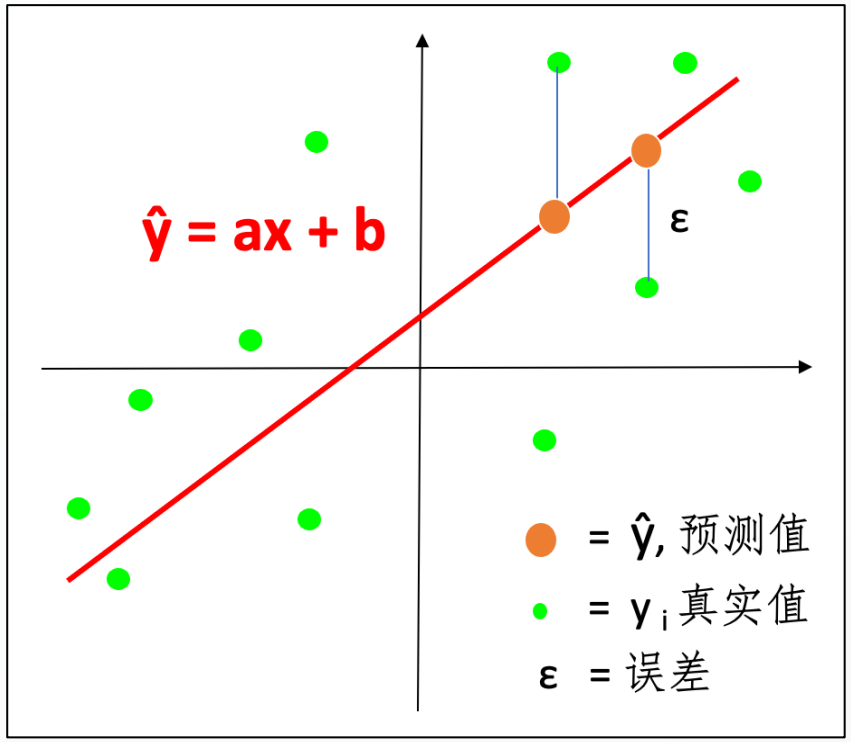
Negative correlation



No correlation

回归分析

- 相关性分析可以告诉我们 属性 X 与 属性 Y 之间是否相互关联
- 回归分析可以让我们进一步根据一个变量 X 的取值 预测 另一个变量 Y 的取值
- 以最简单的线性回归为例，在空间中找到一根直线来近似刻画数据属性之间的联系



回归分析

- 为了找到最优的直线，我们需要尽可能减小预测的误差，需要求解一下的优化问题

模型直线: $\hat{y} = ax + b$ a = 斜率, b = 截距

误差 $(\varepsilon) = y - \hat{y}$

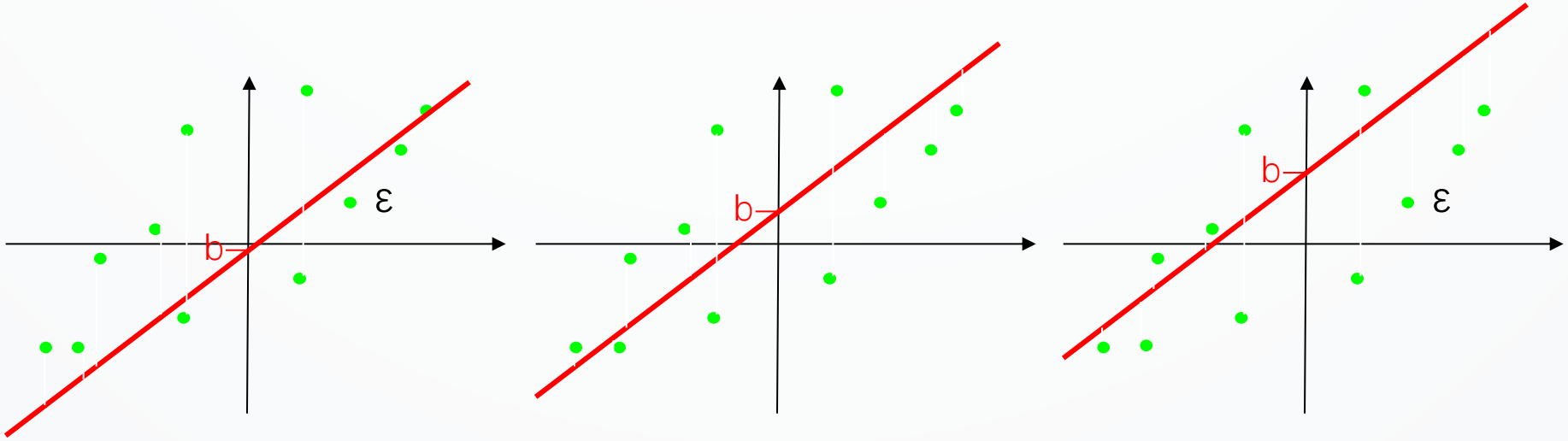
误差的平方和 $= \Sigma (y - \hat{y})^2$

找到使误差的平方和最小的 斜率a 及 截距b

$\min \Sigma (y - \hat{y})^2$

Finding b

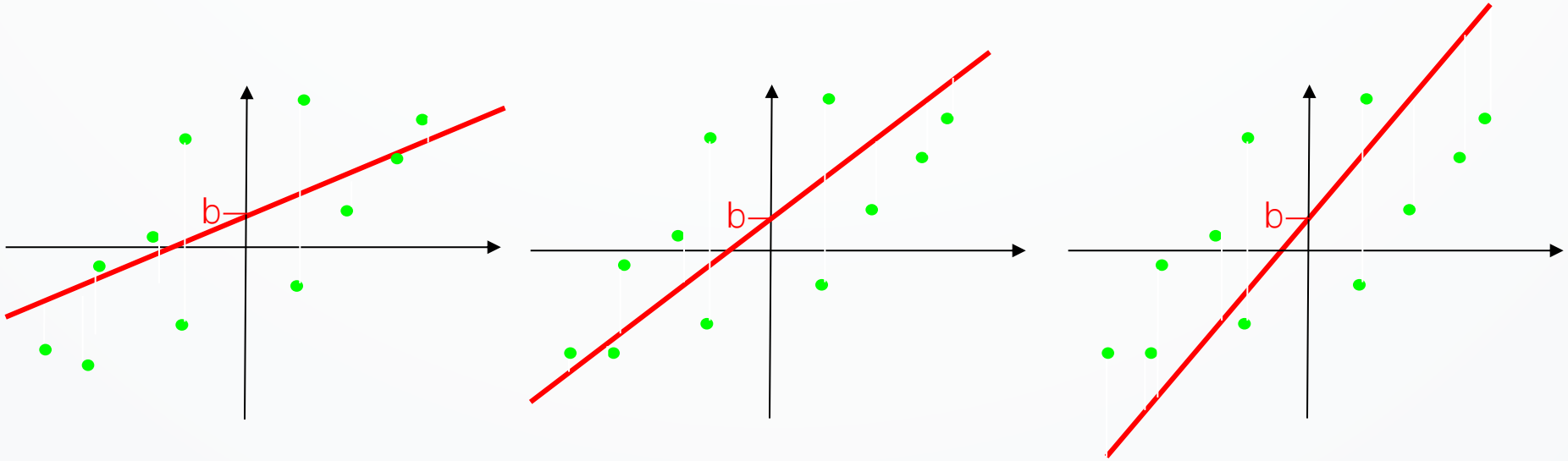
First we find the value of b that gives the min sum of squares



Trying different values of b is equivalent to shifting the line up and down the scatter plot

Finding a

Find the value of a that gives the min sum of squares



Trying out different values of a is equivalent to changing the slope of the line, while b stays constant

课程总结

- 本节课
  - 数据元素及元素的属性
  - 数据元素之间的差异与聚类分析
  - 数据属性之间的关联

- 下节课
  - 可视化的基本设计准则