无监督学习-聚类算法

信息技术学院

夫鸟同翼者而聚居，兽同足者而俱行。

——《战国策》

聚类

- 聚类(Clustering)是指将不同的对象划分成由多个对象组成的多个类的过程。
- 由聚类产生的数据分组，同一组内的对象具有相似性，不同组的对象具有相异性。
- 聚类待划分的类别未知，即训练数据没有标签。
- 聚类属于无监督学习。

簇

- 簇(cluster)是由距离邻近的对象组合而成的集合。聚类的最终目标是获得紧凑、独立的簇集合。一般采用相似度作为聚类的依据，两个对象的距离越近，其相似度就越大。

常见聚类算法

- 按照簇的定义和聚类的方式，聚类大致分为以下几种：
  - K-Means为代表的簇中心聚类、
  - 基于连通性的层次聚类、
  - 以EM算法为代表的概率分布聚类、
  - 以DBSCAN为代表的基于网格密度的聚类，
  - 以及高斯混合聚类等。

6.1.2 K-Means聚类

- K-Means聚类算法也称为K均值聚类算法，是典型的聚类算法。
- 对于给定的数据集和需要划分的类数k，算法根据距离函数进行迭代处理，动态地把数据划分成k个簇（即类别），直到收敛为止。
- 每个样本到其所属簇的中心的距离最小。
- 簇中心(cluster center)也称为聚类中心。

K均值聚类

- K 均值聚类的策略是通过目标函数的最小化选取最优的划分
- 使用欧氏距离平方作为样本之间的距离 $d(x_i, x_j)$

$$d(x_i, x_j) = \sum_{k=1}^{m} (x_{ki} - x_{kj})^2$$

$$= \|x_i - x_j\|^2$$

K均值聚类

目标函数

- 定义为样本与其所属簇的中心之间的距离的总和

$$W(C)=\sum_{l=1}^{k}\sum_{C(i)=l}\|x_i-\bar{x}_l\|^2$$

- 函数 $W(C)$ 也称为能量，表示相同簇中的样本相似的程度。

K均值聚类

- K 均值聚类就是求解最优化问题：

$$C^* = \arg\min_C W(C)$$

$$= \arg\min_C \sum_{l=1}^{k} \sum_{C(i)=l} \|x_i - \bar{x}_l\|^2$$

- 相似的样本被聚到同簇时，目标函数值最小，这个目标函数的最优化能达到聚类的目的。

迭代算法

- K均值聚类的算法是一个迭代的过程，每次迭代包括两个步骤。

- 首先选择 K 个簇的中心，将样本逐个指派到与其最近的中心的簇中，得到一个聚类结果；
- 然后更新每个簇的样本的均值，作为簇的新的中心。
- 重复以上步骤，直到收敛为止。

聚类的运算流程

随机选择 k个数据点 -> 起始簇中心
While数据点的分配结果发生改变:
    for数据集中的每个数据点p:
        for 循环访问每个簇中心c:
            computer_distance(p,c)
            将数据点p分配到最近的簇
    for 每一个簇:
        簇中心更新为簇内数据点的均值

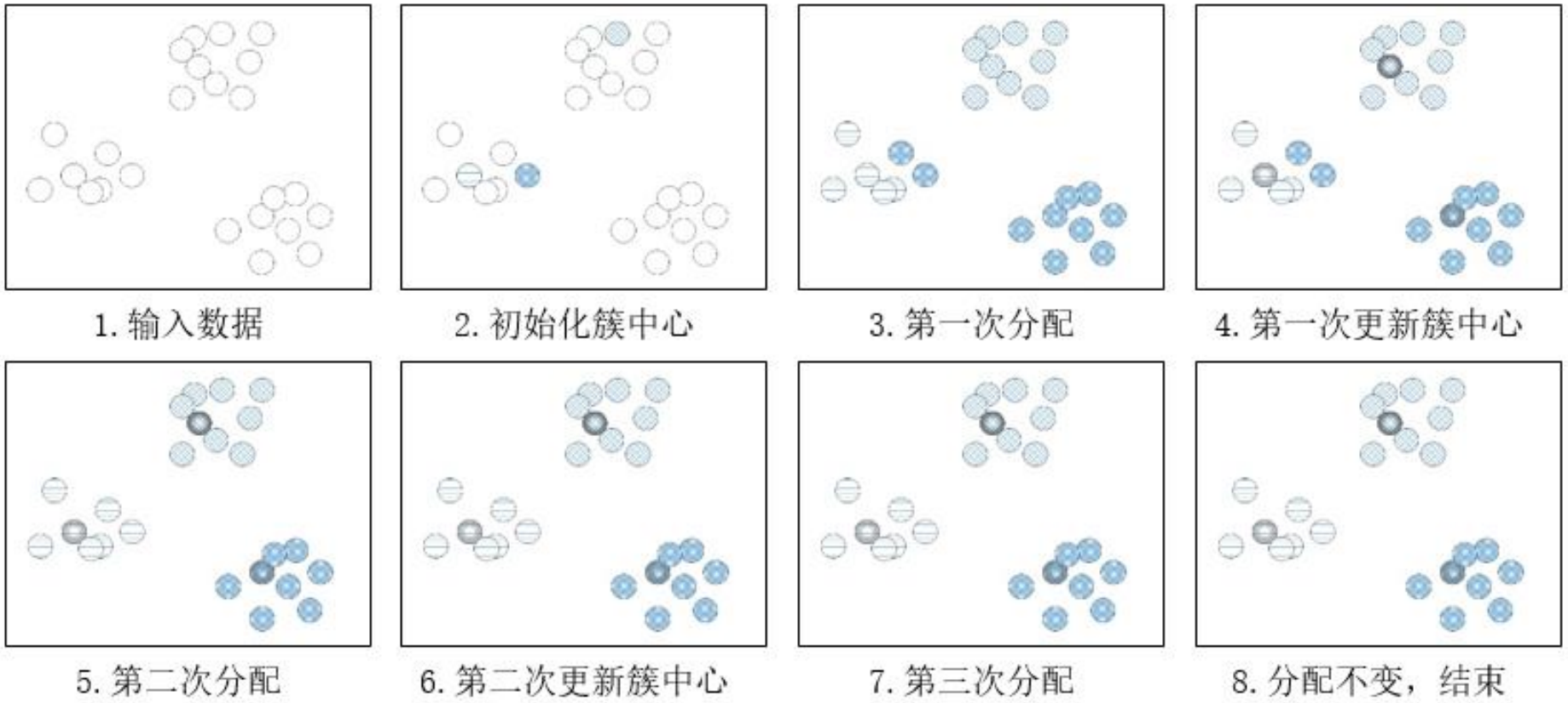

1. 输入数据
2. 初始化簇中心
3. 第一次分配
4. 第一次更新簇中心
5. 第二次分配
6. 第二次更新簇中心
7. 第三次分配
8. 分配不变，结束

图 6.1 聚类过程示意图

初始中心

- K 均值聚类属于启发式方法，不能保证收敛到全局最优，初始中心的选择会直接影响聚类结果。
- 注意，簇中心在聚类的过程中会发生移动，但是往往不会移动太大，因为在每一步，样本被分到与其最近的中心的簇中。

- 选择不同的初始中心，会得到不同的聚类结果。
- 初始中心的选择，比如可以用层次聚类对样本进行聚类，得到k个簇时停止。然后从每个簇中选取 一个与中心距离最近的点。

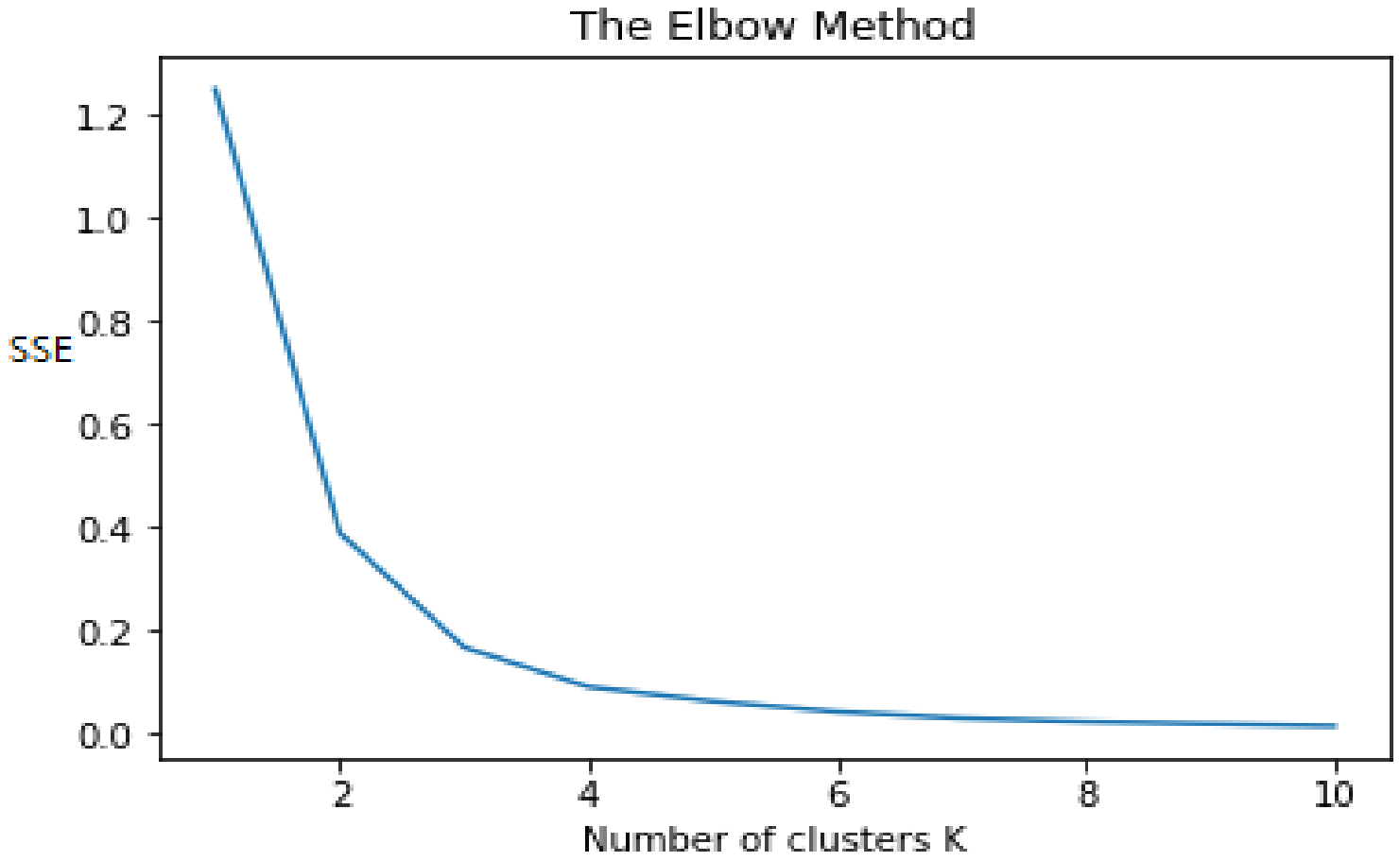
科学确定k值

研究人员提出了很多确定k值的方法，常见的如：

1. 经验值
2. 观测值
3. 肘部方法（Elbow Method）
4. 性能指标法

肘部方法

- 示例：误差平方和拐点出现在k=4位置。



K均值法特点

- 簇的数量 K 事先指定
- 以欧氏距离平方表示样本之间的距离，以中心或样本的均值表示簇
- 以样本和其所属簇的中心之间的距离的总和为最优化的目标函数

- 得到的类别是平坦的、非层次化的
- 算法是迭代算法，不能保证得到全局最优。

- K-Means聚类的优点是算法简单、运算速度快，即便数据集很大计算起来也便捷。
- 不足之处是如果数据集较大，容易获得局部最优的分类结果。而且所产生的类的大小相近，对噪声数据也比较敏感。

K 值

- K 均值聚类中的簇数 K 值需要预先指定，而在实际应用中最优的 K 值是不知道的。
- 尝试用不同的 K 值聚类，检验得到聚类结果的质量，推测最优的 K 值。

- 聚类结果的质量可以用簇的平均直径来衡量。
- 一般地，簇的数量变小时，平均直径会增加；
- 簇的数量变大超过某个值以后，平均直径会不变，而这个值正是最优的 K 值。

K均值应用

图像分割

objects in cluster 1

objects in cluster 2

objects in cluster 3

使用SKlearn实现K-Means聚类

Scikit-learn的cluster模块中提供的KMeans类可以实现K-均值聚类，构造函数如下：
sklearn.cluster.KMeans(n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='auto', verbose=0, random_state=None, copy_x=True, n_jobs=None, algorithm='auto')
主要参数含义：
- n_clusters：可选，默认为8。要形成的簇的数目，即类的数量。
- n_init：默认为10，用不同种子运行k-均值算法的次数。
- max_iter：默认300，单次运行的k-均值算法的最大迭代次数。

返回KMeans对象的属性包括：
- cluster_centers_：数组类型，各个簇中心的坐标。
- labels_：每个数据点的标签。
- inertia_：浮点型，数据样本到它们最接近的聚类中心的距离平方和。
- n_iter_：运行的迭代次数。

例

◆使用sklearn.cluster.KMeans进行k-均值聚类。

训练数据： 六个数据点 [1, 2], [1, 4], [1, 0], [4, 2], [4, 4], [4, 0]。

测试数据： [0, 0], [4, 4]

◆聚类结果：

```
k labels are: [1 0 1 0 0 0]
predict results are: [1 0]
cluster centers are: [[3.25 2.5 ]
 [1.   1.  ]]
```