

无监督学习

黄书剑



- 无监督学习
- 聚类分析
 - k均值聚类
- 关联规则
- 异常检测

无监督学习 v.s. 有监督学习

- **Un-Supervised learning** is the machine learning task of learning **a function that maps an input to an output** based on example input-output pairs. (监督学习/有指导学习/指导学习)
 - mapping input to output
 - with input-output pairs
- Input x / \vec{x} , Output y
- Input output pair (x, y)
- Examples (x_i, y_i)

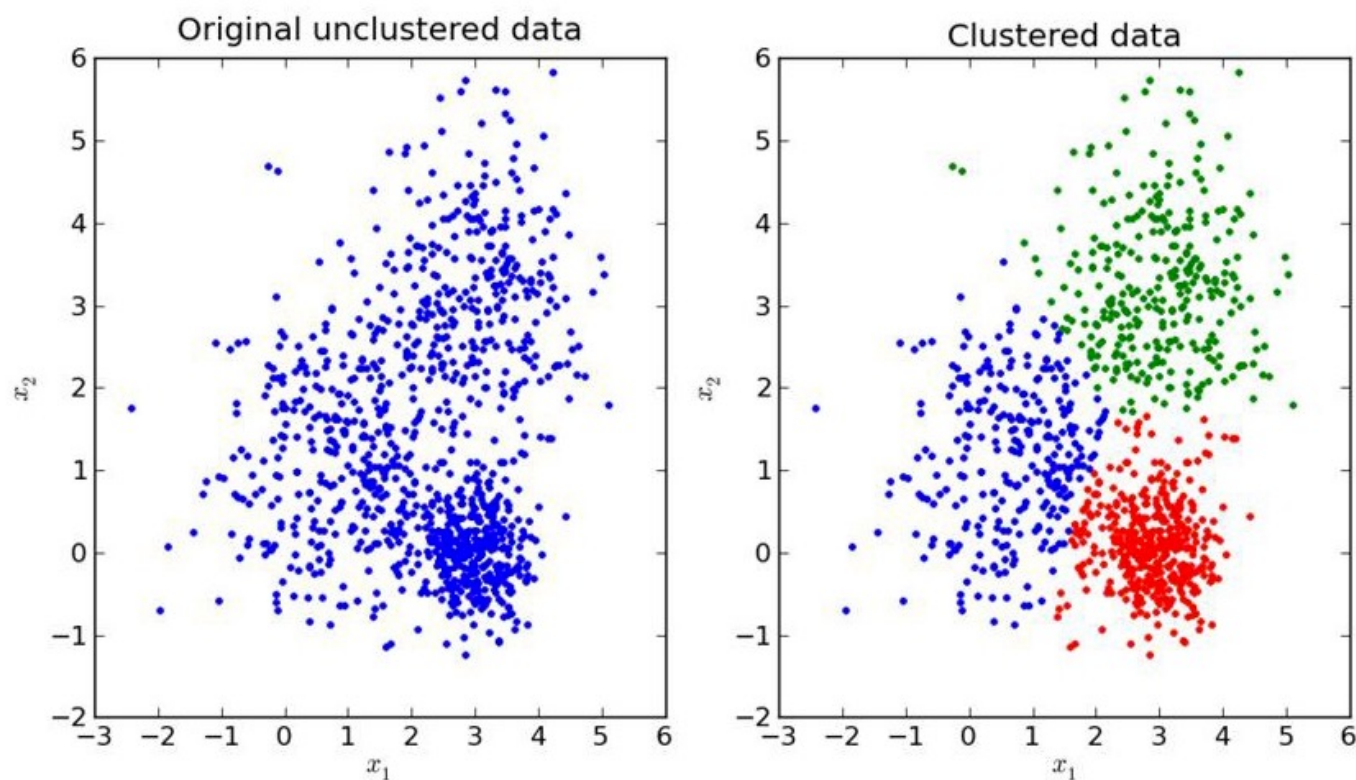
- It may seem somewhat mysterious to imagine what the machine could possibly learn given that it doesn't get any feedback from its environment.
- However, it is possible to develop of formal framework for unsupervised learning based on the notion that the machine's goal is to **build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc.**

Ghahramani (2004) [Unsupervised Learning](#). In Bousquet, O., Raetsch, G. and von Luxburg, U. (eds) Advanced Lectures on Machine Learning LNAI 3176. Springer-Verlag.

无监督学习

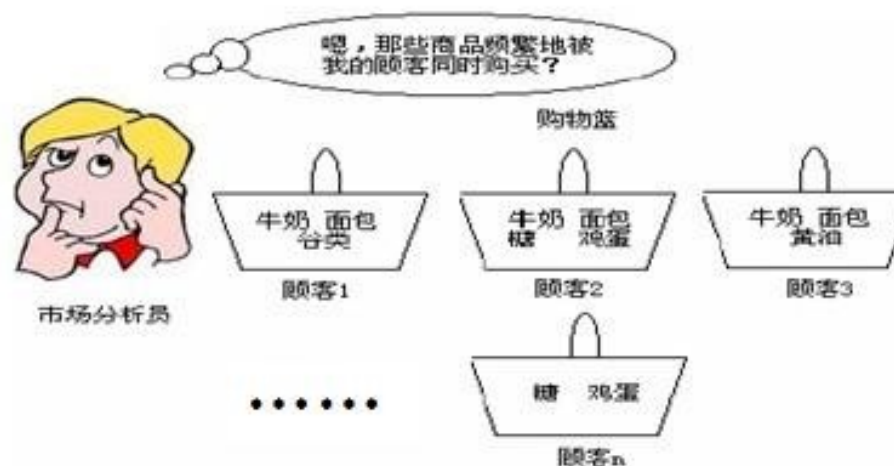
- **Unsupervised machine learning algorithms infer patterns from a dataset without reference to known, or labeled, outcomes.**
- **“Mining”/ infer patterns from examples x_i**
- **维度约简 Dimension Reduction**
- **聚类 Clustering**
- **关联规则 Association Rule Mining**
- **异常检测 Anomaly Detection**

- 将输入按照其分布划分为若干不同的类别



关联规则

- 发掘元素集合中潜在的关联性
 - 商品布局、购物习惯分析

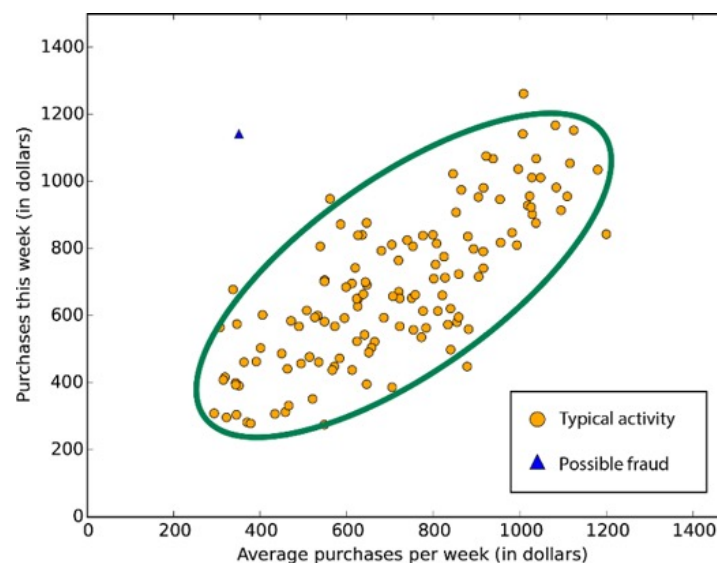
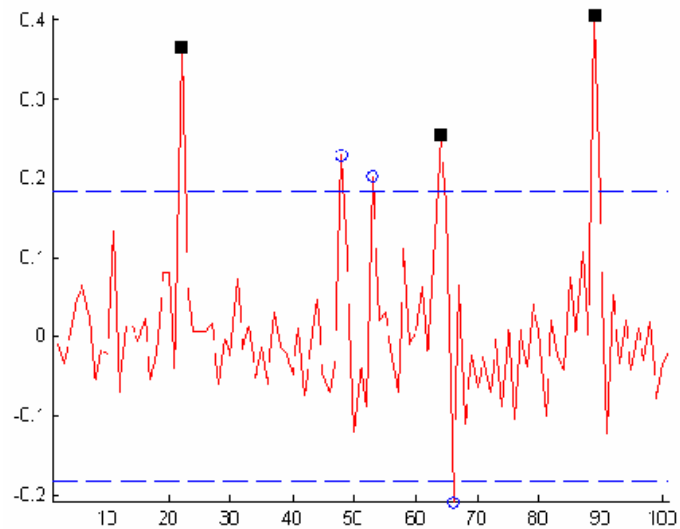


TID	Items
T1	{牛奶,面包}
T2	{面包,尿布,啤酒,鸡蛋}
T3	{牛奶,尿布,啤酒,可乐}
T4	{面包,牛奶,尿布,啤酒}
T5	{面包,牛奶,尿布,可乐}
...	...

➡ {牛奶,面包,尿布} !

异常检测

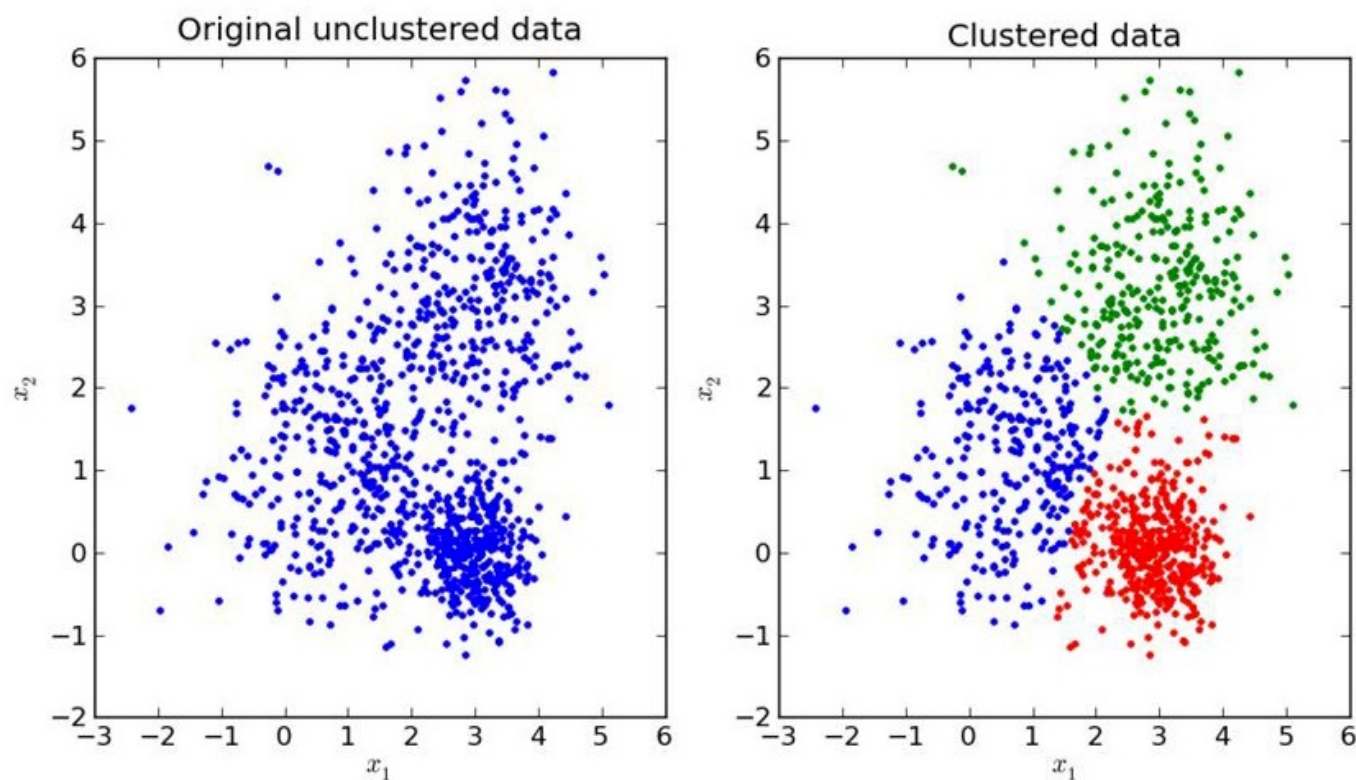
- 发掘数据中包含的不一致性
 - 消除噪音干扰，提高分析精度
 - 检测异常行为（系统故障、欺诈等）



- 无监督学习
- 聚类分析
 - k均值聚类
- 关联规则
- 异常检测

聚类（回顾）

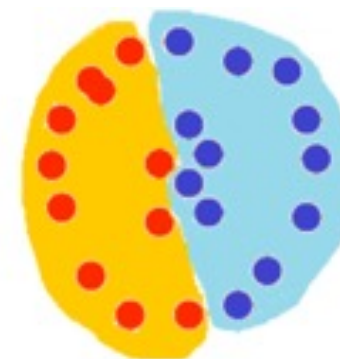
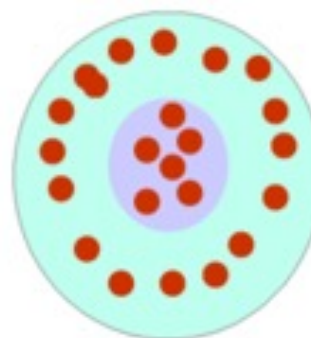
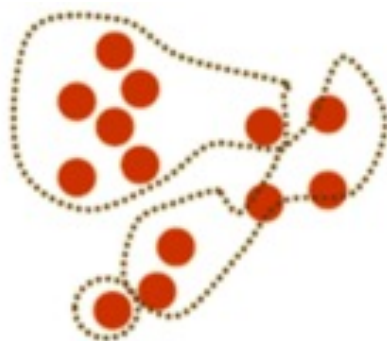
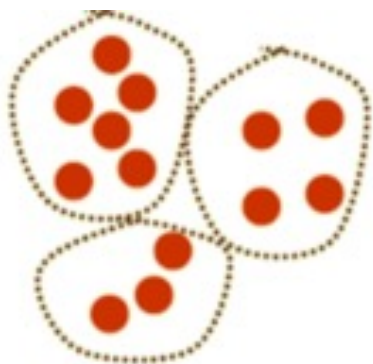
- 将输入按照其分布划分为若干不同的类别



聚类

- 目标

- 将样本划分为若干类别
- 原则：邻近的样本可能关系比较紧密



聚类的评价方法（内部指标）

- 仅考察当前聚类结果
 - 簇内相似度高intra-cluster similarity
 - 簇间相似度高inter-cluster similarity

$$\text{avg}(C) = \frac{2}{|C|(|C| - 1)} \sum_{1 \leq i < j \leq |C|} \text{dist}(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{diam}(C) = \max_{1 \leq i < j \leq |C|} \text{dist}(\mathbf{x}_i, \mathbf{x}_j)$$

$$d_{\min}(C_i, C_j) = \min_{\mathbf{x}_i \in C_i, \mathbf{x}_j \in C_j} \text{dist}(\mathbf{x}_i, \mathbf{x}_j)$$

$$d_{\text{cen}}(C_i, C_j) = \text{dist}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$$

聚类的评价方法（外部指标）

- 已知一个预期(oracle)，考察聚类是否符合预期

- 比较两个结果是否相同

- 定义辅助数值（oracle用*标识），如：

$$a = |SS|, \quad SS = \{(\mathbf{x}_i, \mathbf{x}_j) | \lambda_i = \lambda_j, \lambda_i^* = \lambda_j^*, i < j\}$$

$$b = |SD|, \quad SD = \{(\mathbf{x}_i, \mathbf{x}_j) | \lambda_i = \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\}$$

$$c = |DS|, \quad DS = \{(\mathbf{x}_i, \mathbf{x}_j) | \lambda_i \neq \lambda_j, \lambda_i^* = \lambda_j^*, i < j\}$$

$$d = |DD|, \quad DD = \{(\mathbf{x}_i, \mathbf{x}_j) | \lambda_i \neq \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\}$$

- 可以计算指标如Jaccard系数：

$$JC = \frac{a}{a + b + c}$$

聚类的评价方法

- 一个好的聚类方法：
 - 应该有较好的聚类表现（内部指标）
 - 不一定与预期相吻合（外部指标）
- 结合实际情况进行分析和改进！

k-Means聚类

- 一种针对聚类中心进行优化的方法：

$$E = \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2$$

step 1. 初始化k个聚类中心 $\boldsymbol{\mu}_i$

step 2. 将每个x划入到距离最近的聚类中心 $\boldsymbol{\mu}_i$

step 3. 重新计算每个类的中心 $\boldsymbol{\mu}_i$

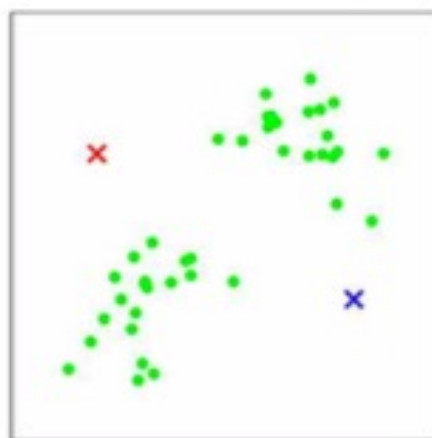
step 4. 转至2继续执行，直至聚类不发生变化

- 迭代进行聚类中心和类别选择的改进！

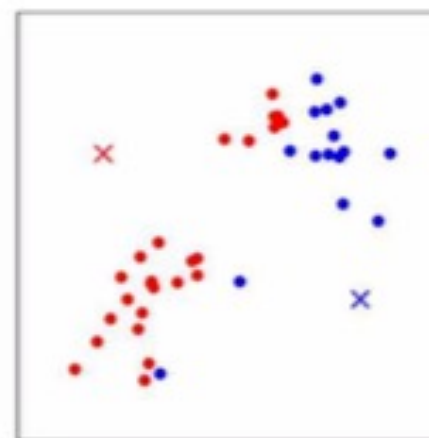
K means的中心变化



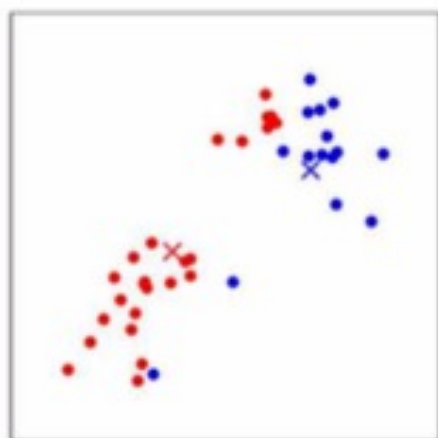
(a)



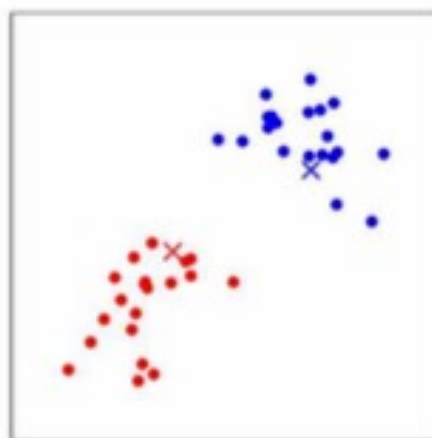
(b)



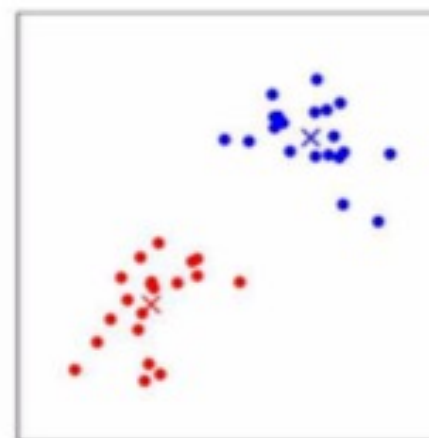
(c)



(d)



(e)



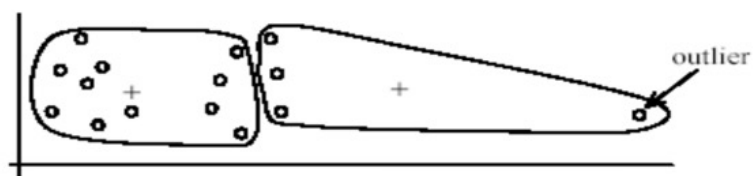
(f)

一些简单的讨论

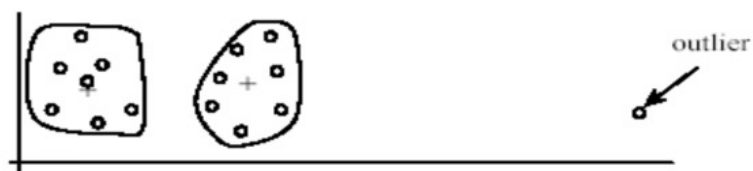
- 聚类数目



- 异常点



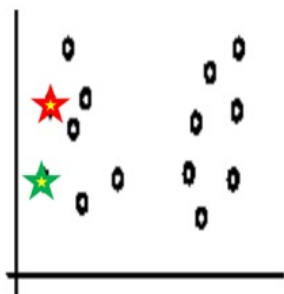
(A): Undesirable clusters



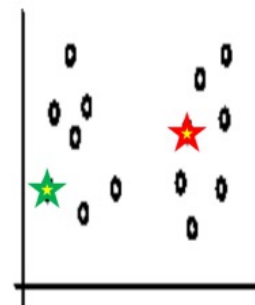
(B): Ideal clusters

一些简单的讨论

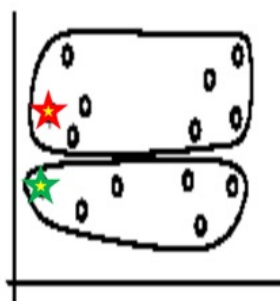
- 初始点



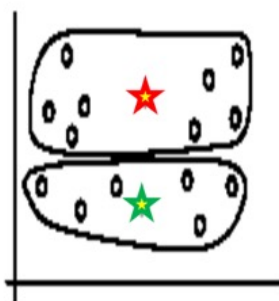
Random selection of seeds (centroids)



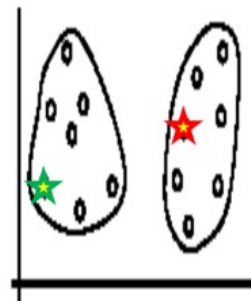
Random selection of seeds (centroids)



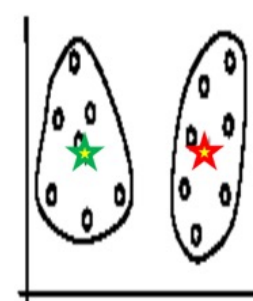
Iteration 1



Iteration 2



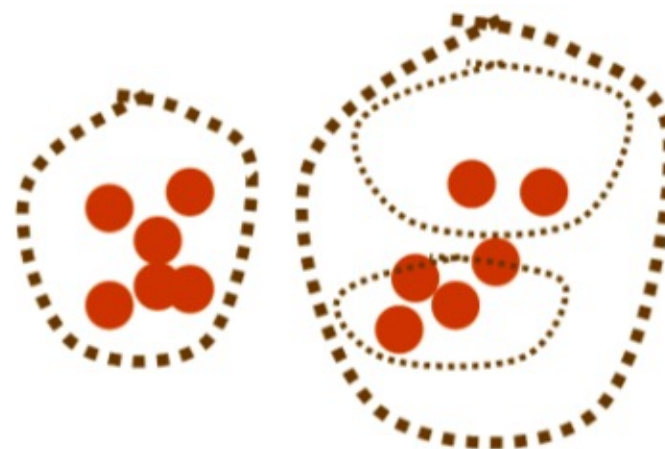
Iteration 1



Iteration 2

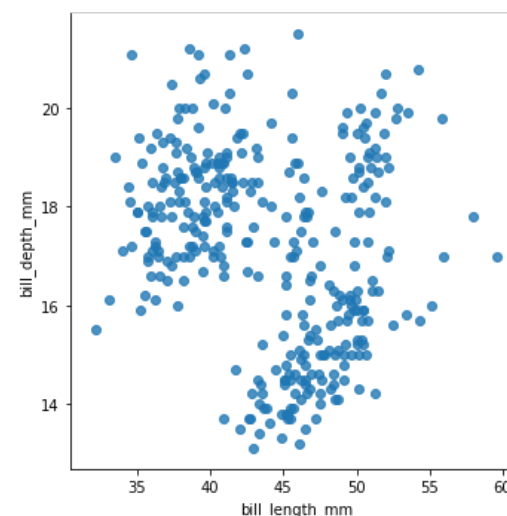
一些简单的讨论

- 类别层次性（层次聚类）
 - 不断改进已有的聚类结果去得到新的更好的聚类
 - 基于聚合（不断组合）、基于划分（不断分割）



有监督的聚类

- 已知部分样本之间的关联
 - 如"must link" and "cannot link"
 - constrained k-means
- 已知部分样本的类别信息
 - constrained seed k-means
- 一个好的聚类方法：
 - 应该有较好的聚类表现（内部指标）
 - 不一定与预期相吻合（外部指标）



聚类小结

- 从数据中发掘分布相关的特点
- K-Means聚类：简单有效的迭代式方法
- 注意以下因素对结果的影响：
 - 类别数目
 - 初始点选择
 - 噪音

- 无监督学习
- 聚类分析
 - k均值聚类
- 关联规则
- 异常检测

关联分析

- 发掘元素集合中潜在的关联性
 - 商品布局、购物习惯分析
 - 网页访问日志
 - 基因关联性

TID	Items
t1	{牛奶,面包}
t2	{面包,尿布,啤酒,鸡蛋}
t3	{牛奶,尿布,啤酒,可乐}
t4	{面包,牛奶,尿布,啤酒}
t5	{面包,牛奶,尿布,可乐}
...	...



{牛奶,面包,尿布} !

{牛奶,面包} → {尿布} !

基本概念

- 项/元素 (item)
 - 如：面包、牛奶
- 项集 (itemset)
 - 如：{面包、牛奶}
- k-项集 (k-itemset)
 - 有k个项的项集
- 事务 (transaction)
 - 如： t_2 : {面包,尿布,啤酒,鸡蛋}
 - 事务中项的个数，也称为事务的宽度
 - 给定一系列事务的集合记为T

TID	Items
t1	{牛奶,面包}
t2	{面包,尿布,啤酒,鸡蛋}
t3	{牛奶,尿布,啤酒,可乐}
t4	{面包,牛奶,尿布,啤酒}
t5	{面包,牛奶,尿布,可乐}
...	...

基本概念

- 项/元素、项集、k-项集、事务
- 关联分析
 - 从给定事务集合T中发掘：频繁项集 (Frequent Itemset) 和关联规则 (Association Rule)
- 关联规则
 - $X \rightarrow Y$: X和Y是两个不相交的项集
 - 如：{牛奶,面包} \rightarrow {尿布}

基本概念

- 项/元素、项集、k-项集、事务、关联规则
- 关联分析: 频繁项集和关联规则
- 重要程度:
 - 在T中出现次数计为 σ :
 - $\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$
 - 支持度support:
 - 给定事务集合T中出现的频繁程度 (概率 $p(X)$)
 - $s(X) = \frac{\sigma(X)}{N}$, $s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$
 - 置信度confidence:
 - 关联规则的可靠程度 (条件概率 $p(Y|X)$)
 - $c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

实例:

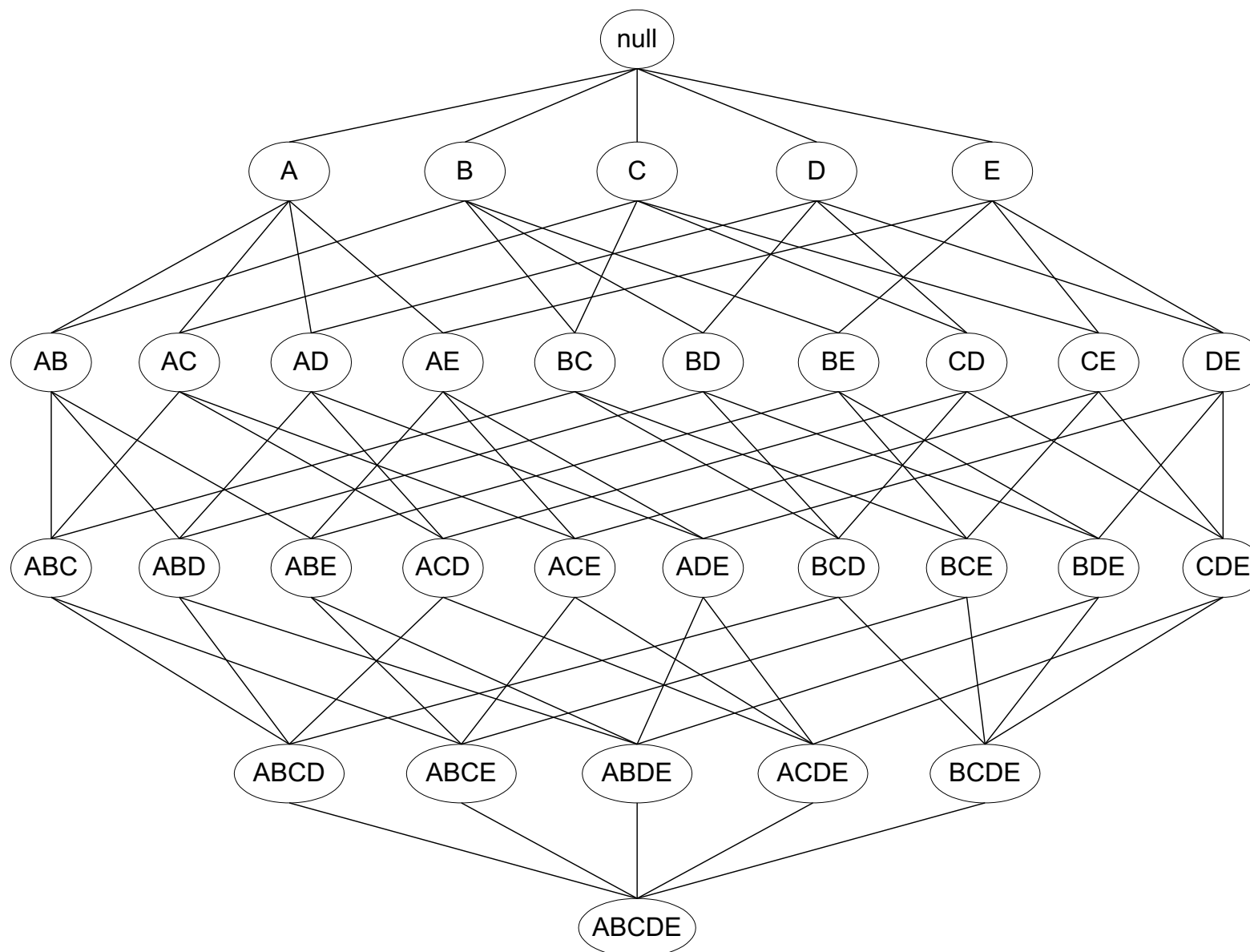
- 给定右图的事务集合
- 要求 $s > 0.5$, $c > 0.5$

TID	Items
t1	{牛奶,面包}
t2	{面包,尿布,啤酒,鸡蛋}
t3	{牛奶,尿布,啤酒,可乐}
t4	{面包,牛奶,尿布,啤酒}
t5	{面包,牛奶,尿布,可乐}

- 频繁项集:
 - {牛奶} 0.8、{面包} 0.8、{尿布} 0.8、{啤酒} 0.6
 - {牛奶,面包} 0.6、{牛奶,尿布} 0.6、{面包,尿布} 0.6、{啤酒,尿布} 0.6
- 关联规则:
 - {啤酒} \rightarrow {尿布} 0.6, 1
 - {尿布} \rightarrow {啤酒} 0.6, 0.75
 - {牛奶} \rightarrow {面包} 0.6, 0.75
 - {面包} \rightarrow {牛奶} 0.6, 0.75
 -

注意: $A \rightarrow B$ 和 $B \rightarrow A$ 具有不同的置信度

首先考虑频繁项集

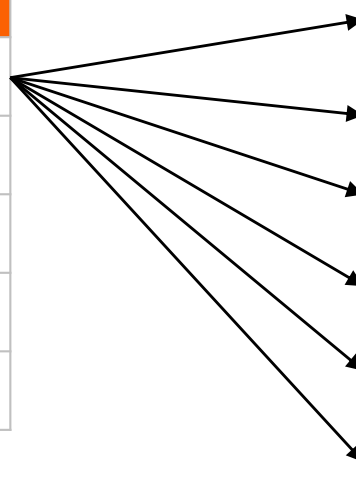


蛮力方法 (Brute-force)

- 穷举所有可能的项集，尝试匹配所有候选项集，依次为成功的匹配计数
 - $O(NMw)$
 - N 为事务数、 w 为事务的宽度
 - M 为项集候选数 ($2^n - 1$)， n 为元素个数

TID	Items
t1	{牛奶,面包}
t2	{面包,尿布,啤酒,鸡蛋}
t3	{牛奶,尿布,啤酒,可乐}
t4	{面包,牛奶,尿布,啤酒}
t5	{面包,牛奶,尿布,可乐}

候选项集	计数
{xxx}	
{xxx}	
{xxx}	
{xxx}	
{xxx}	
...	...



Apriori原理

- 频繁项集的子集一定是频繁的 (Any subset of a frequent itemset must be frequent)
 - 例如：如果{牛奶,尿布,啤酒}是频繁的，{尿布,啤酒}一定是频繁的
 - 简单证明：
 - 任何包含某项集的事务，一定包含其子项集
 - 子项集的频繁程度一定高于其超集

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

- 等价于：不频繁项集的超集一定是不频繁的

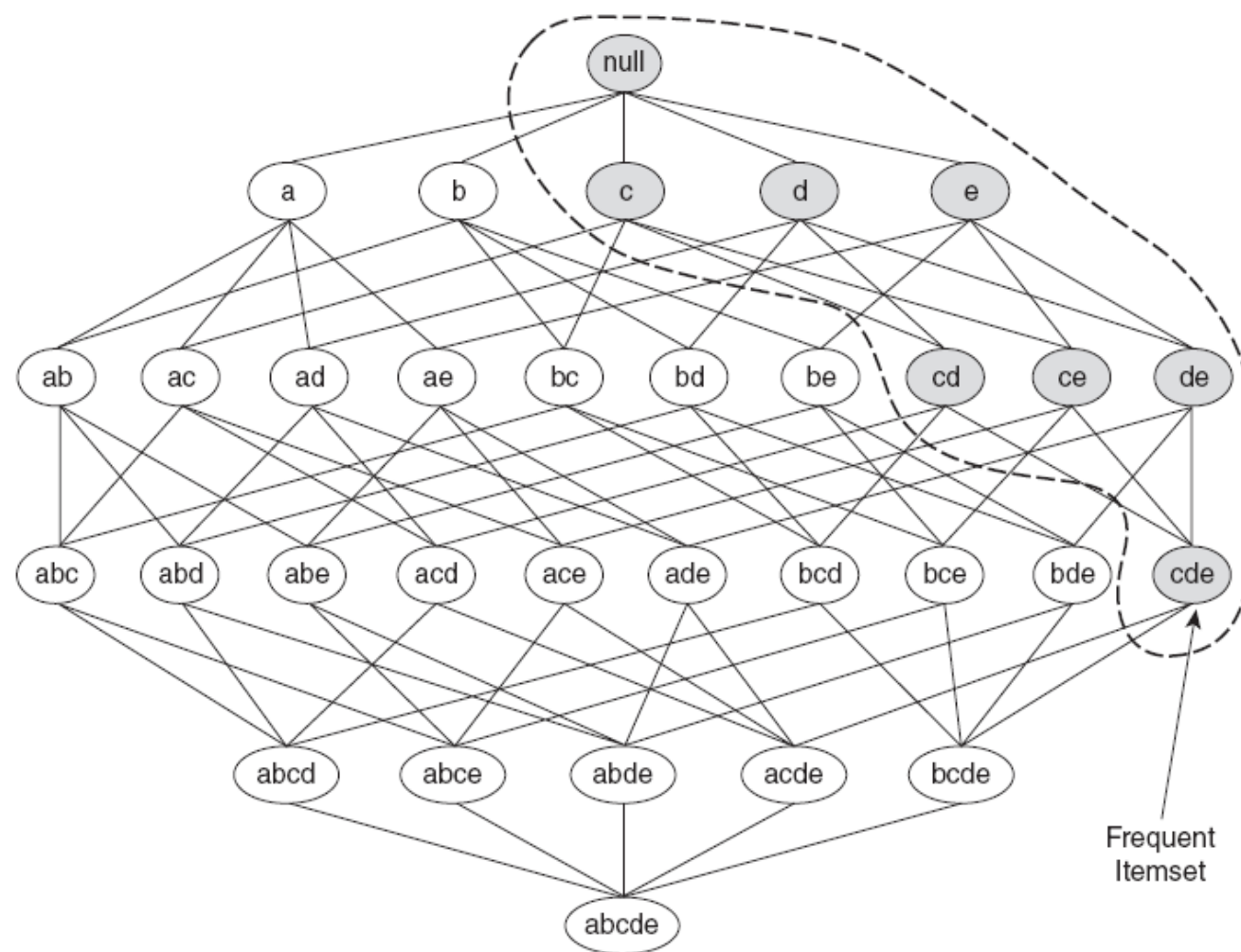
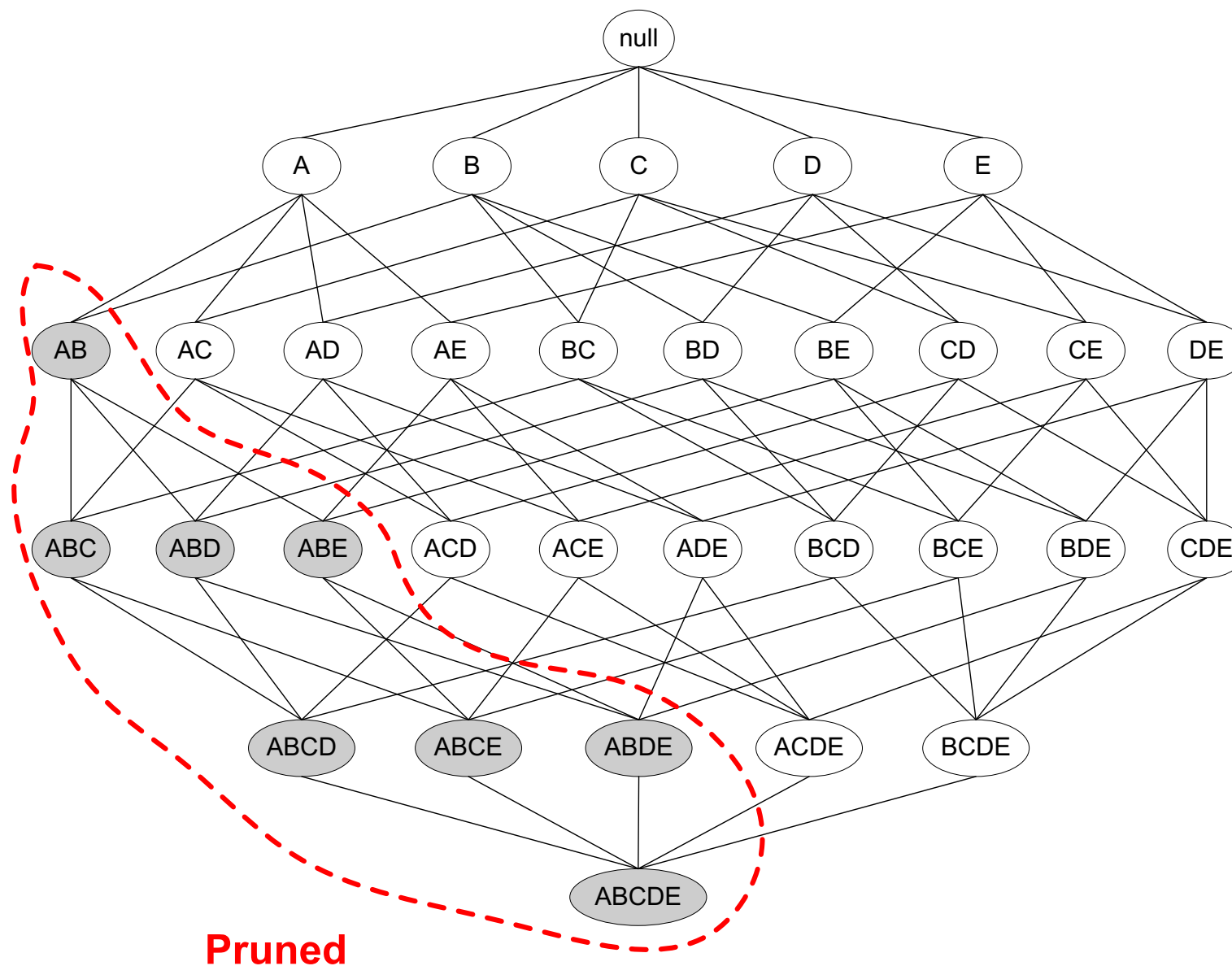


Figure 6.3. An illustration of the *Apriori* principle. If $\{c, d, e\}$ is frequent, then all subsets of this itemset are frequent.



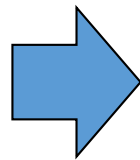
Apriori算法:

- 基于频繁项集候选的生成和检测的方法
 - Apriori原理表明: 如果一个项集是非频繁的, 那么它的所有超集都不应该被生成或者检测
- 流程:
 - 扫描数据, 生成宽度为1的频繁项集
 - 从k-频繁项集生成k+1-频繁项集的候选 (Apriori)
 - 扫描数据, 对候选进行测试, 得到k+1-频繁项集
 - 重复直至不再生成新的频繁项集或者新的候选

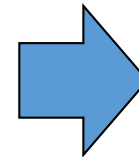
TID	Items
t1	{牛奶,面包}
t2	{面包,尿布,啤酒,鸡蛋}
t3	{牛奶,尿布,啤酒,可乐}
t4	{面包,牛奶,尿布,啤酒}
t5	{面包,牛奶,尿布,可乐}

项集	计数
{鸡蛋}	1
{可乐}	2
{面包}	4
{尿布}	4
{牛奶}	4
{啤酒}	3

$$s > 0.5$$



项集	计数
{面包, 尿布}	3
{面包, 牛奶}	3
{面包, 啤酒}	2
{尿布, 牛奶}	3
{尿布, 啤酒}	3
{牛奶, 啤酒}	2



项集	计数
{面包, 尿布, 牛奶}	2
{面包, 尿布, 啤酒}	2
{尿布, 牛奶, 啤酒}	2

候选项集数:

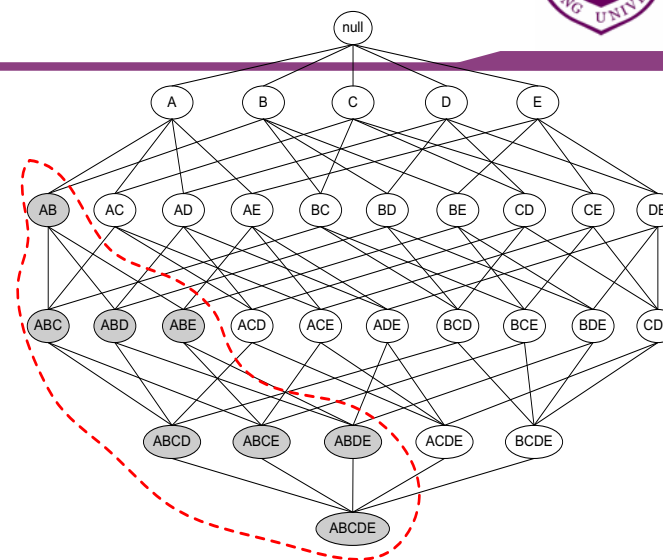
$$C_6^1 + C_6^2 + C_6^3 = 6 + 15 + 20 = 41$$

$$C_6^1 + C_4^2 + 3 = 6 + 6 + 3 = 15 \quad 15/41=36.7\%$$

如何高效的生成候选集合?

- 如何更高效的生成候选?

- 避免重复
- 减少候选数量
- 不能遗漏频繁的候选



- 蛮力方法

- 穷举所有可能，并按照前述剪枝

- $F_{k-1} * F_1$

- 从已有的k-1频繁项集扩展

- $F_{k-1} * F_{k-1}$

- $F_{k-1} * F_{k-1}$ 且 所有的k-1子项都应该是频繁的

F_1 :

项集	计数
{面包}	4
{尿布}	4
{牛奶}	4
{啤酒}	3

$F_{k-1} * F_1$:

项集	计数
{面包, 尿布, 牛奶}	2
{尿布, 牛奶, 啤酒}	2
{面包, 牛奶, 啤酒}	1
{面包, 尿布, 啤酒}	2

F_{k-1} :

项集	计数
{面包, 尿布}	3
{面包, 牛奶}	3
{尿布, 牛奶}	3
{尿布, 啤酒}	3

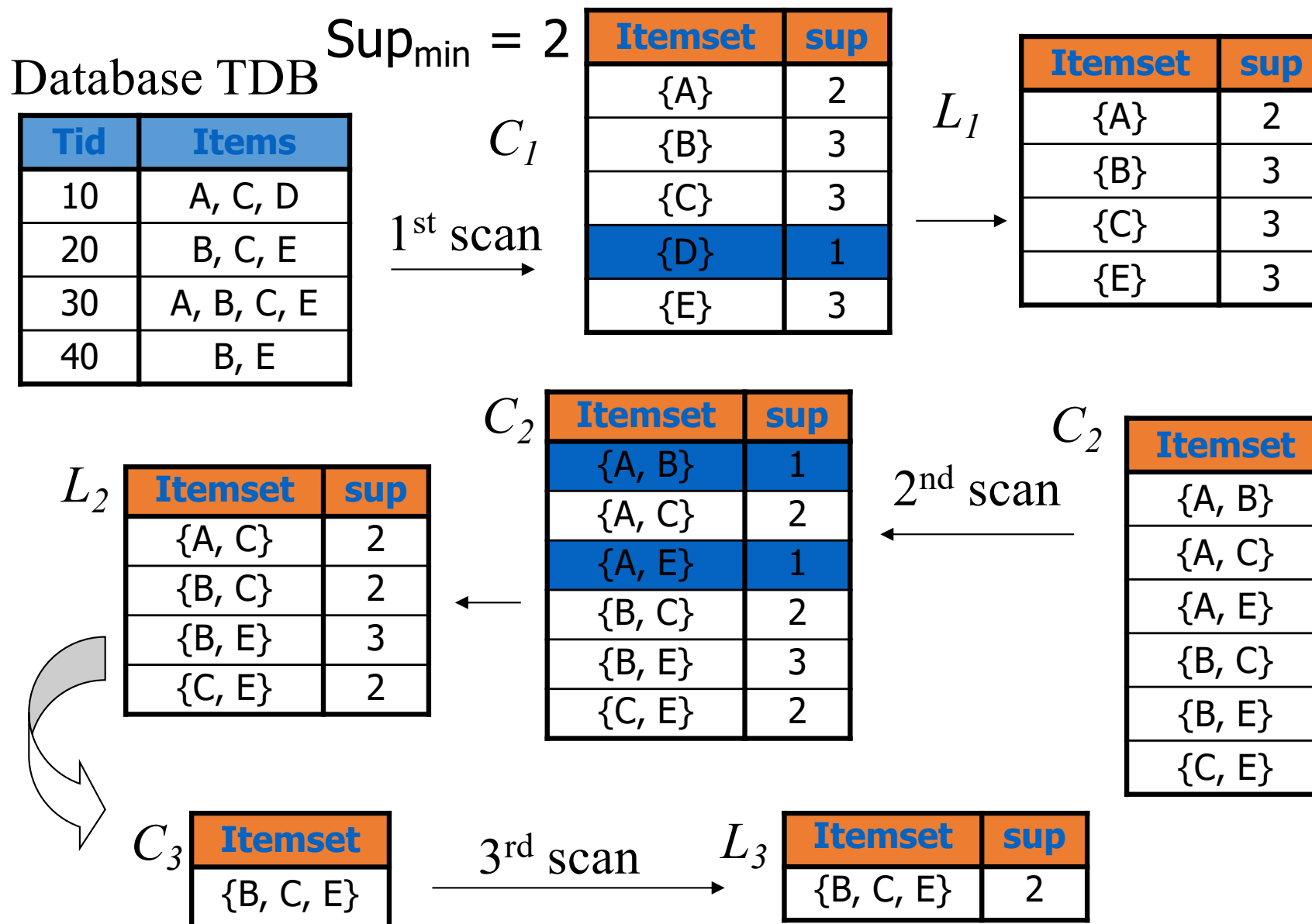
$F_{k-1} * F_{k-1}$:

项集	计数
{面包, 尿布, 牛奶}	2
{尿布, 牛奶, 啤酒}	2
{面包, 尿布, 啤酒}	2

$F_{k-1} * F_{k-1}$ and all subset frequ.:

项集	计数
{面包, 尿布, 牛奶}	2

The Apriori Algorithm—An Example



基本概念（回顾）

- 项/元素、项集、k-项集、事务、关联规则
- 关联分析: 频繁项集和关联规则
- 重要程度:
 - 在T中出现次数计为 σ :
 - $\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$
 - 支持度support:
 - 给定事务集合T中出现的频繁程度（概率 $p(X)$ ）
 - $s(X) = \frac{\sigma(X)}{N}$, $s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$
 - 置信度confidence:
 - 关联规则的可靠程度（条件概率 $p(Y|X)$ ）
 - $c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

关联规则生成

- 关联规则($X \rightarrow Y$)的项集 $X \cup Y$ 是频繁的
 - 首先得到符合支持度要求的k-频繁项集（记为Y）
 - 将该项集划分为两个非空子集X和Y-X，则得到关联规则（ $X \rightarrow Y-X$ ）
 - 逐层生成，每层规则后件的项数增大
 - 检查置信度要求
- 若X'为X的子集，如果规则 $X \rightarrow Y-X$ 不满足置信度要求，则 $X' \rightarrow Y-X'$ 也一定不满足(Apriori原理)
 - $c(X \rightarrow Y - X) = \frac{\sigma(X \cup (Y-X))}{\sigma(X)} = \frac{\sigma(Y)}{\sigma(X)}$

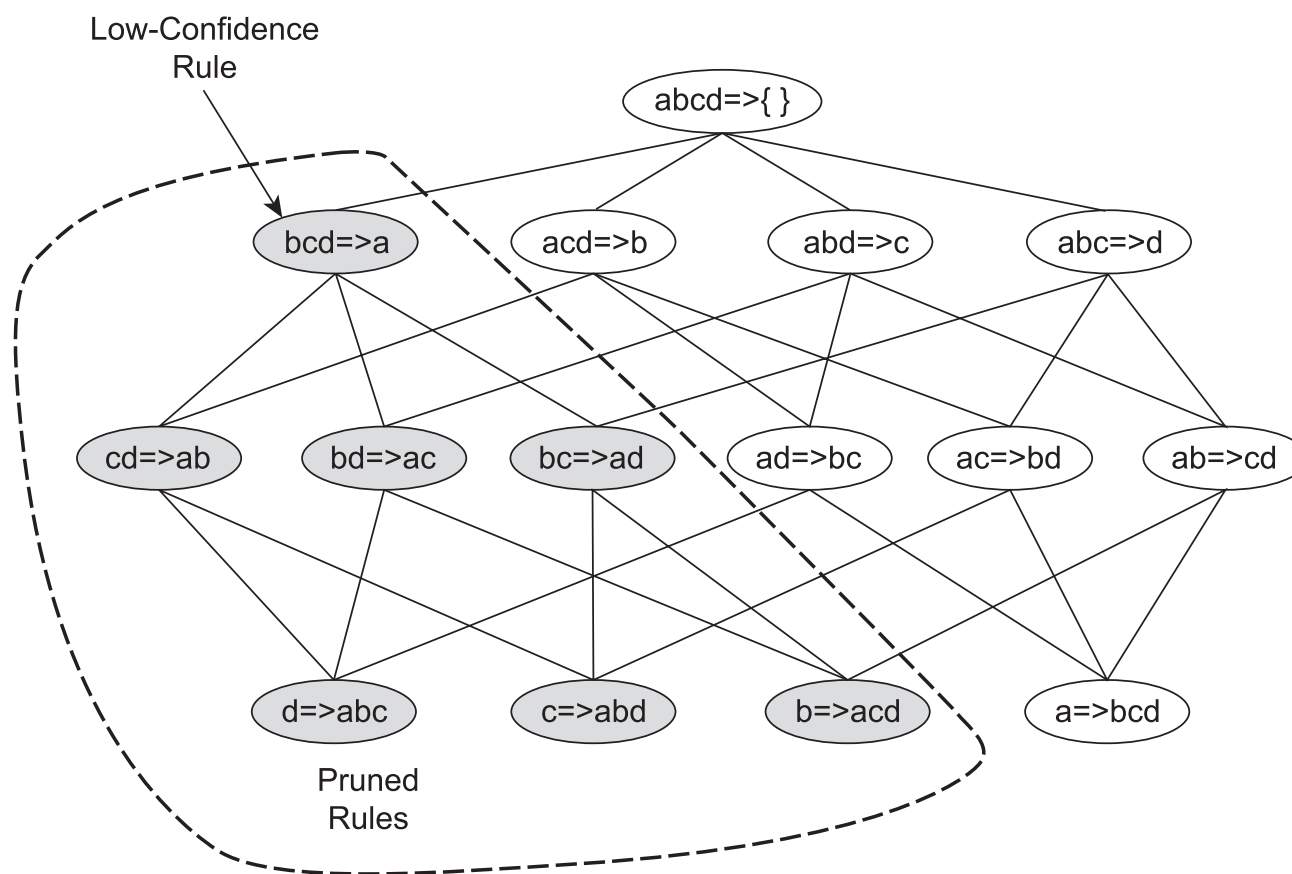


Table 5.3. List of binary attributes from the 1984 United States Congressional Voting Records. Source: The UCI machine learning repository.

1. Republican	18. aid to Nicaragua = no
2. Democrat	19. MX-missile = yes
3. handicapped-infants = yes	20. MX-missile = no
4. handicapped-infants = no	21. immigration = yes
5. water project cost sharing = yes	22. immigration = no
6. water project cost sharing = no	23. synfuel corporation cutback = yes
7. budget-resolution = yes	24. synfuel corporation cutback = no
8. budget-resolution = no	25. education spending = yes
9. physician fee freeze = yes	26. education spending = no
10. physician fee freeze = no	27. right-to-sue = yes
11. aid to El Salvador = yes	28. right-to-sue = no
12. aid to El Salvador = no	29. crime = yes
13. religious groups in schools = yes	30. crime = no
14. religious groups in schools = no	31. duty-free-exports = yes
15. anti-satellite test ban = yes	32. duty-free-exports = no
16. anti-satellite test ban = no	33. export administration act = yes
17. aid to Nicaragua = yes	34. export administration act = no

<https://archive.ics.uci.edu/ml/datasets/congressional+voting+records>

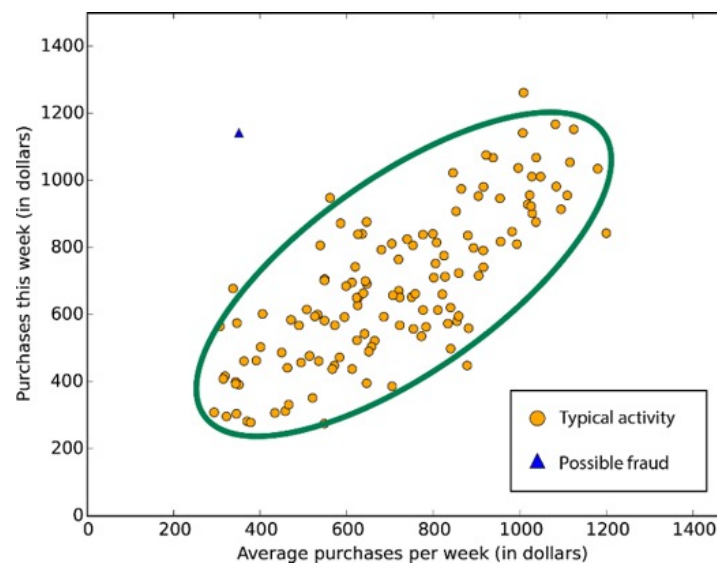
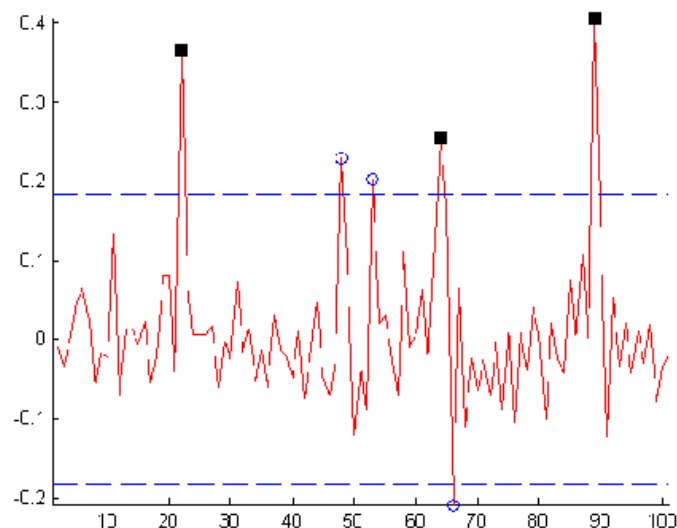
Association Rule	Confidence
{budget resolution = no, MX-missile=no, aid to El Salvador = yes } → {Republican}	91.0%
{budget resolution = yes, MX-missile=yes, aid to El Salvador = no } → {Democrat}	97.5%
{crime = yes, right-to-sue = yes, physician fee freeze = yes} → {Republican}	93.5%
{crime = no, right-to-sue = no, physician fee freeze = no} → {Democrat}	100%

<https://archive.ics.uci.edu/ml/datasets/congressional+voting+records>

- 无监督学习
- 聚类分析
 - k均值聚类
- 关联规则
 - Apriori
- 异常检测

异常检测

- 发掘数据中包含的不一致性、不规律（离群点）
 - 检测异常行为（系统故障、欺诈、入侵等）
 - 检测异常的状态（生态系统失调、流感疫情爆发等）



异常的成因

- 数据来源于不同的类别
 - 数据自然变异
 - 数据测量和收集的误差
 -
-
- 异常v.s.噪音
 - 噪音不一定导致异常的结果
 - 噪音的观察价值较小（倾向于随机发生）

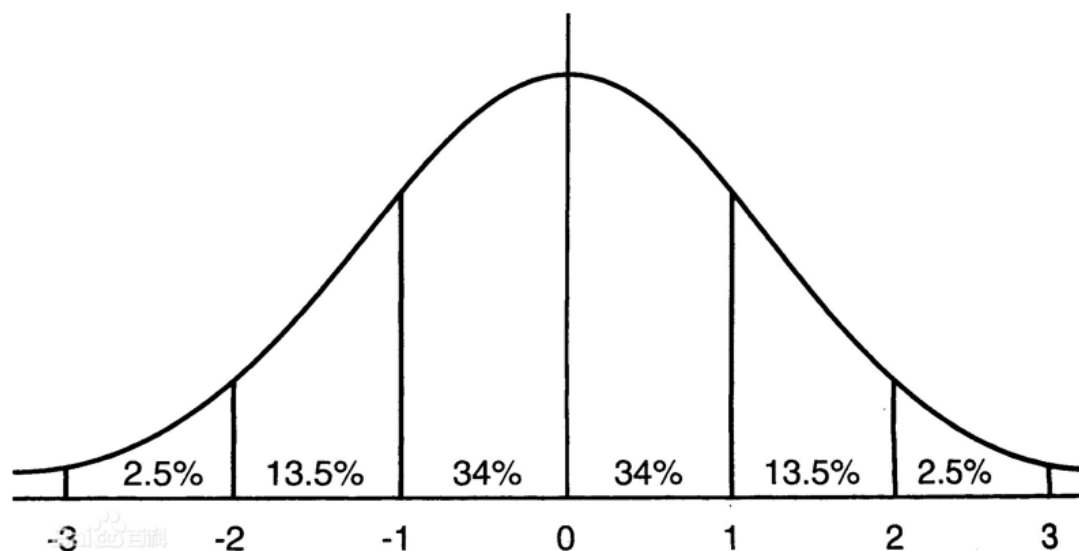
异常的检测

- 有监督的检测
- 无监督的检测
- 基于模型的技术
 - 正态分布
- 基于邻近度的技术
 - k-近邻
- 基于密度的技术

https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap9_anomaly_detection.pptx

基于正态分布的离群点预测

- 标准正态分布，是以0为均值、以1为标准差的正态分布，记为 $N(0, 1)$
 - 离群程度为其出现的概率

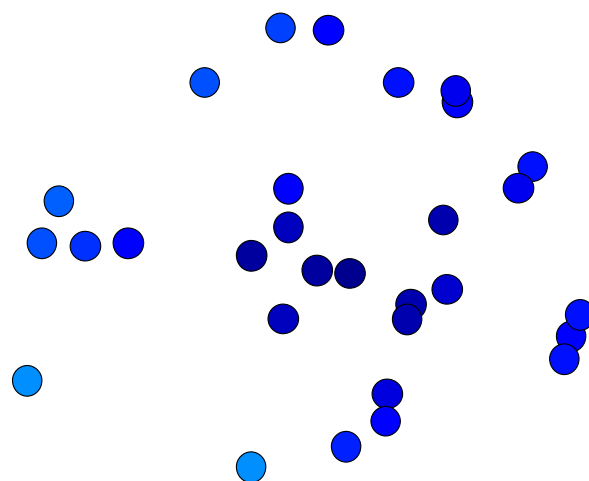


基于邻近密度的离群点检测

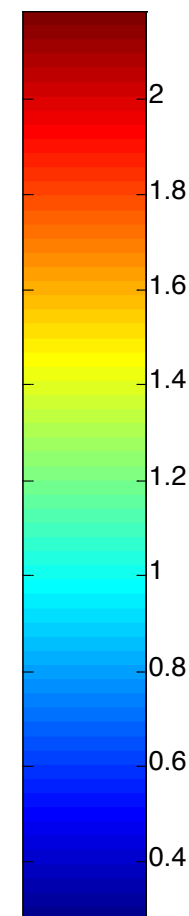
- 离群程度定义为其到k近邻的距离

– k=1

$$\text{density}(\mathbf{x}, k) = \frac{1}{|N(\mathbf{x}, k)|} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} d(\mathbf{x}, \mathbf{y})$$



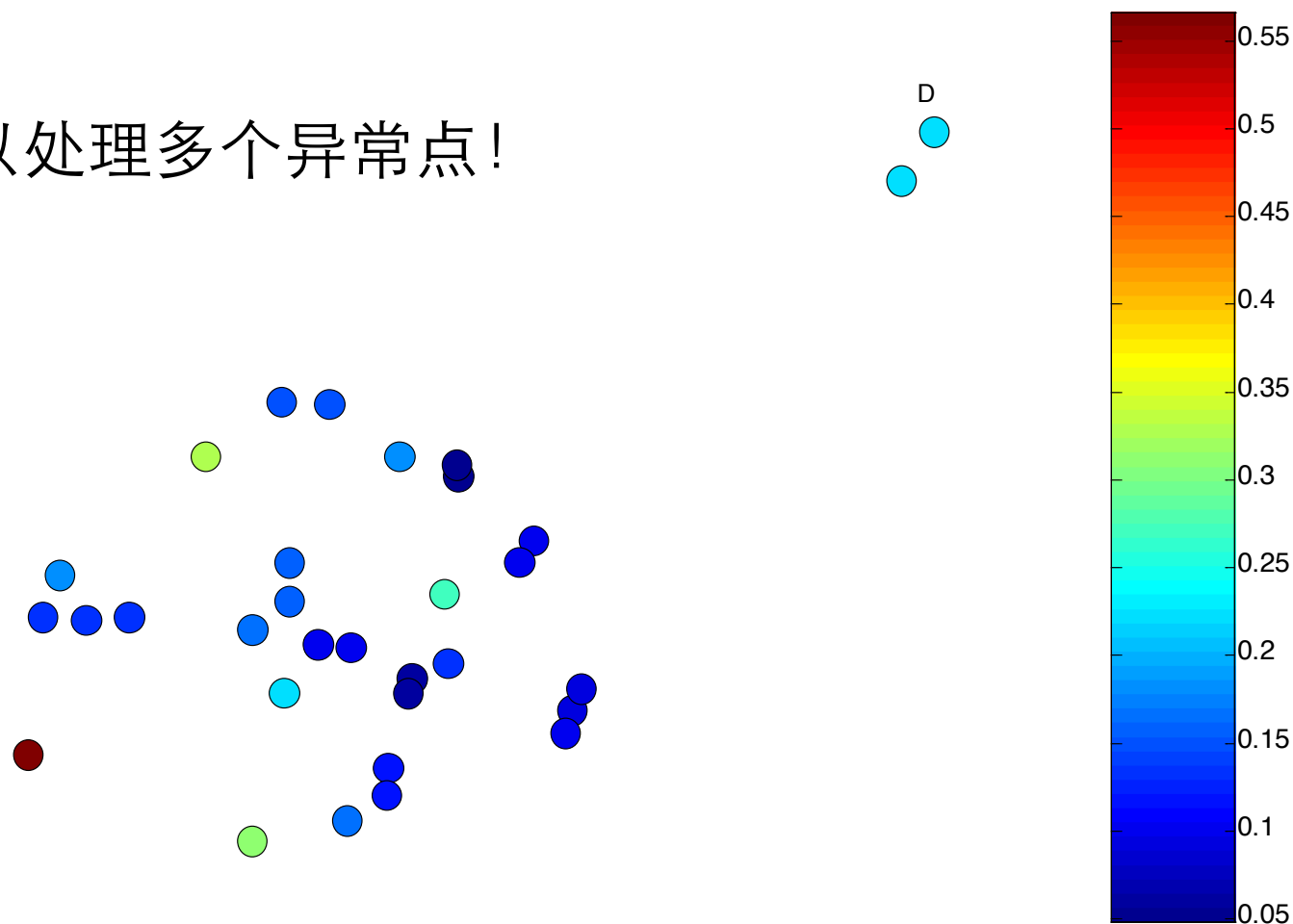
D



- 离群程度定义为其到k近邻的距离

- $k=1$

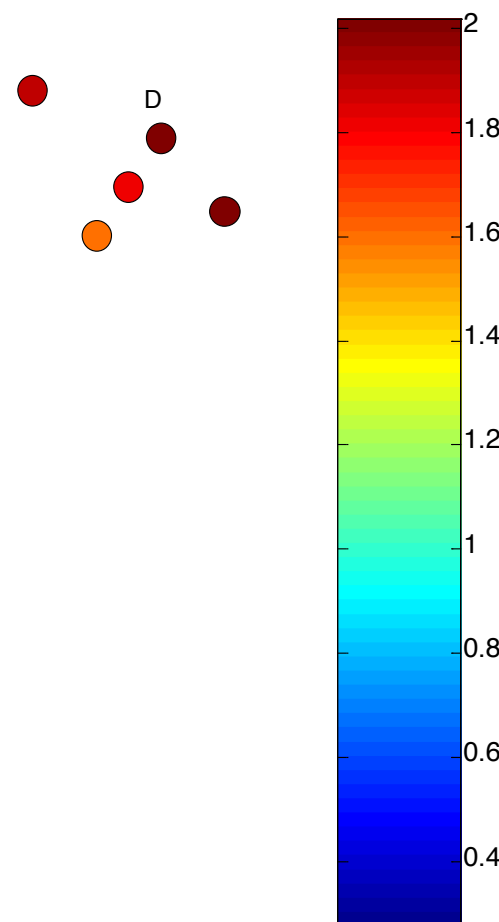
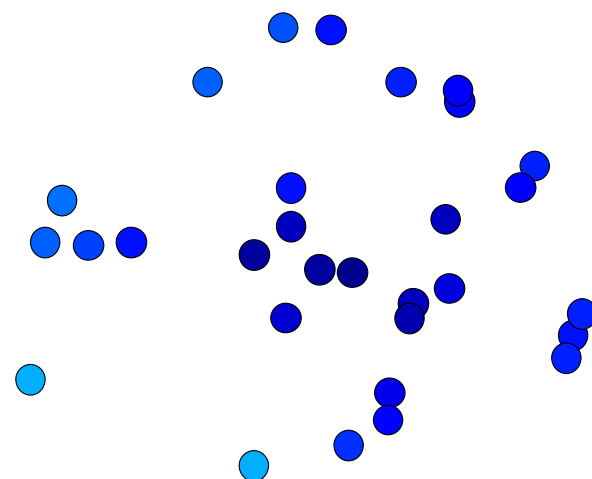
- 难以处理多个异常点!



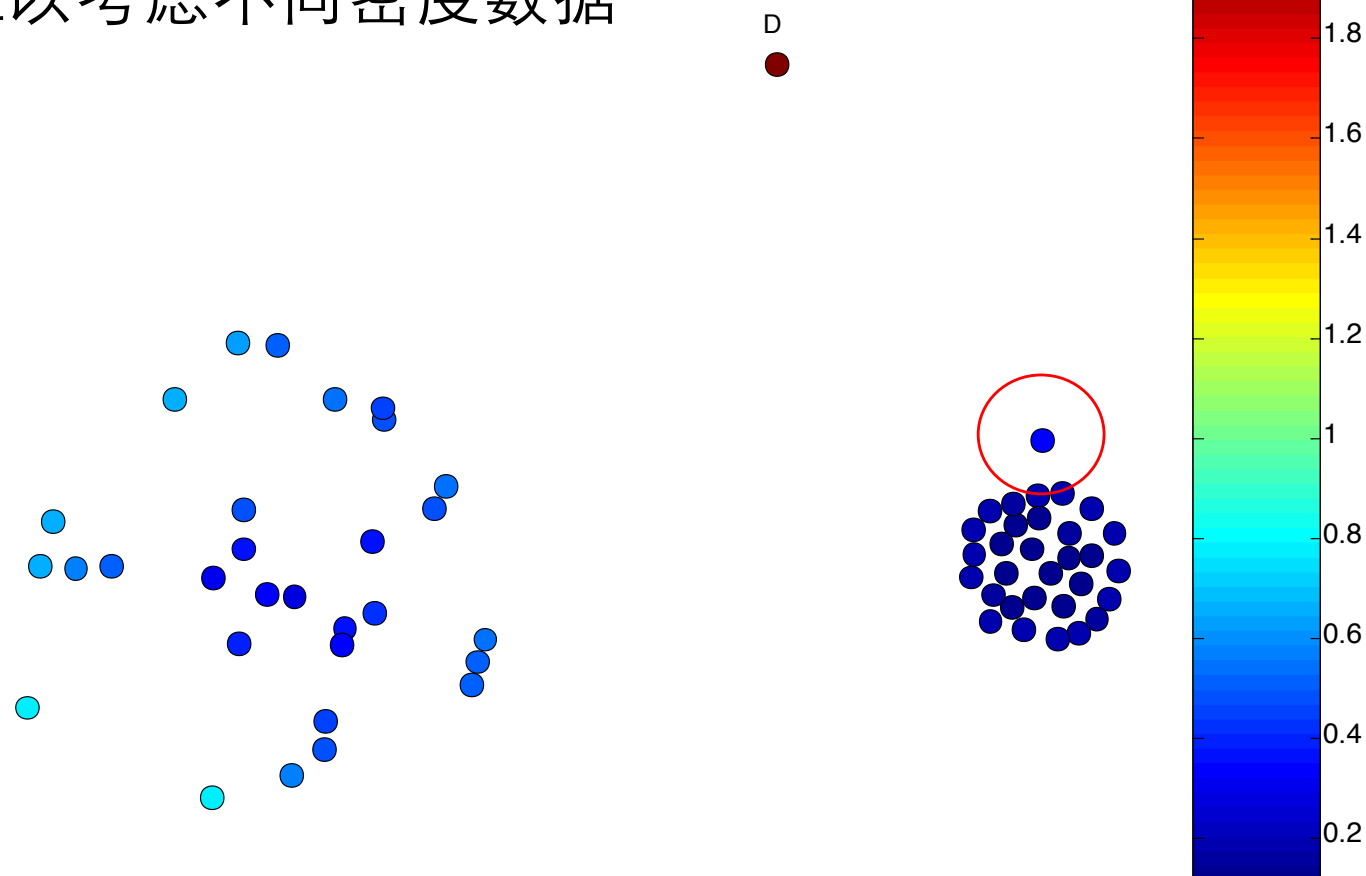
- 离群程度定义为其到k近邻的距离

– k=5

$$\text{density}(\mathbf{x}, k) = \frac{1}{|N(\mathbf{x}, k)|} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} d(\mathbf{x}, \mathbf{y})$$



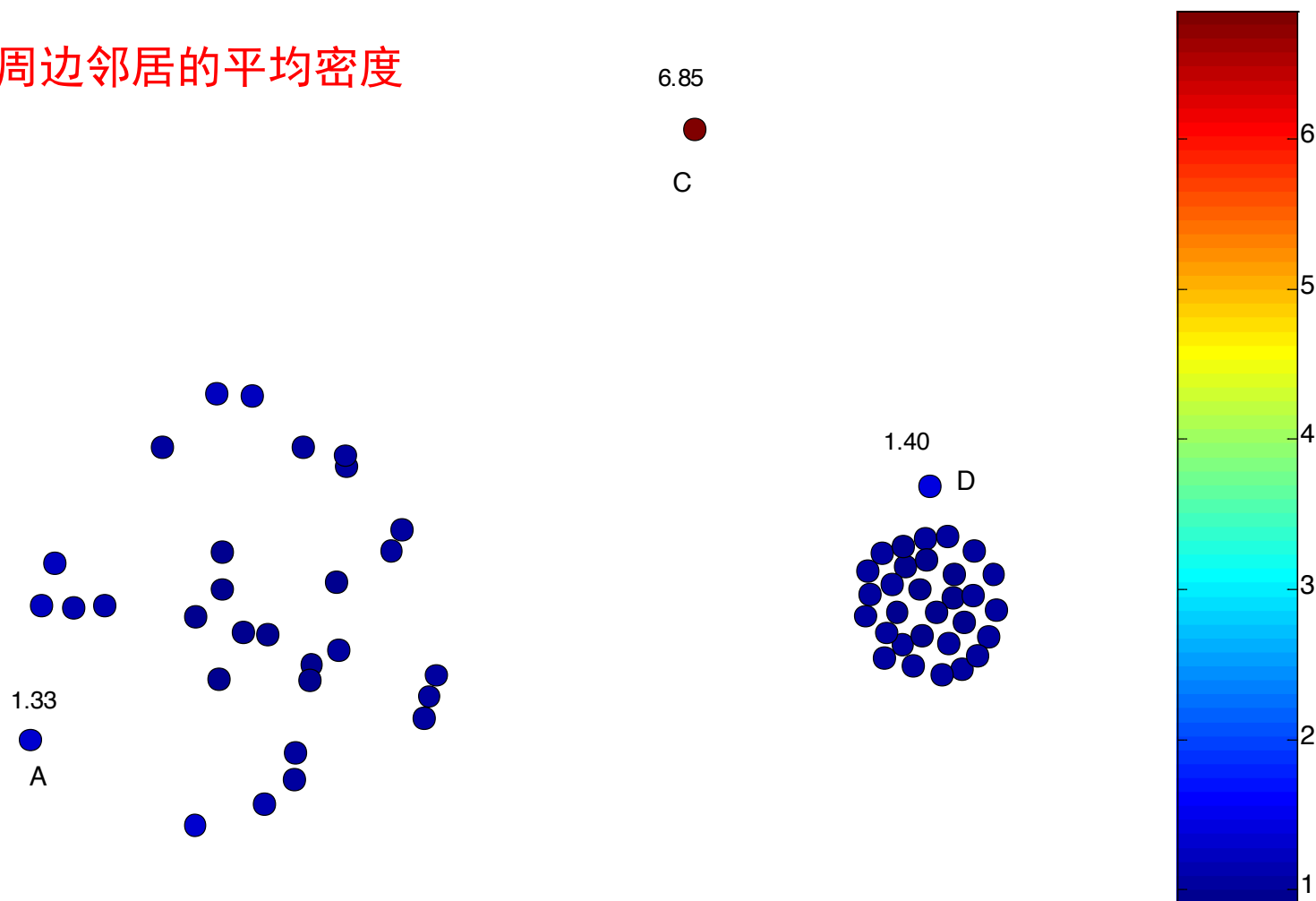
- 离群程度定义为其到k近邻的距离
 - 难以考虑不同密度数据



基于相对密度的离群点检测

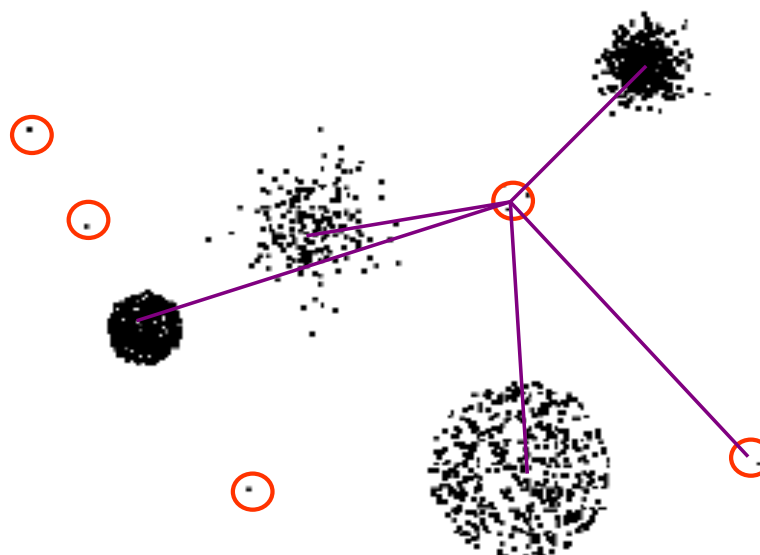
$$\text{average relative density}(\mathbf{x}, k) = \frac{\text{density}(\mathbf{x}, k)}{\sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{density}(\mathbf{y}, k) / |N(\mathbf{x}, k)|}$$

考虑周边邻居的平均密度

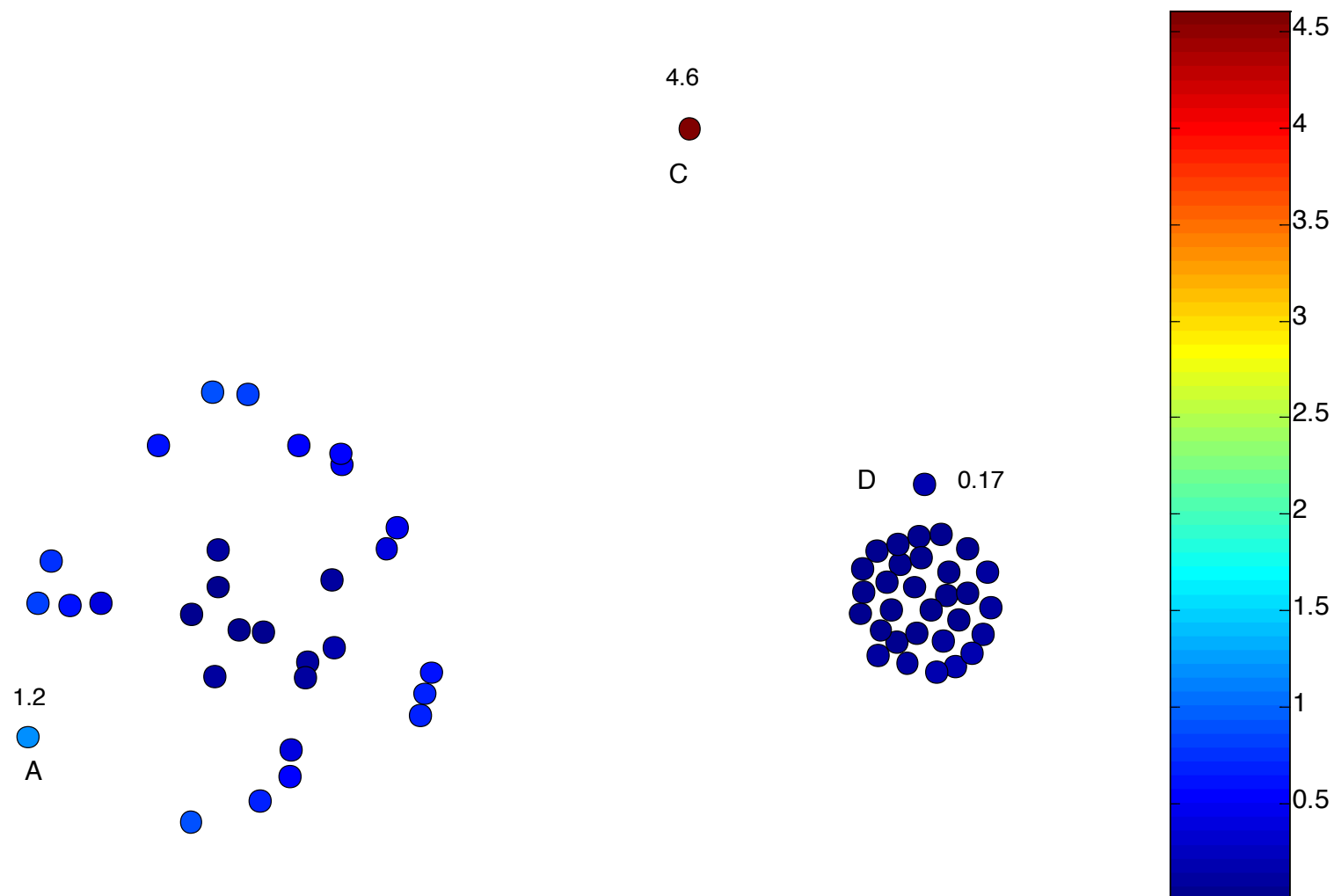


基于聚类的技术

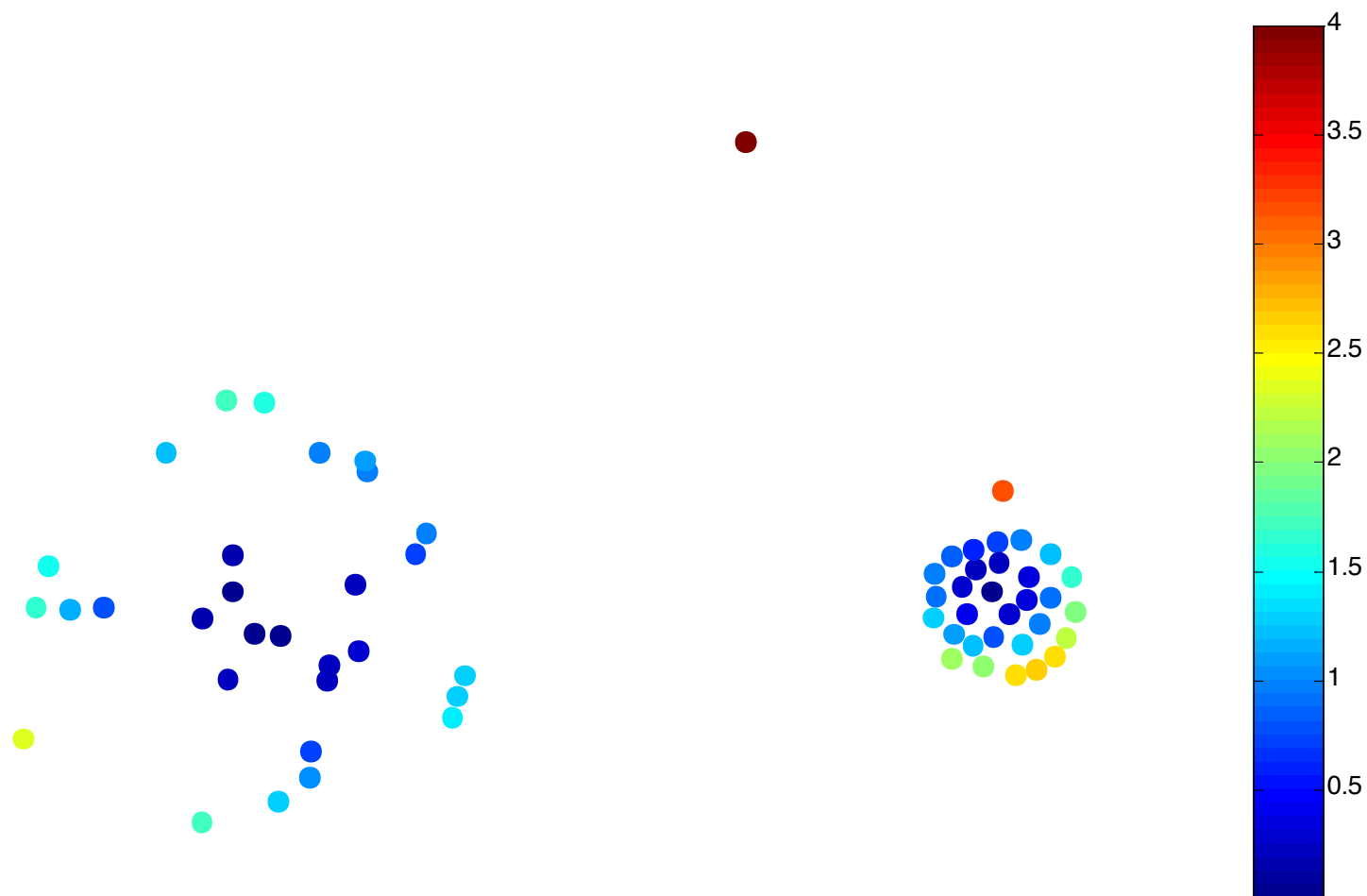
- 综合聚类结果、密度等
 - 远离所有聚类中心的可能是异常点



- 到最近聚类中心点的距离



- 到最近聚类中心点的相对距离



练习三

- 尝试实现k-means聚类算法
- 尝试比较不同的初始点选择、不同的k取值对结果的影响
- 尝试实现一个简单的Apriori算法，比较不同实现的性能差距
- 尝试可视化观察数据中的异常点的分布
(<http://odds.cs.stonybrook.edu/#table1>)
- 是否能够通过简单的距离或者密度计算自动识别出这些异常点？

参考资料

- 机器学习 周志华 清华大学出版社 (Ch2, Ch3)
- Machine Learning Course in stanford
<http://cs229.stanford.edu/>
- <https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-867-machine-learning-fall-2006/lecture-notes/>

参考资料

- 关于关联规则挖掘和异常检测的大部分内容来源于以下两个课程的相关部分：
 - Introduction to Data Mining (Second Edition) <https://www-users.cs.umn.edu/~kumar001/dmbook/index.php>
 - Data Mining: Concepts and Techniques, 3rd ed. https://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm

一些数据资源

- <https://github.com/stedy/Machine-Learning-with-R-datasets>
 - [groceries.csv](#)
- <http://odds.cs.stonybrook.edu/>