第8.1节聚类分析

中国科学院大学 叶齐祥

qxye@ucas.ac.cn

提 纲

- 概述
- K-means聚类算法
- 聚类算法的距离度量
- 层次聚类算法
- 从聚类到Unsupervised Learning

概述

- 物以类聚、人以群分。
  - 但什么是分类的根据呢？
  - 比如，要想把中国城市分成若干类，有很多种分类法；
    - 可以按照自然条件来分，
    - 比如考虑降水、土地、日照、湿度等各方面；
    - 也可以考虑收入、教育水准、医疗条件、基础设施等指标；
    - 既可以用某一项来分类，也可以同时考虑多项指标来分类。

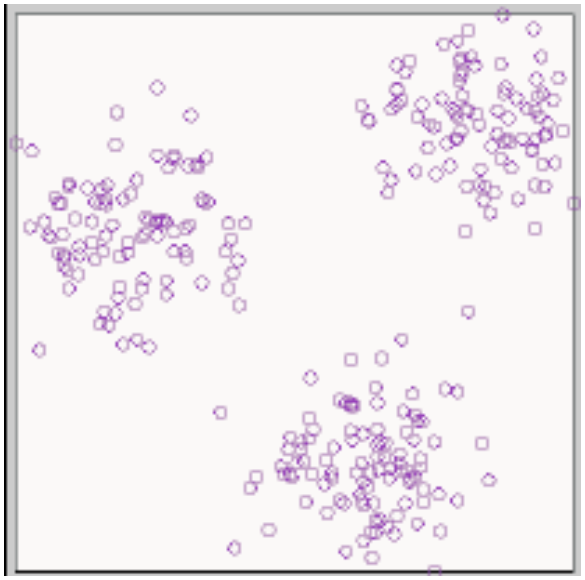
概述

- 聚类分析*：
  - 就是按照一定的规律和要求对事物进行区分和分类的过程，在这一过程中没有任何关于类分的先验知识，没有指导，仅靠事物间的相似性作为类属划分的准则。
  - 一个数据集合分组成几个簇 （Clusters）
  - 聚类分析是一种无监督分类:**没有预定义的类**

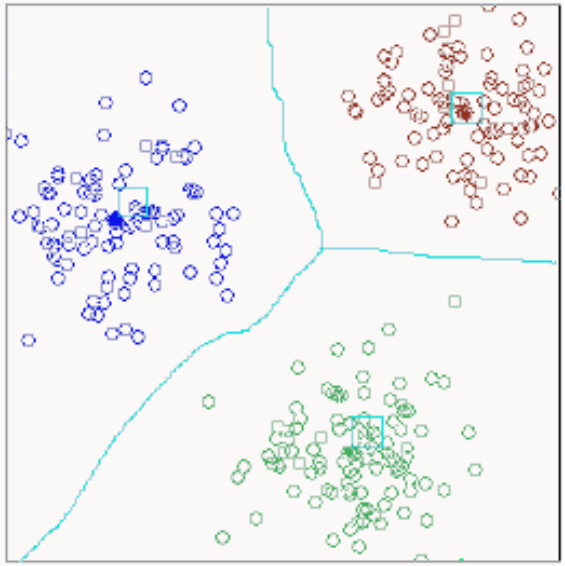
概述

- 聚类是无监督的学习过程，同分类区别在于*
  - 分类是需要事先知道所依据的数据特性，而聚类是要找到这个数据特性
  - 在很多应用中，聚类分析作为一种数据预处理过程，是进一步分析和处理数据的基础。
  - 例如
    - 在电子商务中，帮助市场分析人员从客户基本库中发现不同的客户群，用不同模式来刻画不同的客户群的特征。
    - 在生物学中，聚类分析能用于推导植物和动物的分类，对基因进行分析，获得对种群中固有结构的认识。
    - 聚类分析也能用于分类Web文档来获得信息。

概述



二维数据



空间划分

空间覆盖

提 纲

- 概述
- K-means聚类算法
- 聚类算法的距离度量
- 层次聚类算法
- 从聚类到Unsupervised Learning

K-means聚类算法

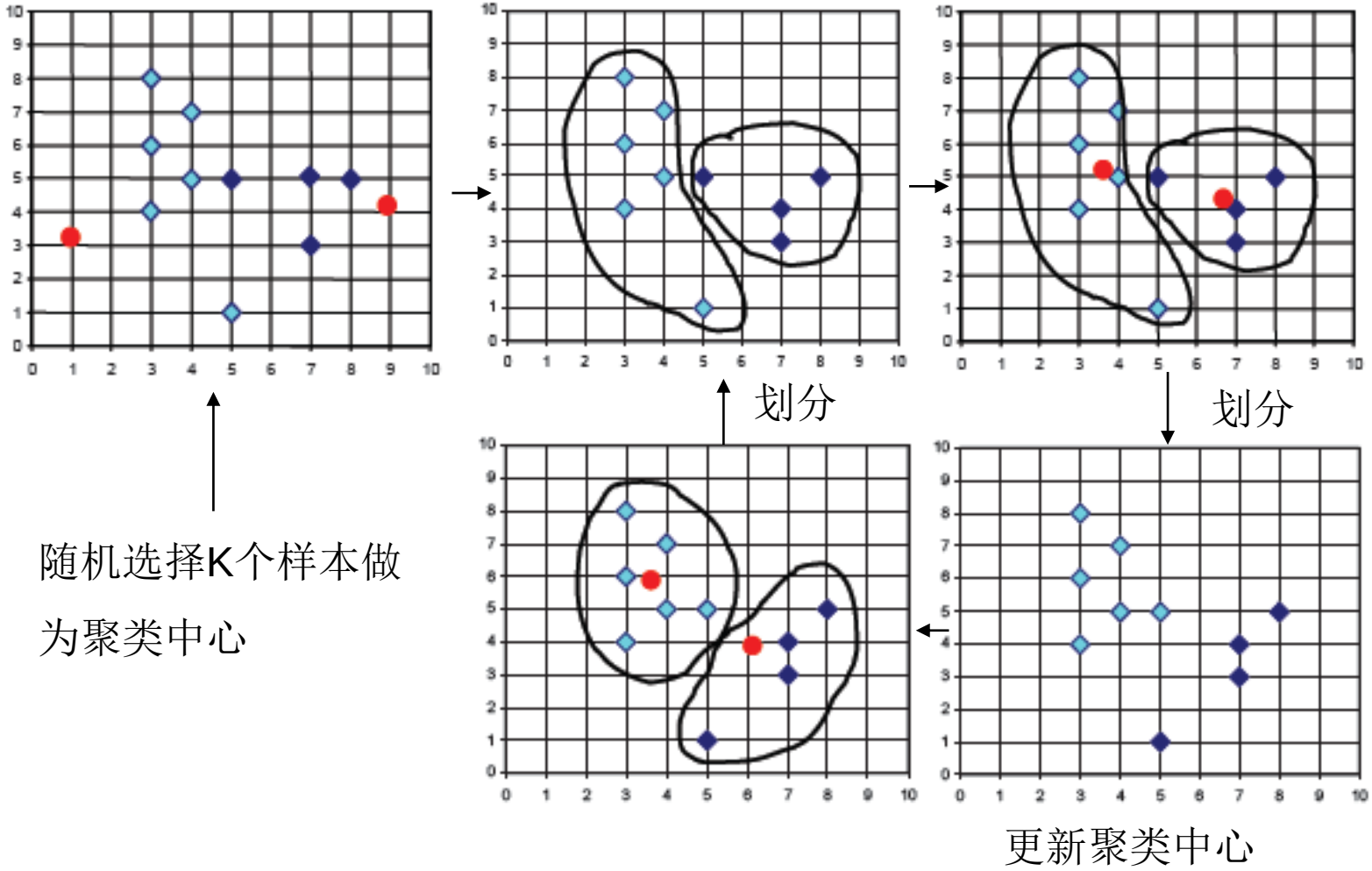
- 定义
  - K-means算法首先随机选择k个对象，每个对象代表一个聚类的质心。
  - 对于其余的每一个对象，根据该对象与各聚类质心之间的距离，把它分配到与之最相似的聚类中。
  - 然后，计算每个聚类的新质心。
  - 重复上述过程，直到准则函数(如标号变化率)收敛。

K-means聚类算法

- 算法描述
  - 1) 从数据集中选择k个质心C^1，C^2，…，C^k作初始聚类中心；
  - 2) 把每个样本划分到与之**最相近**的聚合。每个聚合用其中所有对象的均值来代表，“最相近”就是指距离最小。对于每个点V^i，找出一个质心C^j，使它们之间的距离$d(V^j, C^j)$最小，并把V^i分配到第j个簇（Cluster）；
  - 3) 把所有的样本划分之后，重新计算每个簇的质心；
  - 4) 循环执行第2)步和第3)步，直到数据的划分不再发生变化。

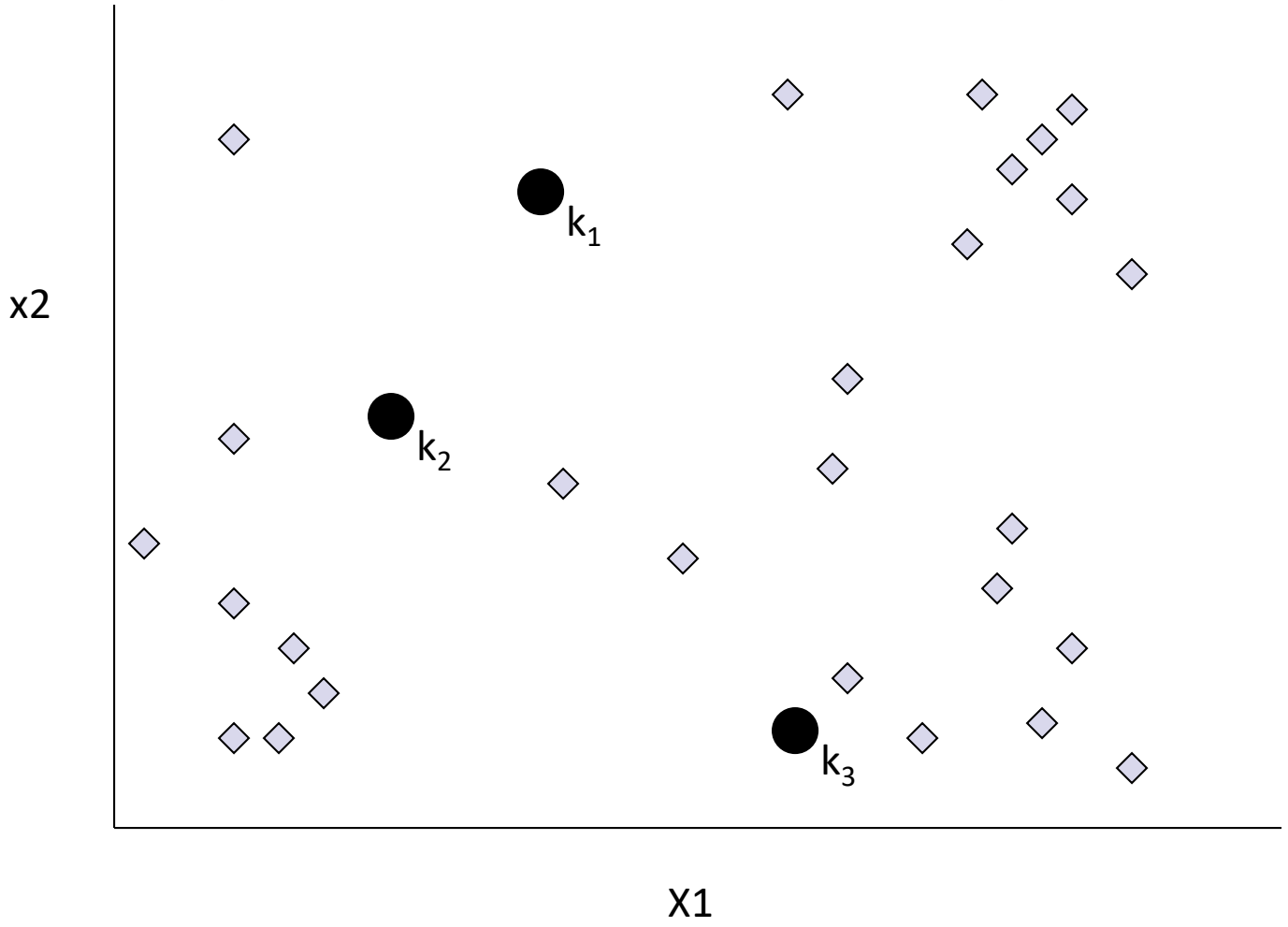
K-means聚类算法

- 算法示意



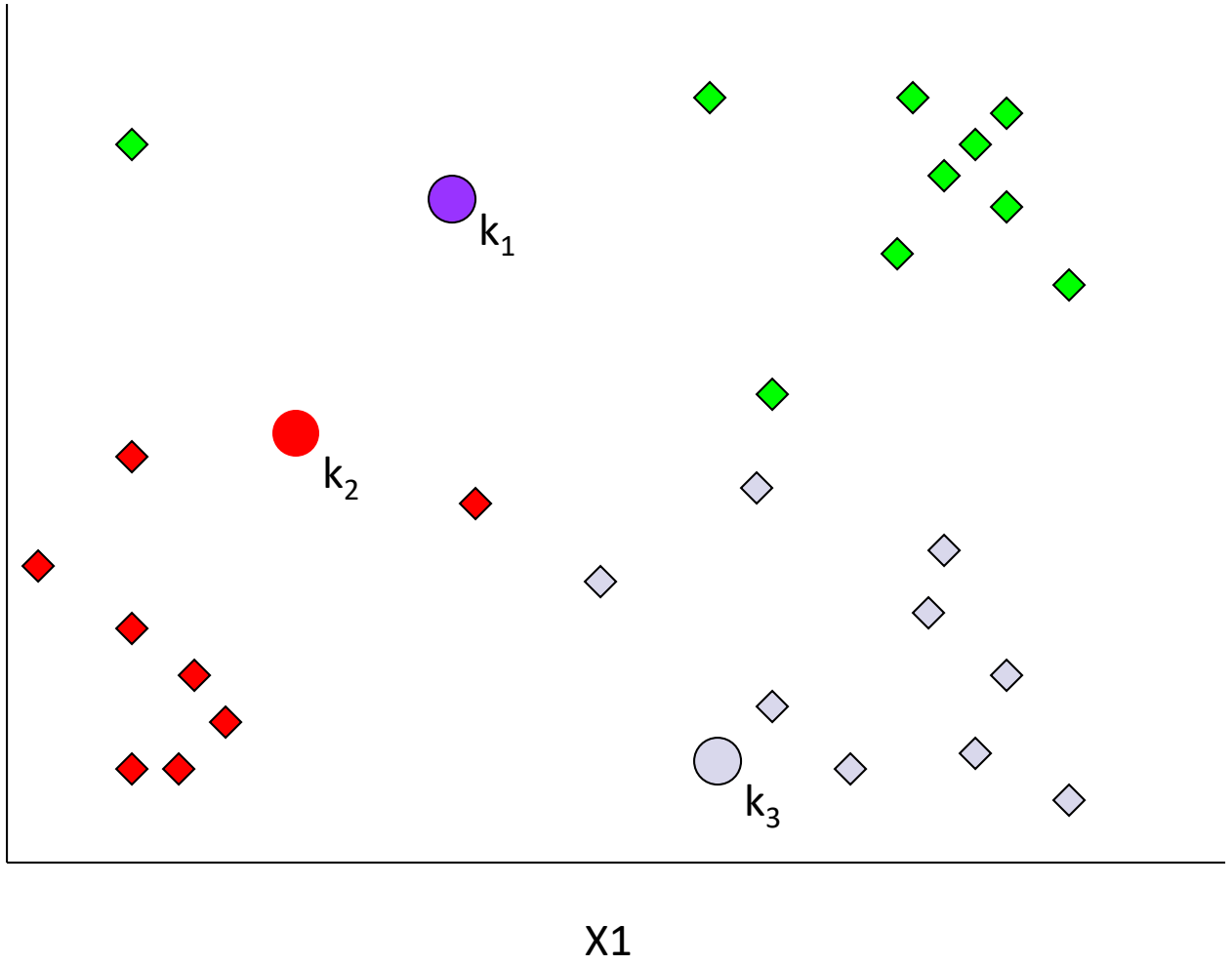
K-means聚类算法

（随机地）选择初始聚类中心

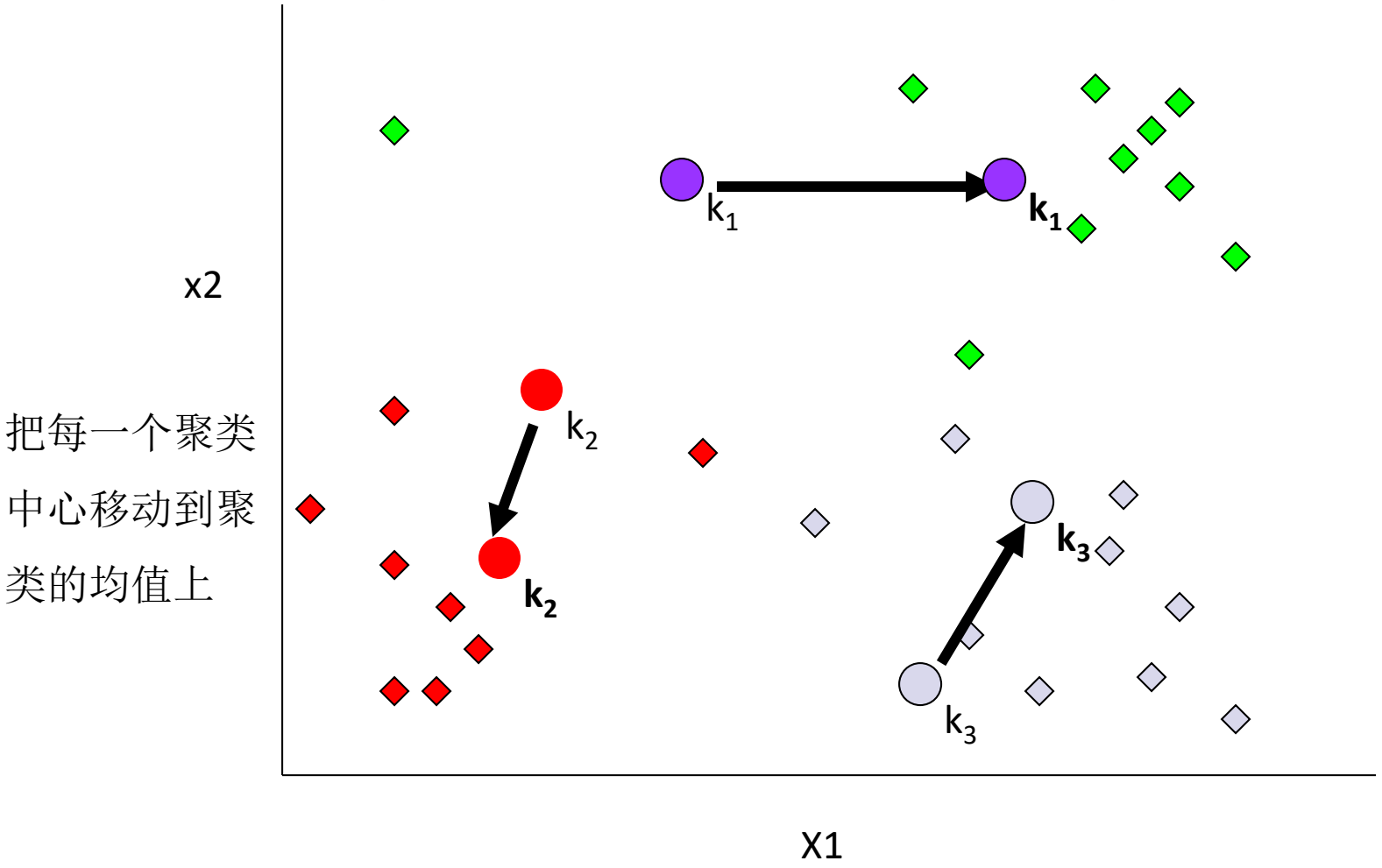


K-means聚类算法

每一个点被划分到离某中心最近的那个聚类



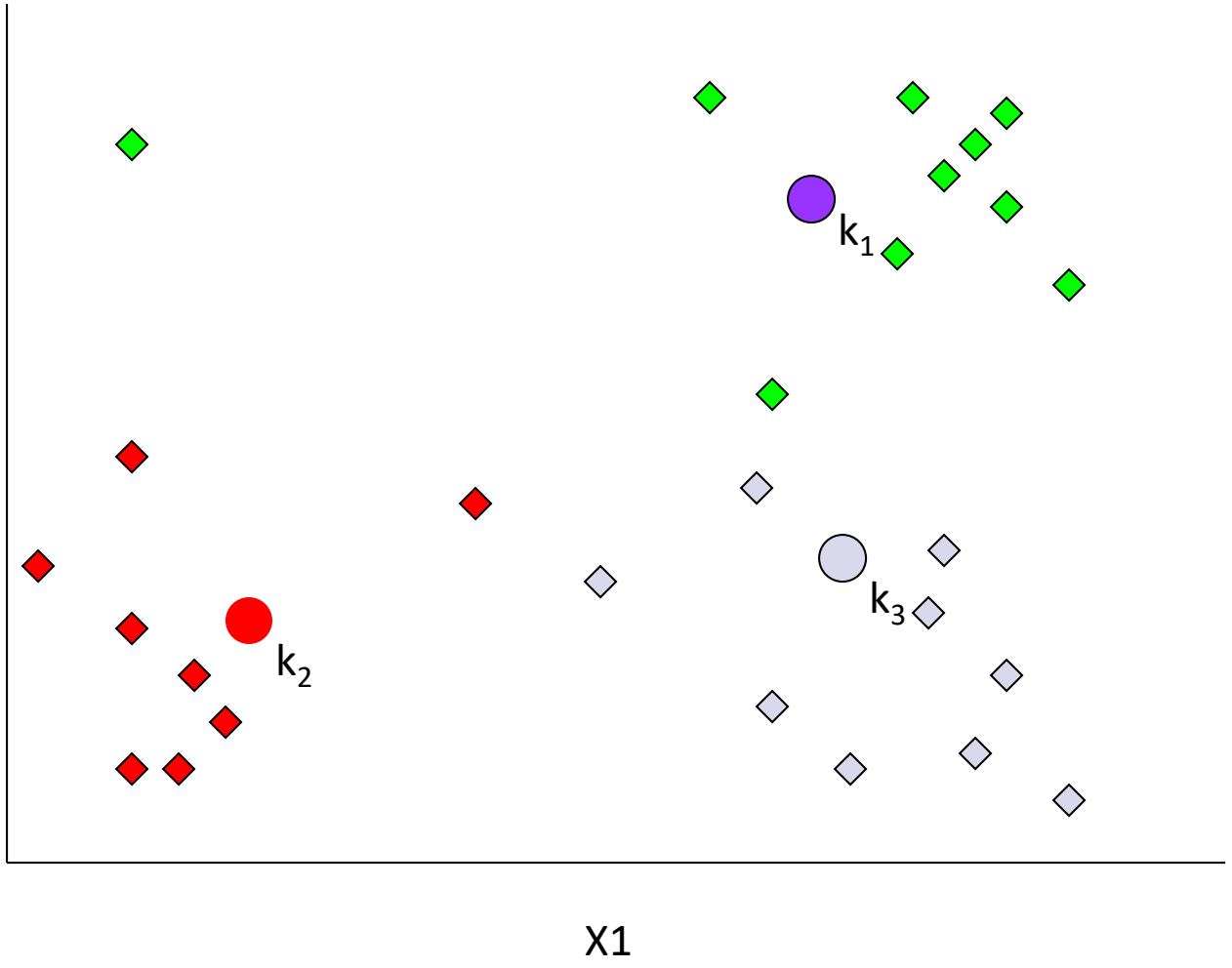
K-means聚类算法



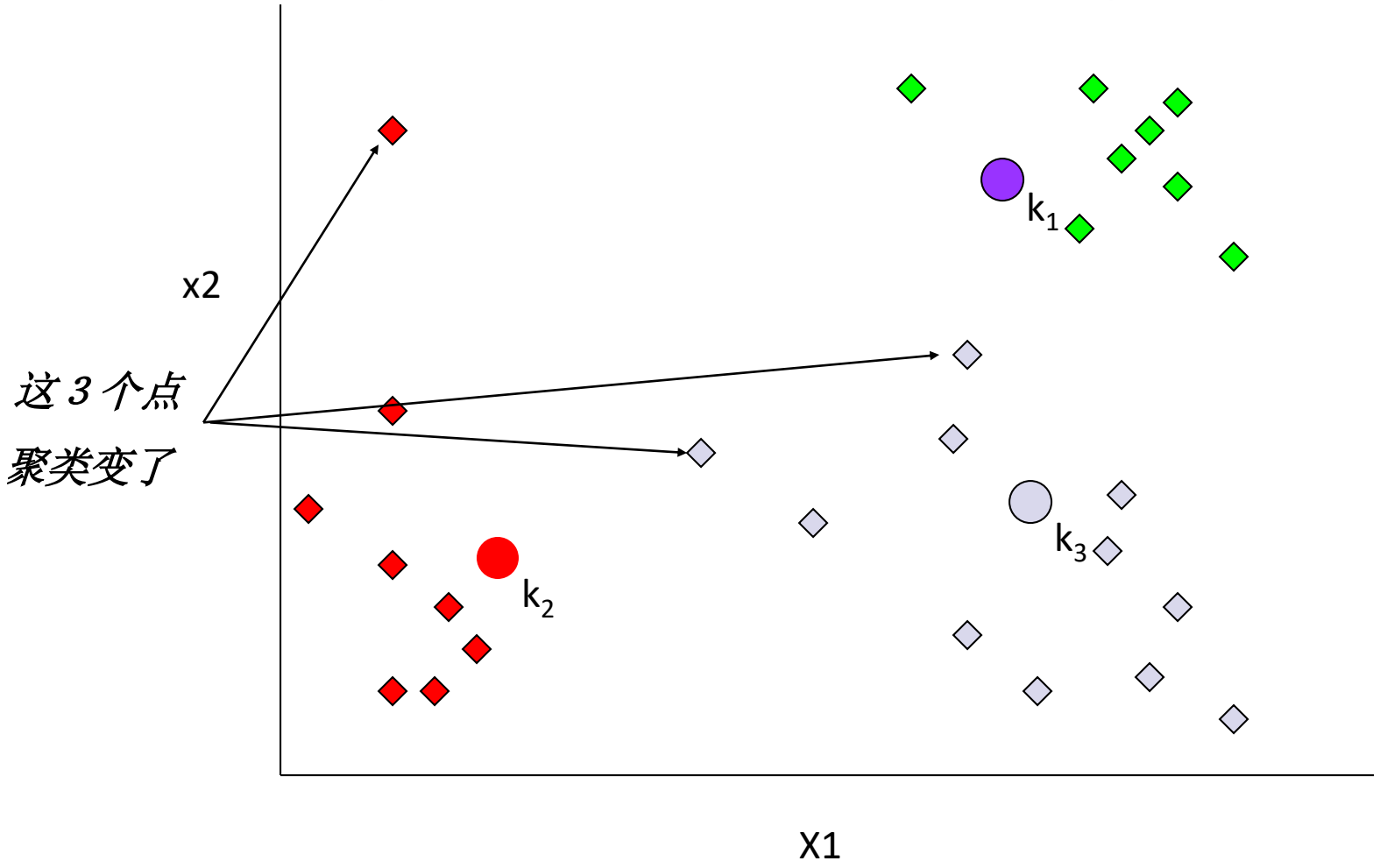
K-means聚类算法

重新把这些点安排到中心最近的聚类

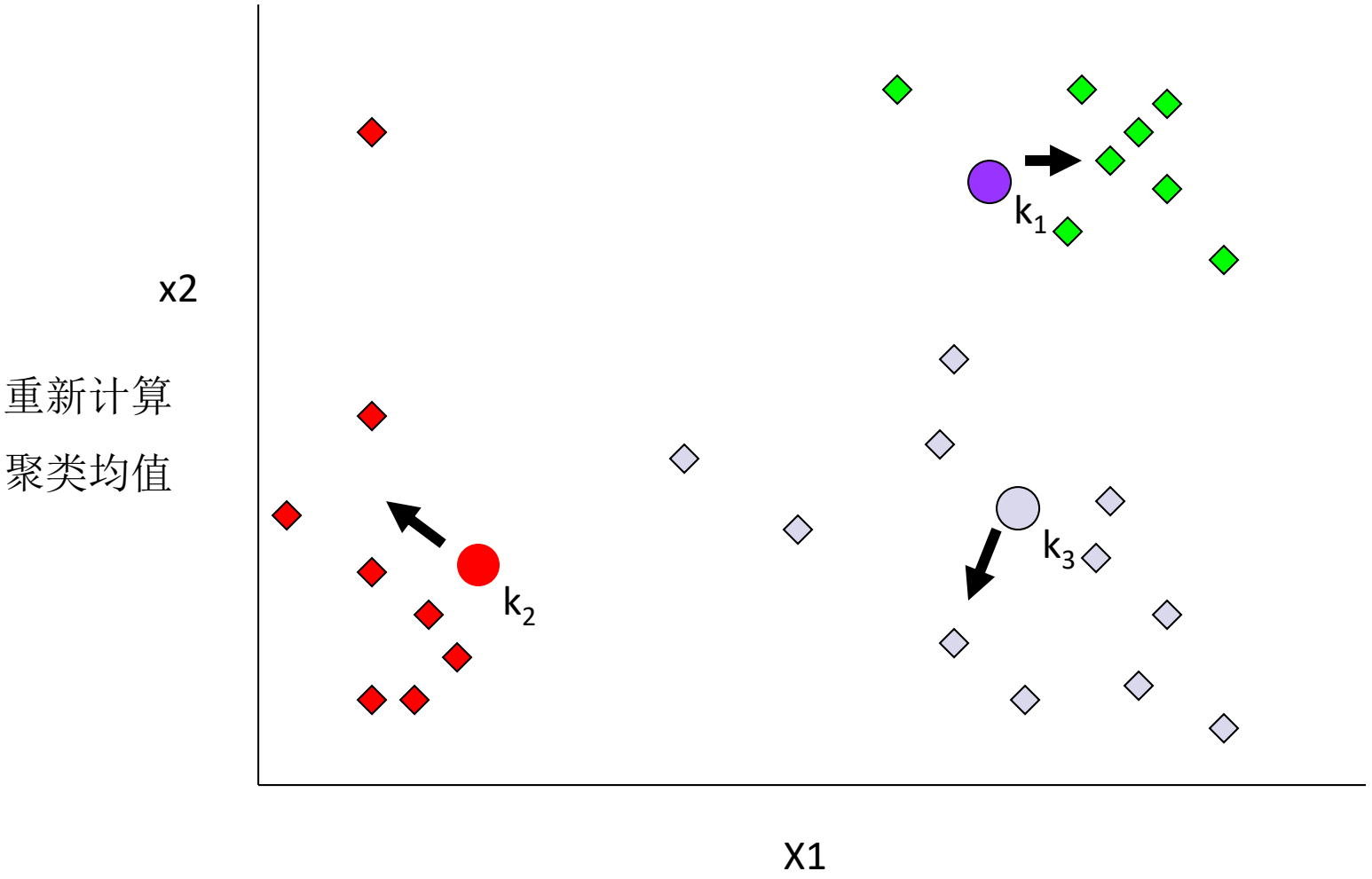
哪些点的聚类变了?



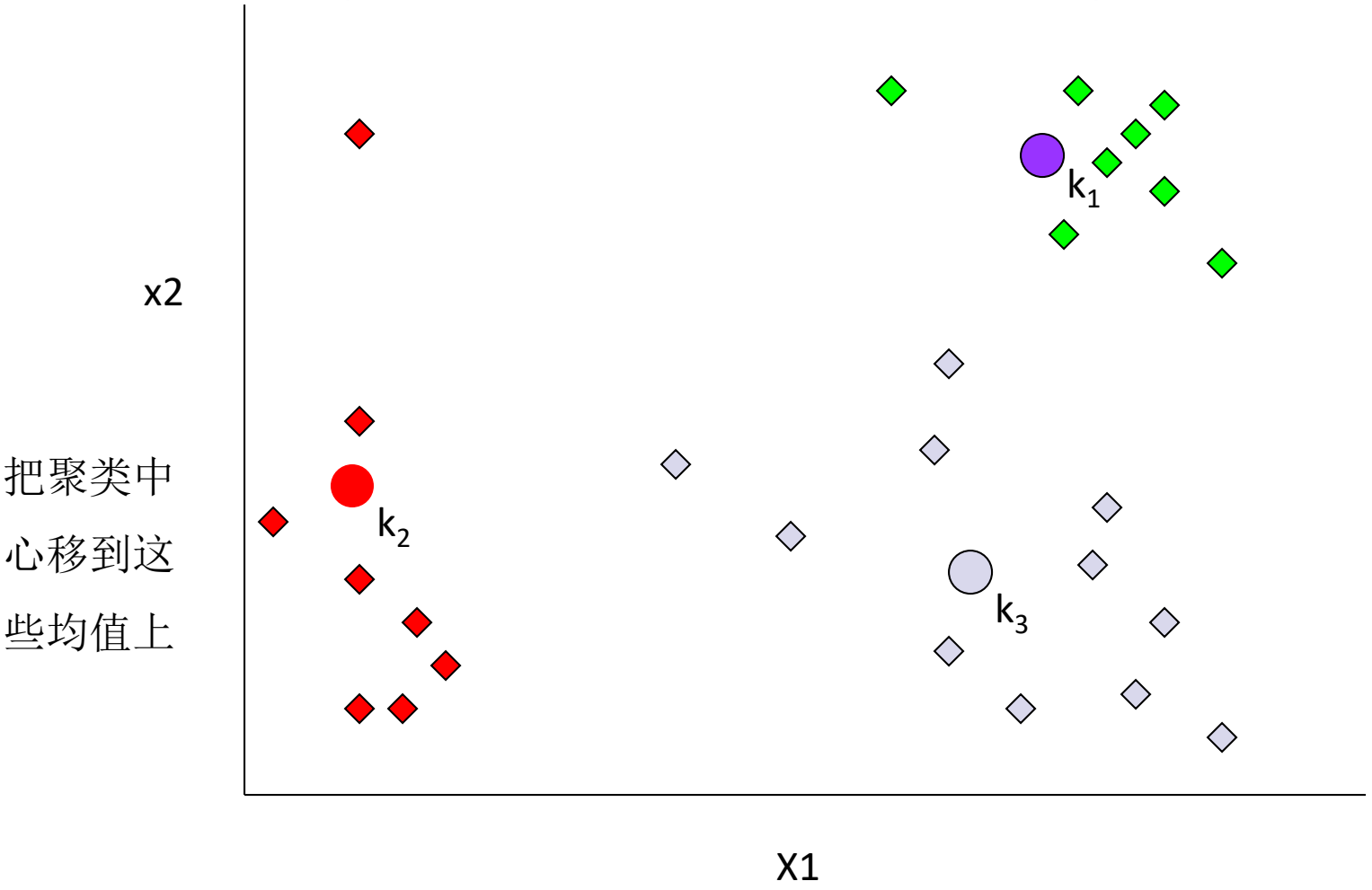
K-means聚类算法



K-means聚类算法

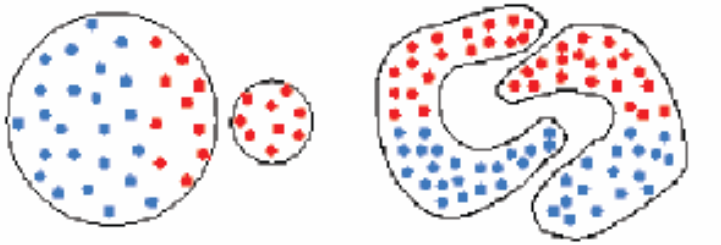


K-means聚类算法



K-means聚类算法

- 优缺点
  - 其计算复杂度为O(*nkt*)，其中，*n*是样本个数，*k*是类别数，*t*是循环迭代的次数。
  - K-means聚类算法的不足之处在于它要多次扫描样本集合
  - 它只能找出球形的类，而不能发现任意形状的类。



  - 初始质心对聚类结果有较大的影响
  - 该算法对噪声很敏感。

K-means聚类算法的改进K-medoids

- 定义
  - K-medoids算法的过程和上述k-means的算法过程相似，不同之处是：
  - k-medoids算法用类中最靠近中心的一个对象来代表该聚类，而k-means算法用质心来代表聚类。
  - 在k-means算法中，对噪声非常敏感，因为一个极大的值会对质心的计算带来很大的影响。而k-medoid算法中，通过用中心来代替质心，可以有效地消除该影响。

提 纲

- 概述
- K-means聚类算法
- 聚类算法的距离度量
- 层次聚类算法
- 从聚类到Unsupervised Learning

K-means聚类算法的距离度量

1. 欧氏距离: $\|x-y\|_2=\left(\sum_i (x_i-y_i)^2\right)^{1/2}$

2. 街区距离: $\|x-y\|_1=\sum_i |x_i-y_i|$

3. 向量角距离: $\cos\theta_{xy}=\frac{x\cdot y}{\|x\|\|y\|}$

4. 相关系数: $r_{xy}=\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2 \sum_i (y_i-\bar{y})^2}}$

提 纲

- 概述
- K-means聚类算法
- 聚类算法的距离度量
- 层次聚类算法
- 从聚类到Unsupervised Learning

典型算法—层次方法

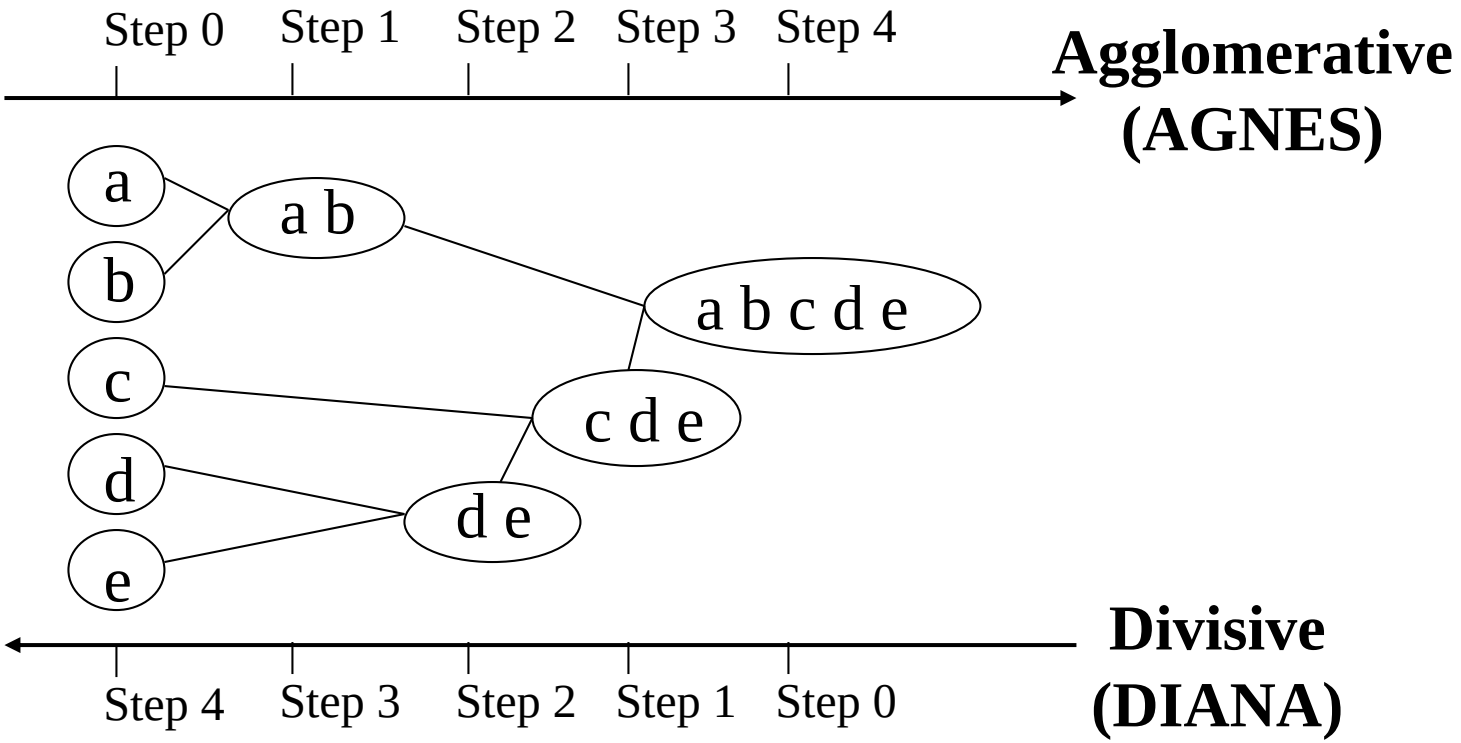
- 定义
  - 层次方法对给定数据对象集合进行层次的分解。根据顺序，层次的方法可以分为凝聚的和分裂的。
  - 凝聚的方法，为自底向上的方法，一开始将每个对象作为单独的一个组，然后相继地合并相近的对象或组，直到所有的组合并为一个，或者达到一个终止条件。
  - 分裂的方法，为自顶向下的方法，一开始将所有的对象置于一个簇中，在迭代的每一步中，一个簇被分裂为更小的簇，直到最终每个对象在单独的一个簇中，或者达到一个终止条件。

典型算法—层次方法

- 层次的聚类方法将数据对象组成一棵聚类的树
- 根据层次分解是自底向上, 还是自顶向下形成, 层次的聚类方法可以进一步分为凝聚的(agglomerative)和分裂的(divisive)层次聚类
- 纯粹的层次聚类方法的聚类质量受限于如下特点：一旦一个合并或分裂被执行，就不能修正
- 最近的研究集中于凝聚层次聚类和迭代重定位方法的集成
- 使用距离矩阵作为聚类标准. 该方法不需要输入聚类数目 k, 但需要终止条件

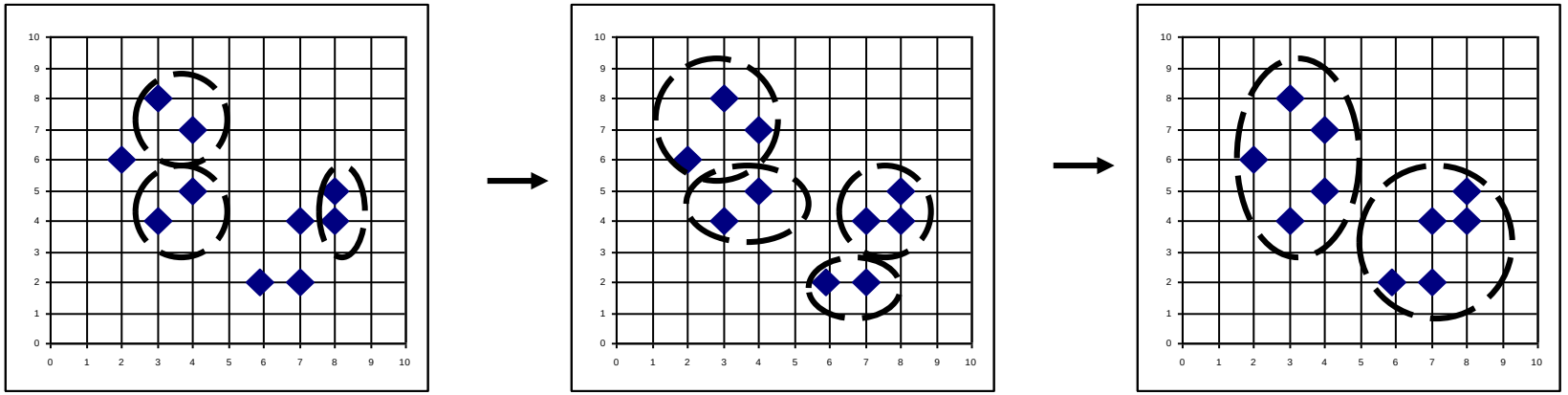
典型算法—层次方法

- 凝聚的(agglomerative)和分裂的(divisive)层次聚类图示



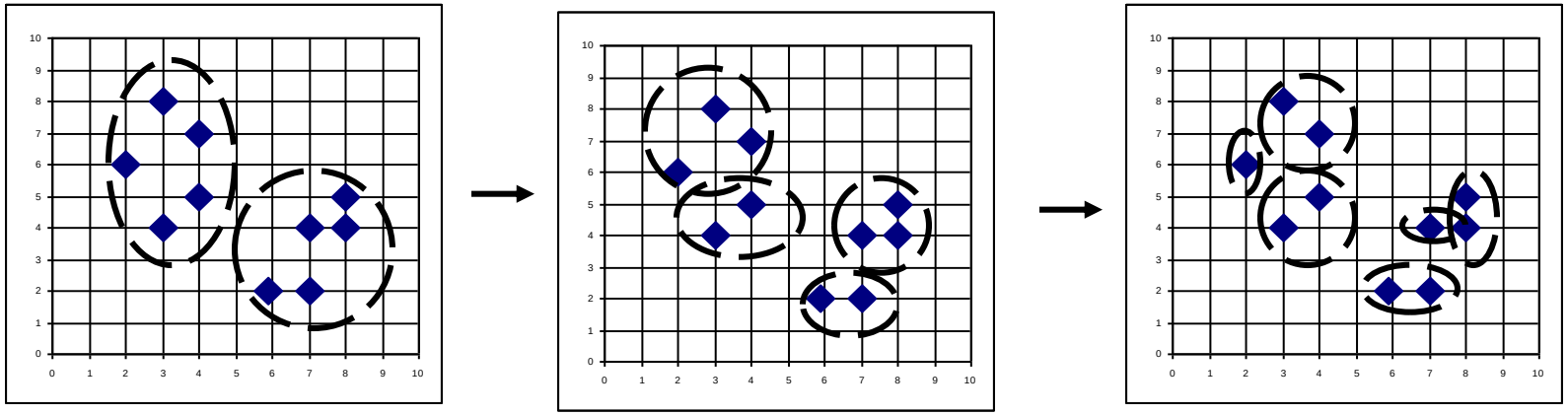
AGNES (Agglomerative Nesting)

- 由 Kaufmann和Rousseeuw提出(1990)
- 使用单链接(Single-Link)方法和距离矩阵
- 合并具有最小距离的节点
- 以非递减的方式继续
- 最终所有的节点属于同一个簇



DIANA (Divisive Analysis)

- 由 Kaufmann和Rousseeuw提出 (1990)
- 是 AGNES的逆
- 最终每个节点自己形成一个簇



典型算法—层次方法 $\|x-y\|_2 = \left(\sum_i (x_i - y_i)^2\right)^{1/2}$

- 四个广泛采用的簇间距离度量方法
  - 最小距离：$d_{min}(C^i,C^j) = min_{\ p\in Cj,\ p'\in Cj} |p-p'|$
  - 最大距离：$d_{max}(C^i,C^j) = max_{\ p\in Ci,\ p'\in Cj} |p-p'|$
  - 平均值的距离：$d_{mean}(C^i,C^j) = | m_i - m_j |$
  - 平均距离：$d_{avg}(C^i,C^j) = \sum_{p\in Ci} \sum_{p'\in Cj} |p-p'| / n_i n_j$

其中, $|p-p'|$是两个向量p和p'之间的距离

m^i是簇C^i的平均值，n^i是簇C^i中对象的数目

典型算法—层次方法

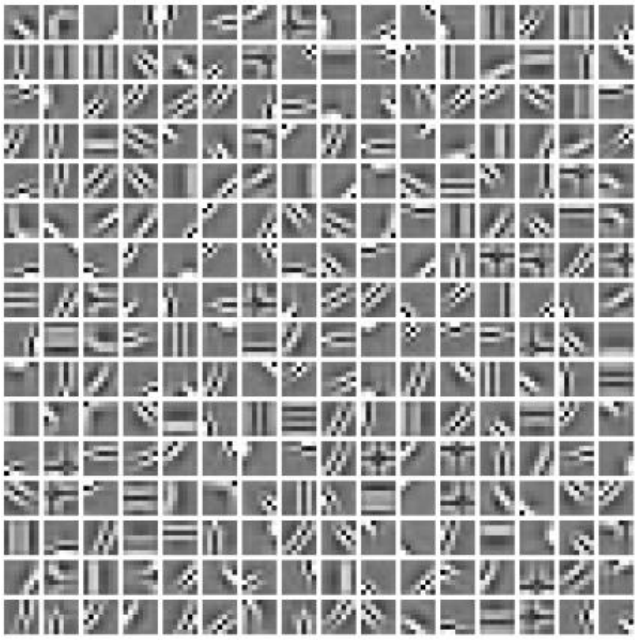
- 层次聚类的主要缺点
  - 不具有很好的可伸缩性: 时间复杂性至少是 $O(n^2)$, 其中 n 样本总数
  - 合并或分裂的决定需要检查和估算大量的对象或簇
  - 不能撤消已做的处理, 聚类之间不能交换对象. 如果某一步没有很好地选择合并或分裂的决定, 可能会导致低质量的聚类结果

提 纲

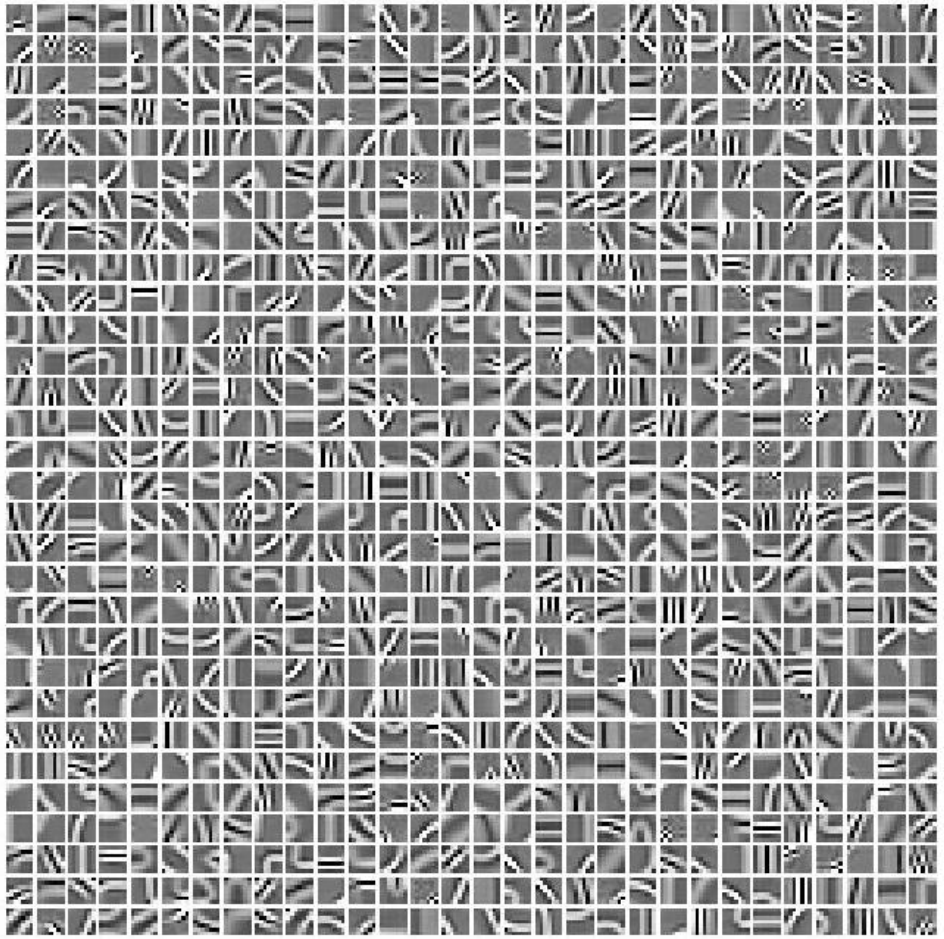
- 概述
- K-means聚类算法
- 聚类算法的距离度量
- 层次聚类算法
- 从聚类到Unsupervised Learning

从聚类到Unsupervised Learning

- Unsupervised learning
  - Whitten (PCA)
  - K-means clustering



For Chinese Text



For English-and-Chinese Text

Adam Coates， Honglak Lee， Adrew. N.G “An Analysis of Single-Layer Networks in Unsupervised Feature Learning,” NIPS 2011.

从聚类到Unsupervised Learning

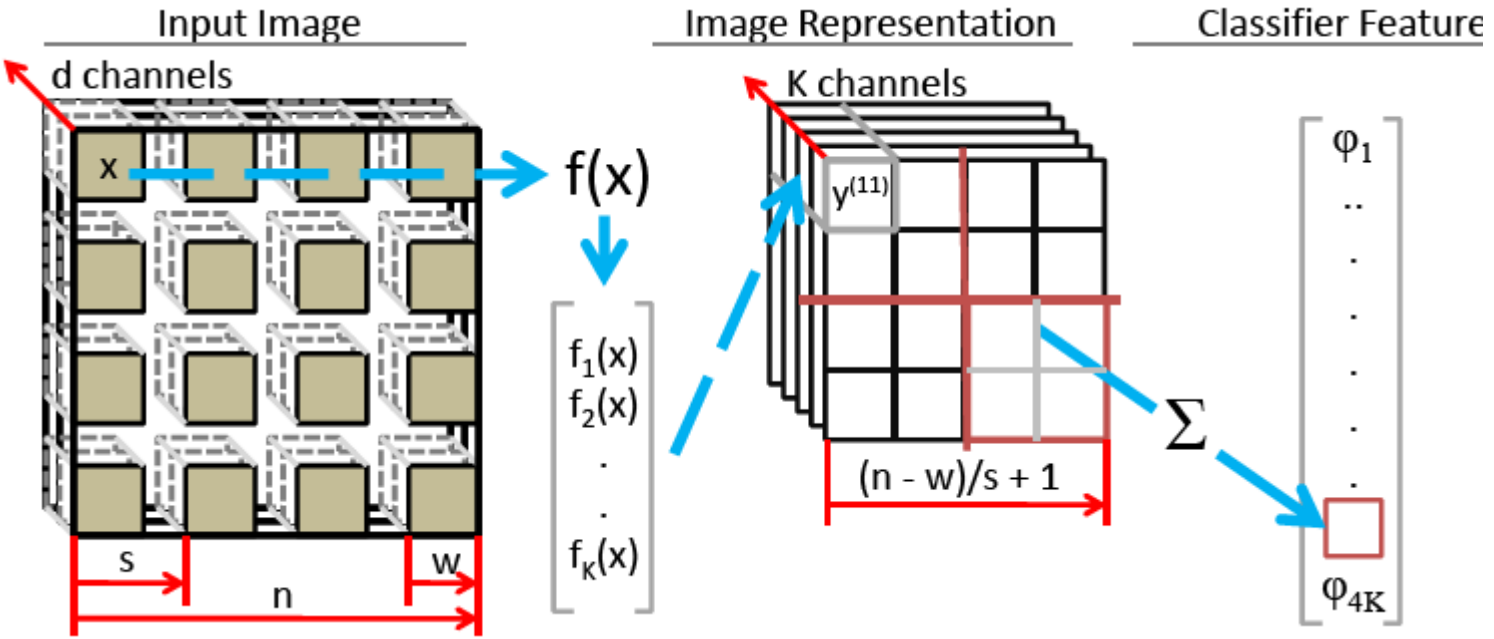
- Learning framework
  - 1. Extract random patches from unlabeled training images.
  - 2. Apply a pre-processing stage to the patches.
  - 3. Learn a feature-mapping using an unsupervised learning algorithm.

- Feature extraction procedures
  - 1. Extract features from equally spaced sub-patches covering the input image.
  - 2. Pool features together over regions of the input image to reduce the number of feature values.
  - 3. Train a linear classifier to predict the labels given the feature vectors.

Adam Coates， Honglak Lee， Adrew. N.G “An Analysis of Single-Layer Networks in Unsupervised Feature Learning,” NIPS 2011.

从聚类到Unsupervised Learning

- Feature Extraction with unsupervised learning
  - Filtering
  - Pooling



Adam Coates， Honglak Lee， Adrew. N.G “An Analysis of Single-Layer Networks in Unsupervised Feature Learning,” NIPS 2011.

从聚类到Unsupervised Learning

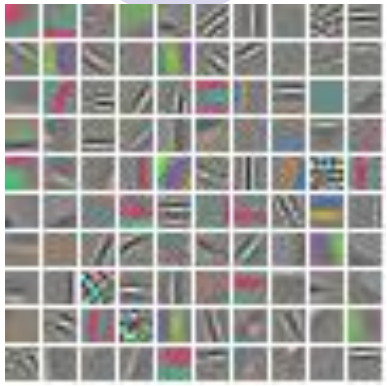
- unsupervised learning algorithms
  - K-means clustering
  - Sparse auto-encoder
  - Sparse restricted Boltzmann machine
  - Gaussian mixtures

Adam Coates， Honglak Lee， Adrew. N.G “An Analysis of Single-Layer Networks in Unsupervised Feature Learning,” NIPS 2011.

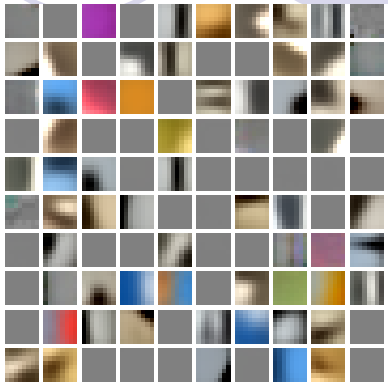
从聚类到Unsupervised Learning

K-means+Whitten

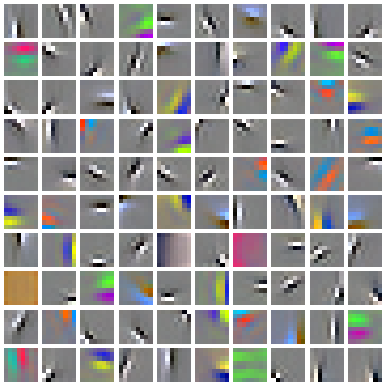
K-means



GMMs+Whitten

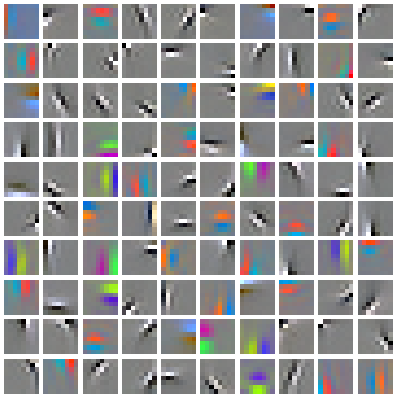


GMMs



Sparse autoencoder +Whitten

Sparse autoencoder



RBM+Whitten

RBM

无监督单层学习在**CIFAR-10**图像数据集合上学习到的特征

Adam Coates， Honglak Lee， Adrew. N.G “An Analysis of Single-Layer Networks in Unsupervised Feature Learning,” NIPS 2011.

从聚类到Unsupervised Learning

Table 1: Test recognition accuracy on CIFAR-10

Algorithm	Accuracy
Raw pixels (reported in [13])	37.3%
3-Way Factored RBM (3 layers) [24]	65.3%
Mean-covariance RBM (3 layers) [23]	71.0%
Improved Local Coord. Coding [33]	74.5%
Conv. Deep Belief Net (2 layers) [14]	78.9%
Sparse auto-encoder	73.4%
Sparse RBM	72.4%
K-means (Hard)	68.6%
K-means (Triangle)	77.9%
K-means (Triangle, 4000 features)	**79.6%**

Table 2: Test recognition accuracy (and error) for NORB (normalized-uniform)

Algorithm	Accuracy (error)
Conv. Neural Network [16]	93.4% (6.6%)
Deep Boltzmann Machine [26]	92.8% (7.2%)
Deep Belief Network [20]	95.0% (5.0%)
(Best result of [11])	94.4% (5.6%)
Deep neural network [27]	**97.13% (2.87%)**
Sparse auto-encoder	96.9% (3.1%)
Sparse RBM	96.2% (3.8%)
K-means (Hard)	96.9% (3.1%)
K-means (Triangle)	97.0% (3.0%)
K-means (Triangle, 4000 features)	**97.21% (2.79%)**

学习到的特征用于监督分类时的性能比较

Adam Coates， Honglak Lee， Adrew. N.G “An Analysis of Single-Layer Networks in Unsupervised Feature Learning,” NIPS 2011.

其他聚类方法

- 基于密度的聚类…
- 基于网格的聚类…
- 统计推断与生成模型
  - 混合高斯模型GMMs
  - **从Laplacian矩阵到谱聚类**