南京大学
自然语言处理研究组
Natural Language Processing Group

南京大学
NANJING UNIVERSITY

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

- Supervised learning is the machine learning task of learning a function that maps an input to ~~an output~~ based on example <u>~~input-output pairs~~</u>. (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, $\cancel{y_i}$)

- 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

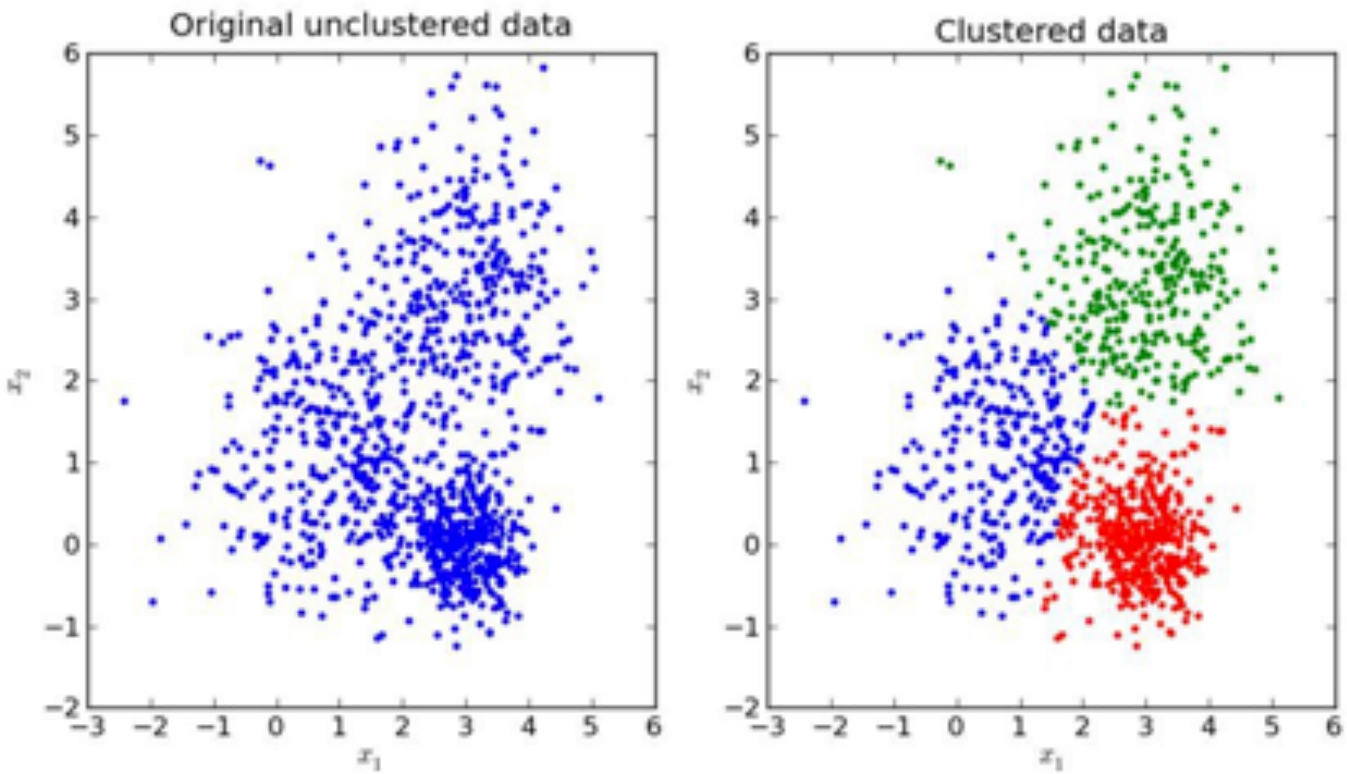
- Unsupervised machine learning algorithms infer patterns from a dataset without reference to known, or labeled, outcomes.

- “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

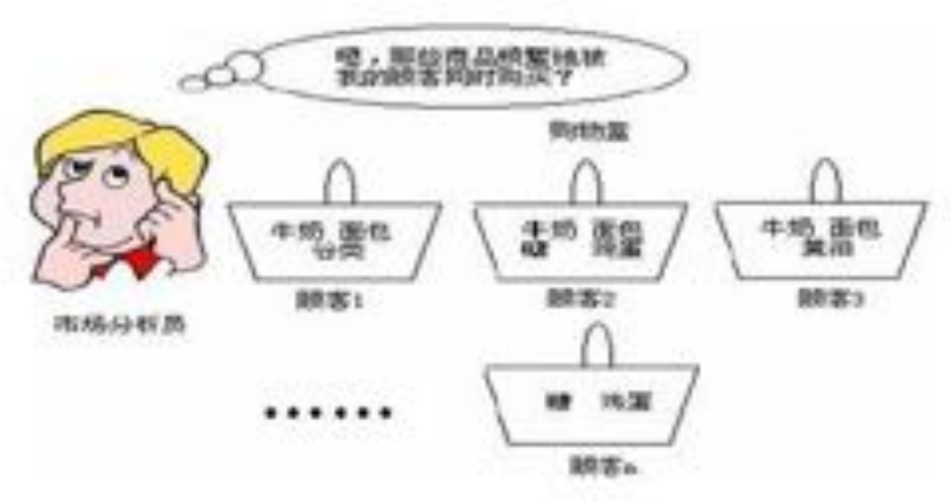
聚类

- 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

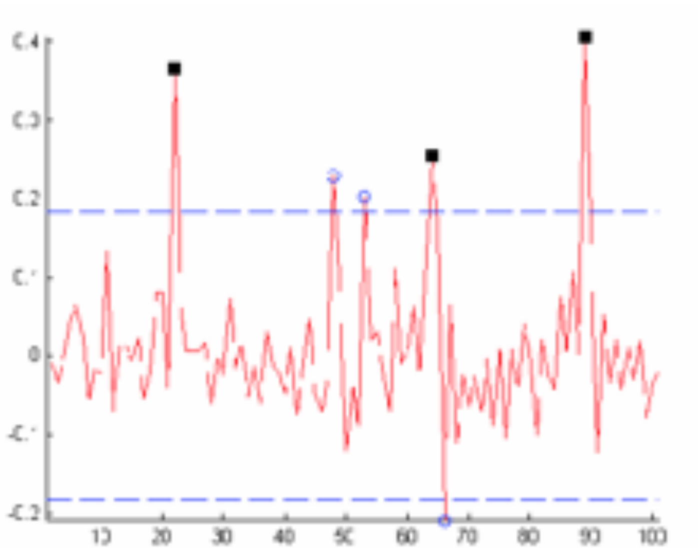


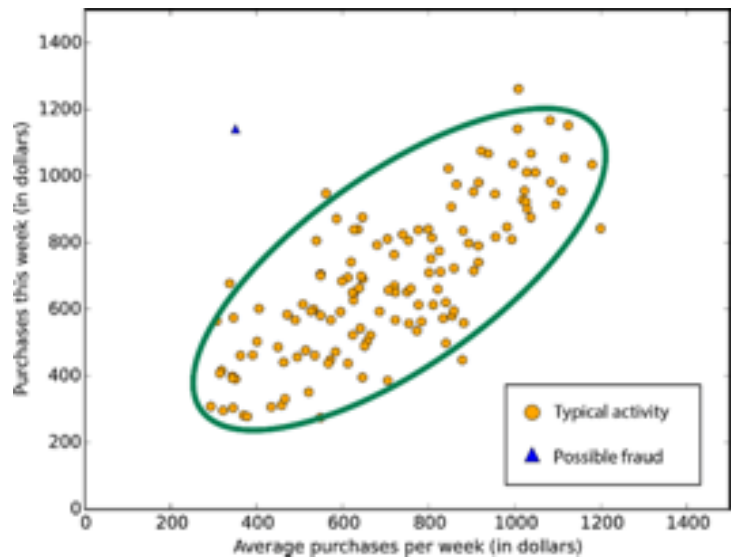
TID	Items
T1	{牛奶,面包}
T2	{面包,尿布,啤酒,鸡蛋}
T3	{牛奶,尿布,啤酒,可乐}
T4	{面包,牛奶,尿布,啤酒}
T5	{面包,牛奶,尿布,可乐}
…	…

→ {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）

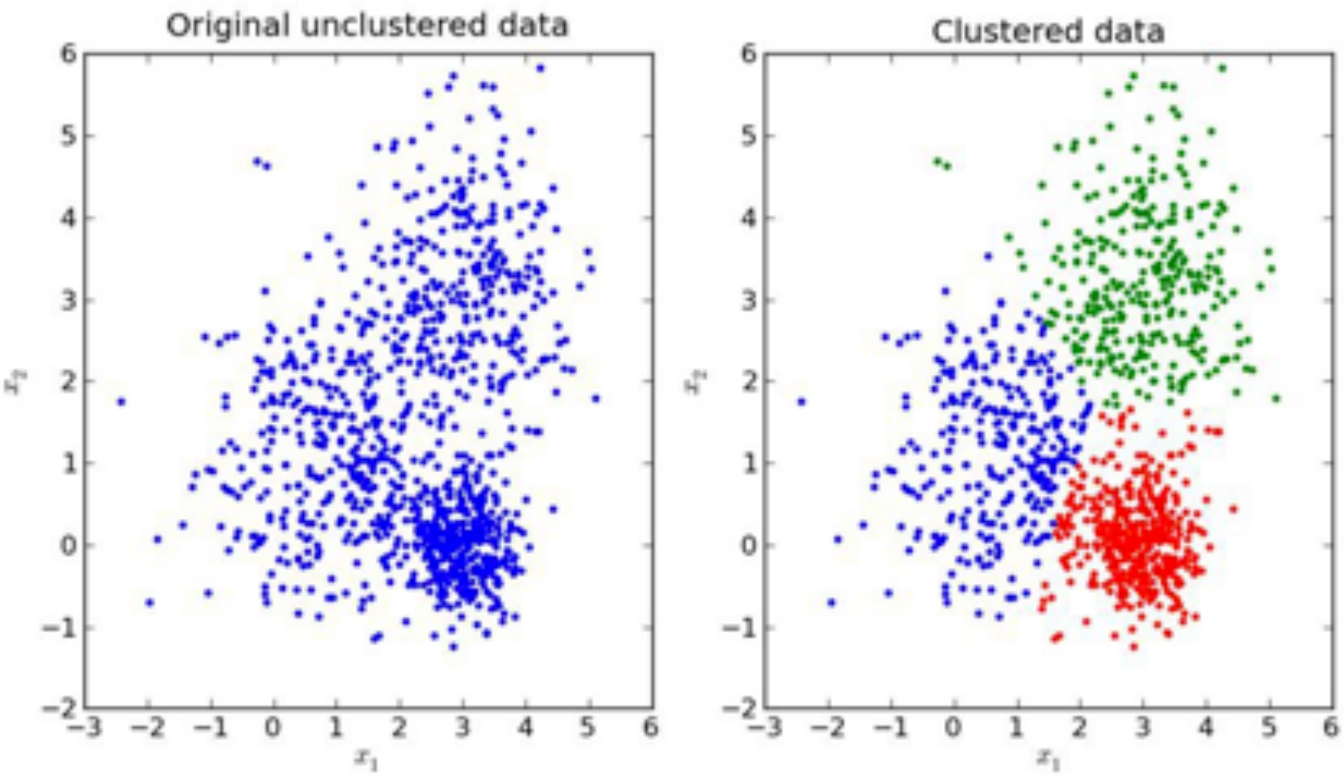




- 无监督学习
- 聚类分析
  - k均值聚类
- 关联规则
- 异常检测

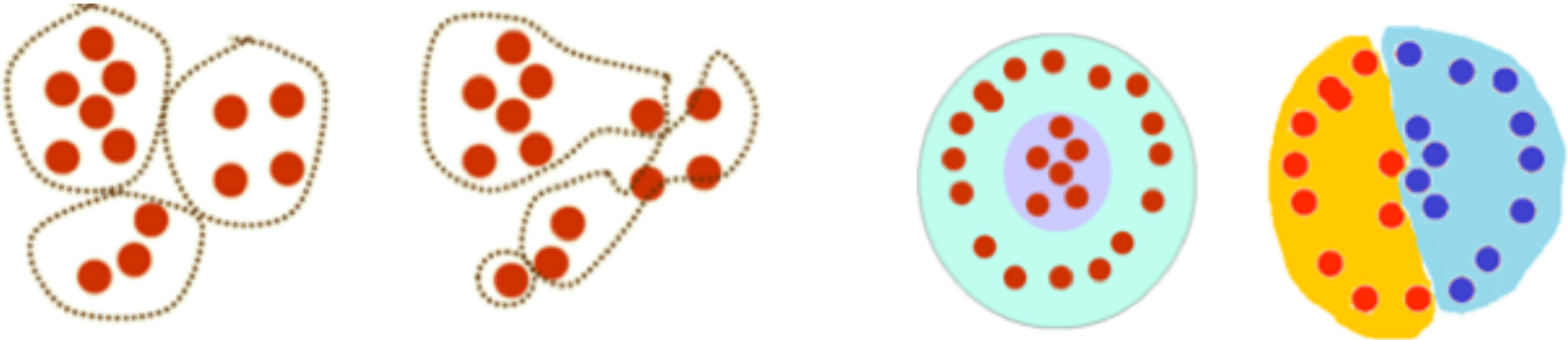
聚类（回顾）

- 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

- 已知一个oracle的聚类结果（外部指标）
  - 比较两个结果是否相同
  - 定义辅助数值如:

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

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

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

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

  - 可以计算指标如Jaccard系数:

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

聚类的评价方法

- 仅考察当前聚类结果（内部指标）
  - 簇内相似度高intra-cluster similarity
  - 簇间相似度低inter-cluster similarity

$$\operatorname{avg}(C)=\frac{2}{|C|(|C|-1)}\sum_{1\leqslant i<j\leqslant|C|}\operatorname{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

$$\operatorname{diam}(C)=\max_{1\leqslant i<j\leqslant|C|}\operatorname{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

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

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

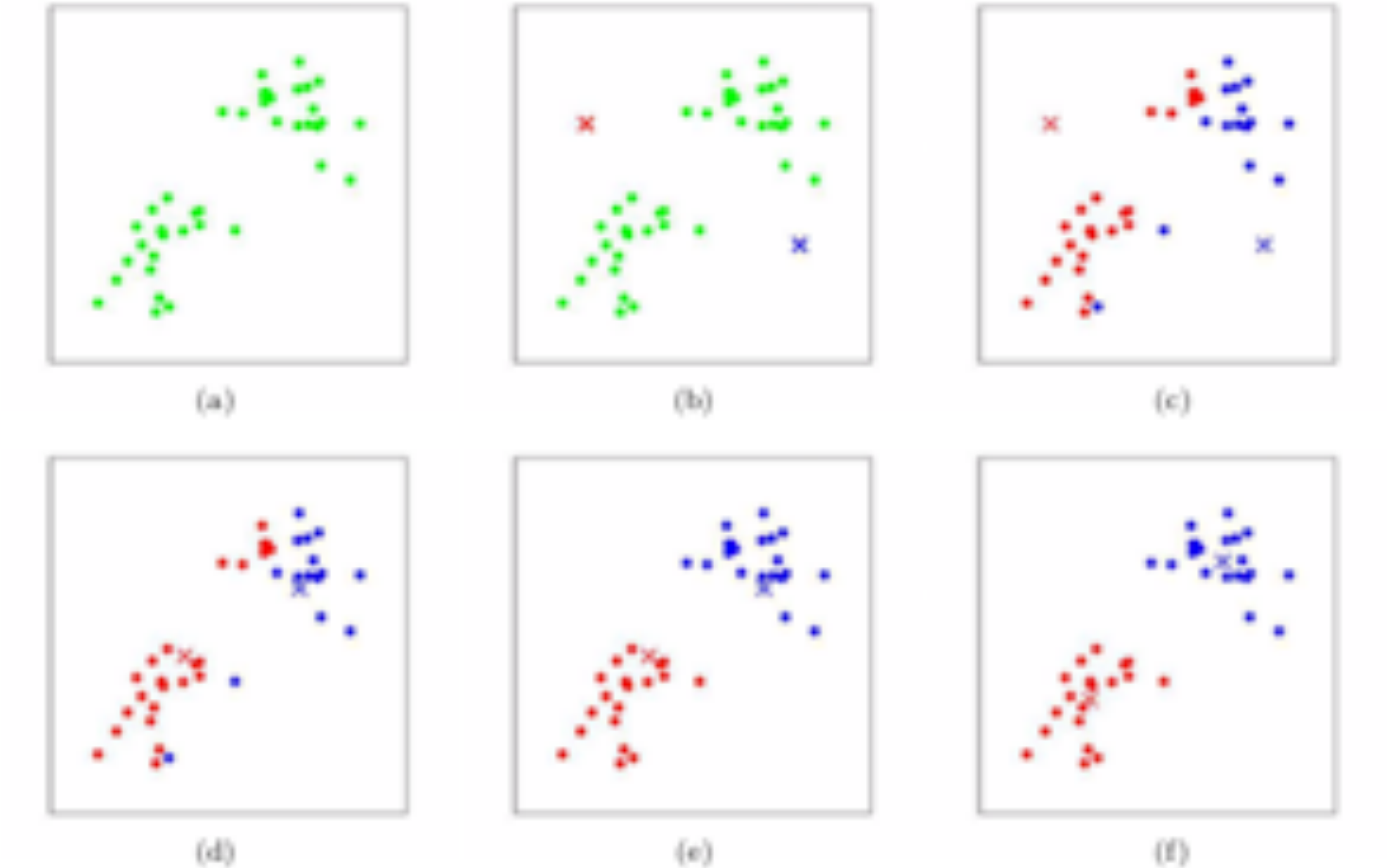
k-Means聚类

- 一种针对聚类中心进行优化的方法：

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

- 1. 初始化k个聚类中心$\boldsymbol{\mu}_i$
- 2. 将每个x划入到距离最近的聚类中心$\boldsymbol{\mu}_i$
- 3. 重新计算每个类的中心$\boldsymbol{\mu}_i$
- 4. 转至2继续执行，直至聚类不发生变化

K means的中心变化

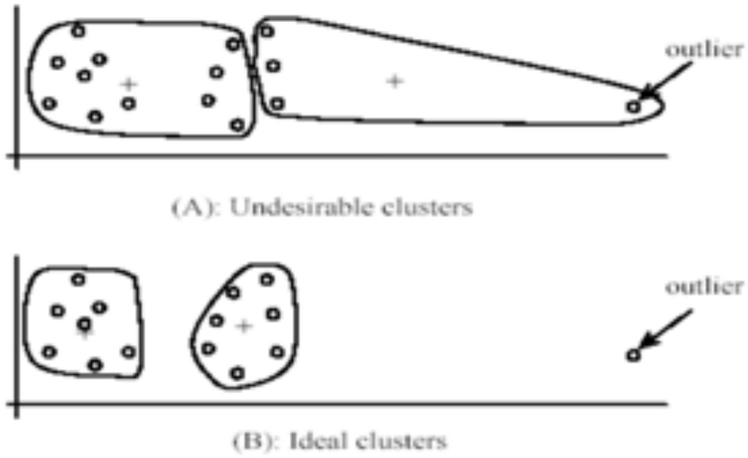


http://cs229.stanford.edu/notes-spring2019/cs229-notes7a.pdf

一些简单的讨论

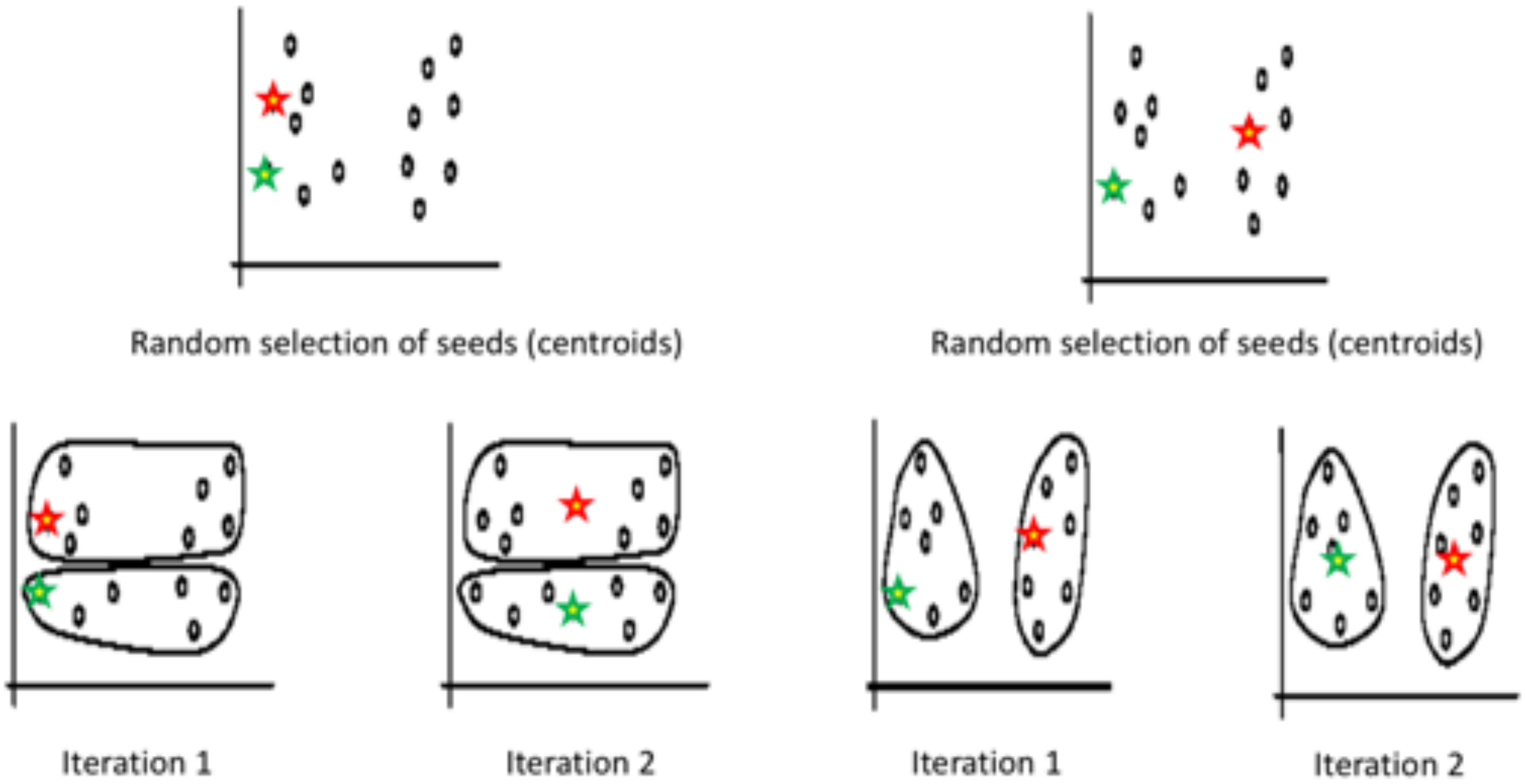
- 聚类数目

- 异常点



一些简单的讨论

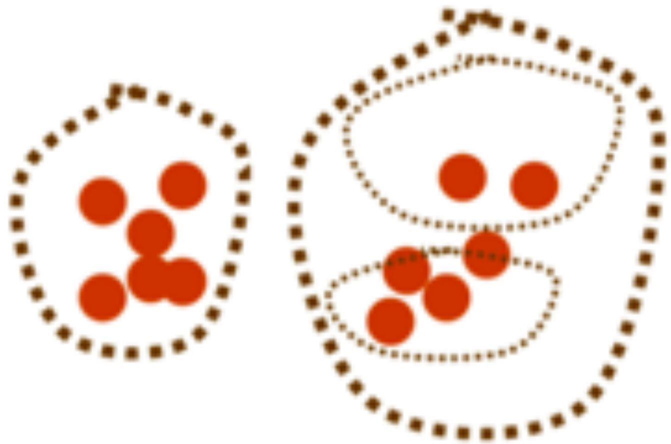
- 初始点



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

练习四

- 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

参考资料

- 机器学习 周志华 清华大学出版社（Ch2，Ch3）
- Machine Learining 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/