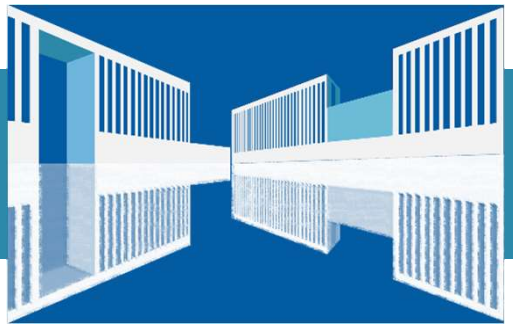
北京航空航天大学
BEIHANGUNIVERSITY

人工智能及其航空应用

第十章 聚类——航天器零部件包装标准化方法



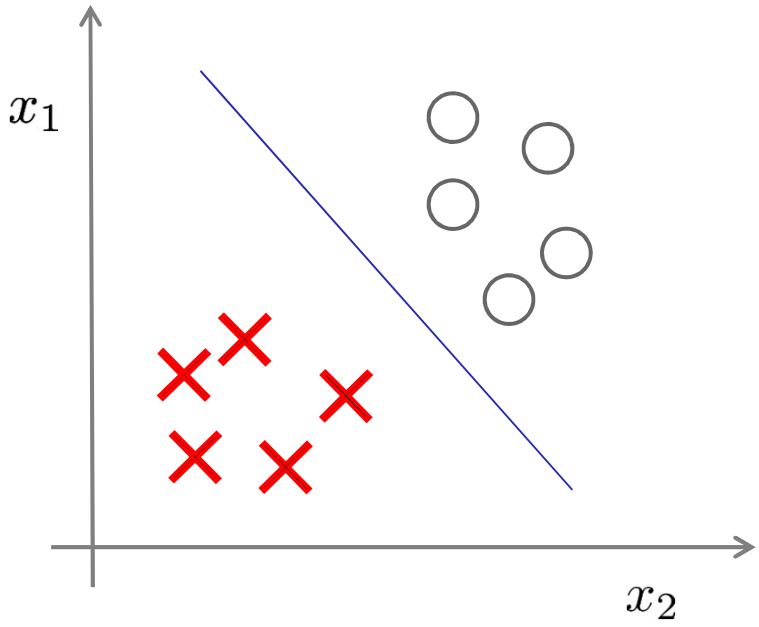
目录

0. 介绍
1. 航天器零部件货运装箱背景
2. 聚类理论
3. K均值算法
4. 问题示例
5. 总结与作业
6. 知识扩展

0 介绍

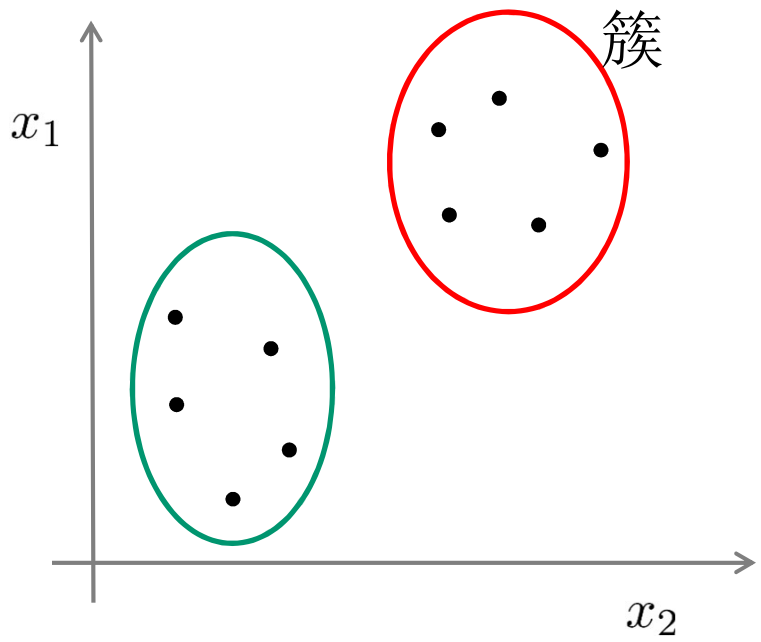
- 机器学习中有两大类问题：
  - 监督学习（分类）
    - 训练样本中包含一些已知的标记信息，其目标是使学习模型通过对有标记信息样本的学习，实现对未知标记的样本进行分类
  - 无监督学习（聚类）
    - 训练样本不包含标记信息，其目标是使学习模型通过对无标记训练样本的学习来获得数据之间的内部联系或性质

监督学习



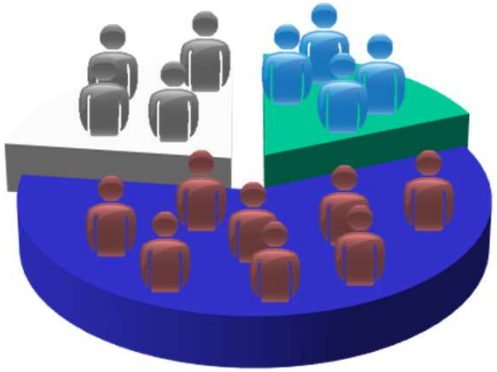
训练数据集:$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

无监督学习

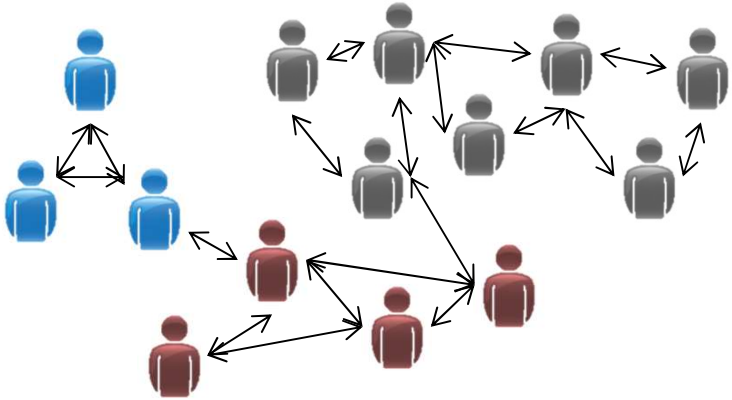


训练数据集：$\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$

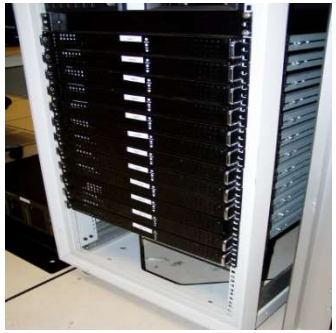
聚类应用



市场划分



社交网络分析



计算机集群组织

天文数据分析

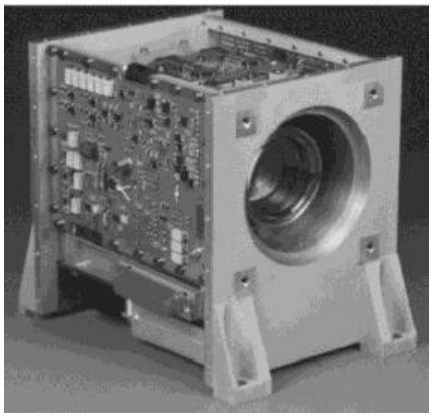
聚类应用

- 航空航天中的应用
  - 零部件的故障诊断
  - 航空航天大数据分析
  - 飞行安全评估
  - 遥感数据分析
  - 航空发动机寿命预测与维修决策
  - 货运网络布局

航天器的生产加工过程中，对零部件运输中的包装尺寸、材料、方式等进行标准化，使包装规格种类适当、形状规范，提高装箱效率降低物流运输成本。

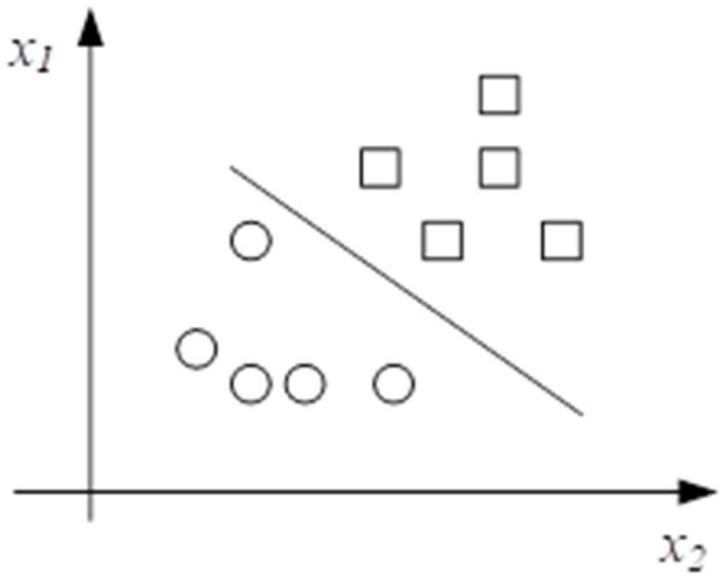
1 航天器零部件货运装箱背景

- 航天器零部件多，测试、生产过程中，需在零部件生产车间、测试工厂、装配车间之间运输
- 根据零部件结构、尺寸、材料等特性，使用多种不同规格的包装材料，包括纸箱、木箱、塑料包装箱等，实现零部件产品的外观与功能不受破坏，实现产品的高效周转

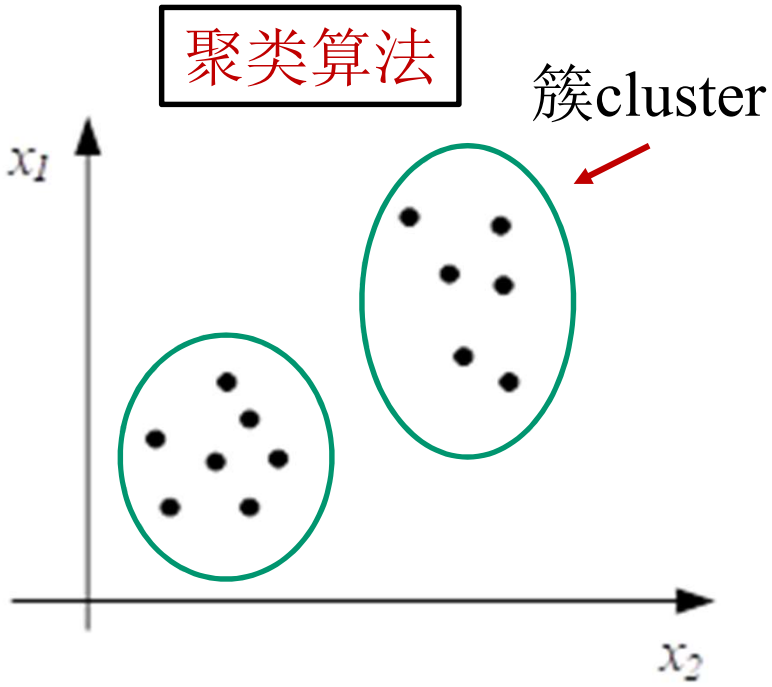


2 聚类理论

聚类的任务：将数据集中的样本划分为不同的分组



（1）监督学习

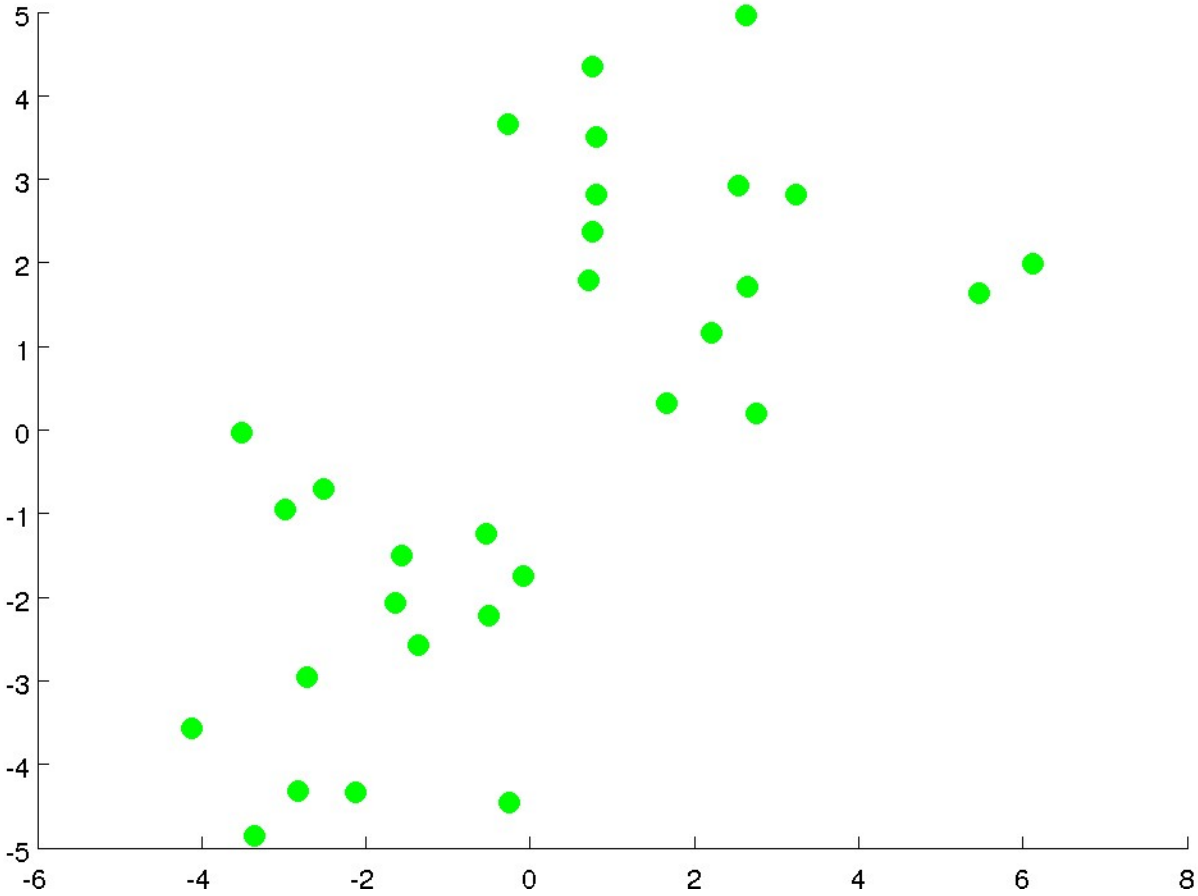


（2）无监督学习

目录

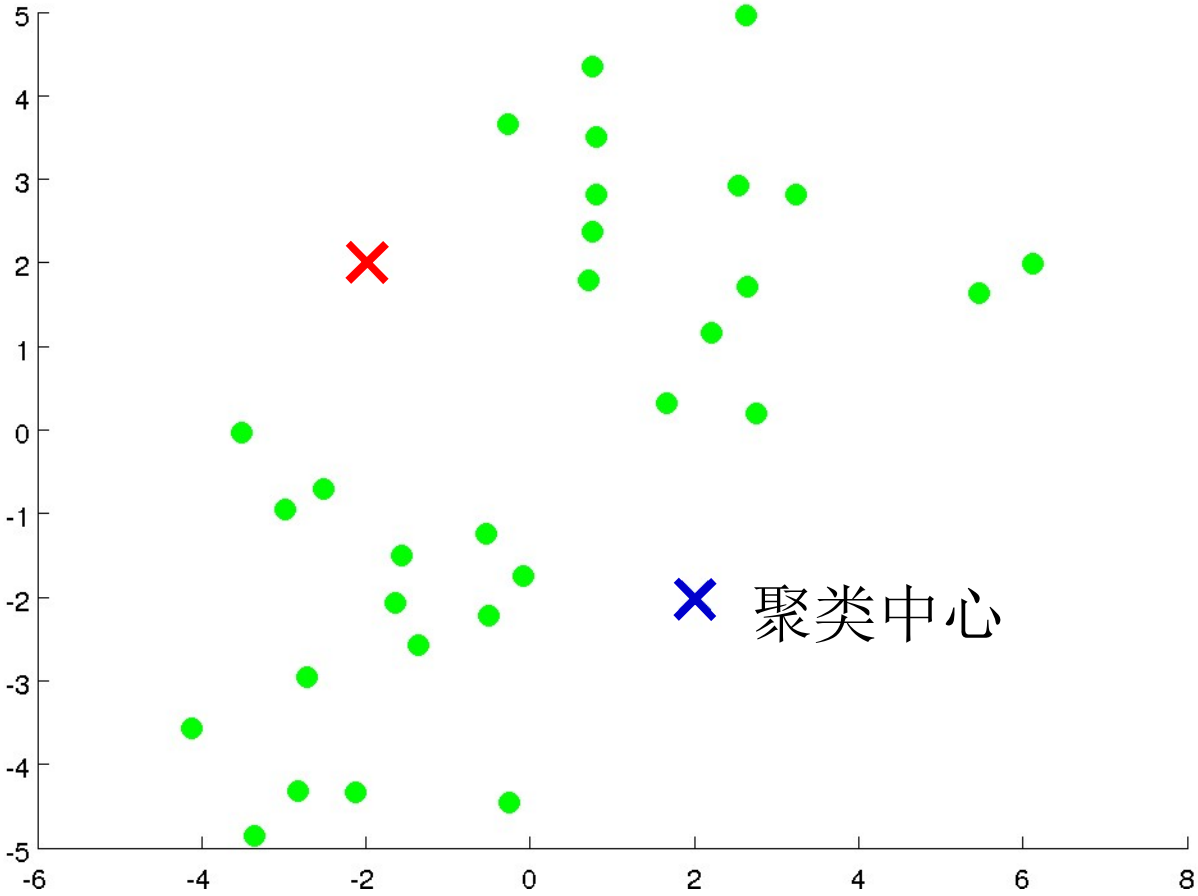
0. 介绍
1. 航天器零部件货运装箱背景
2. 聚类理论
3. K均值算法
4. 问题示例
5. 总结与作业
6. 知识扩展

K-均值算法



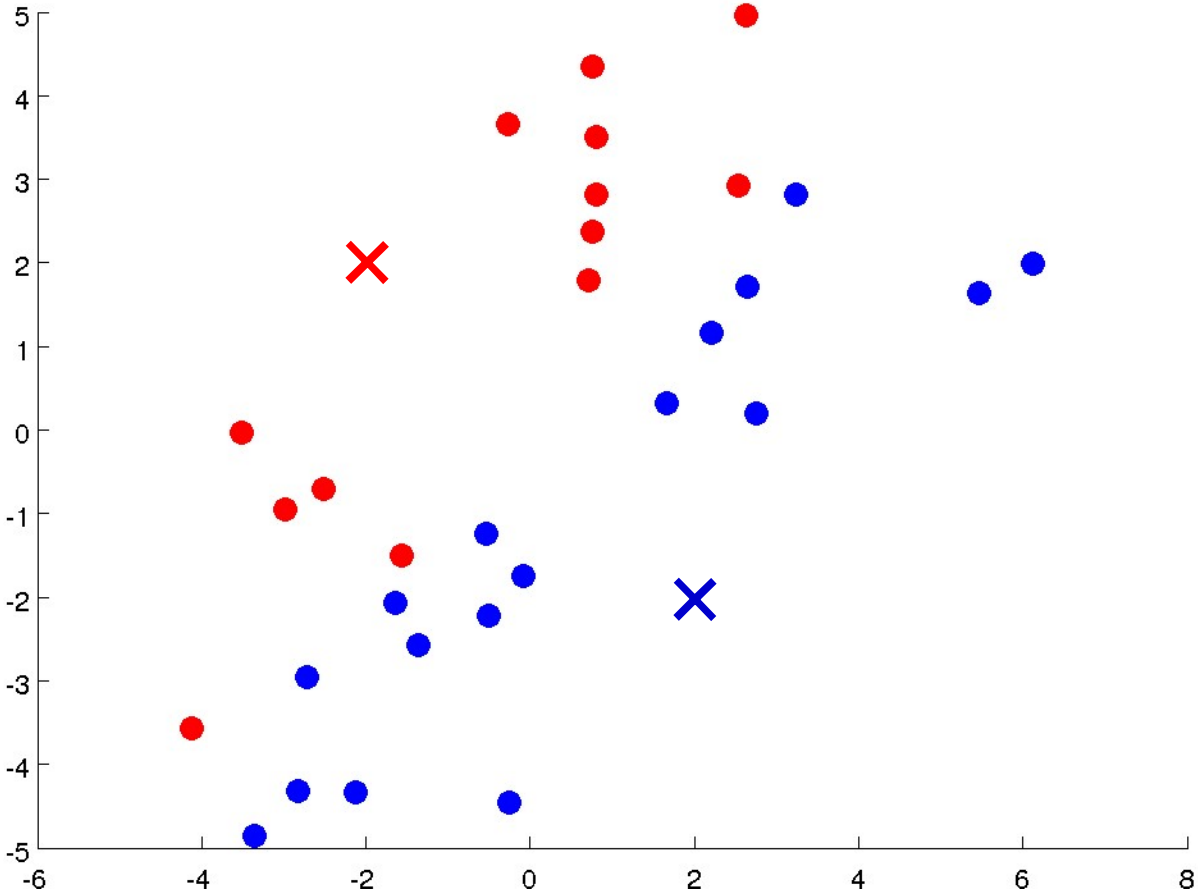
有一个无标签数据集

K-均值算法



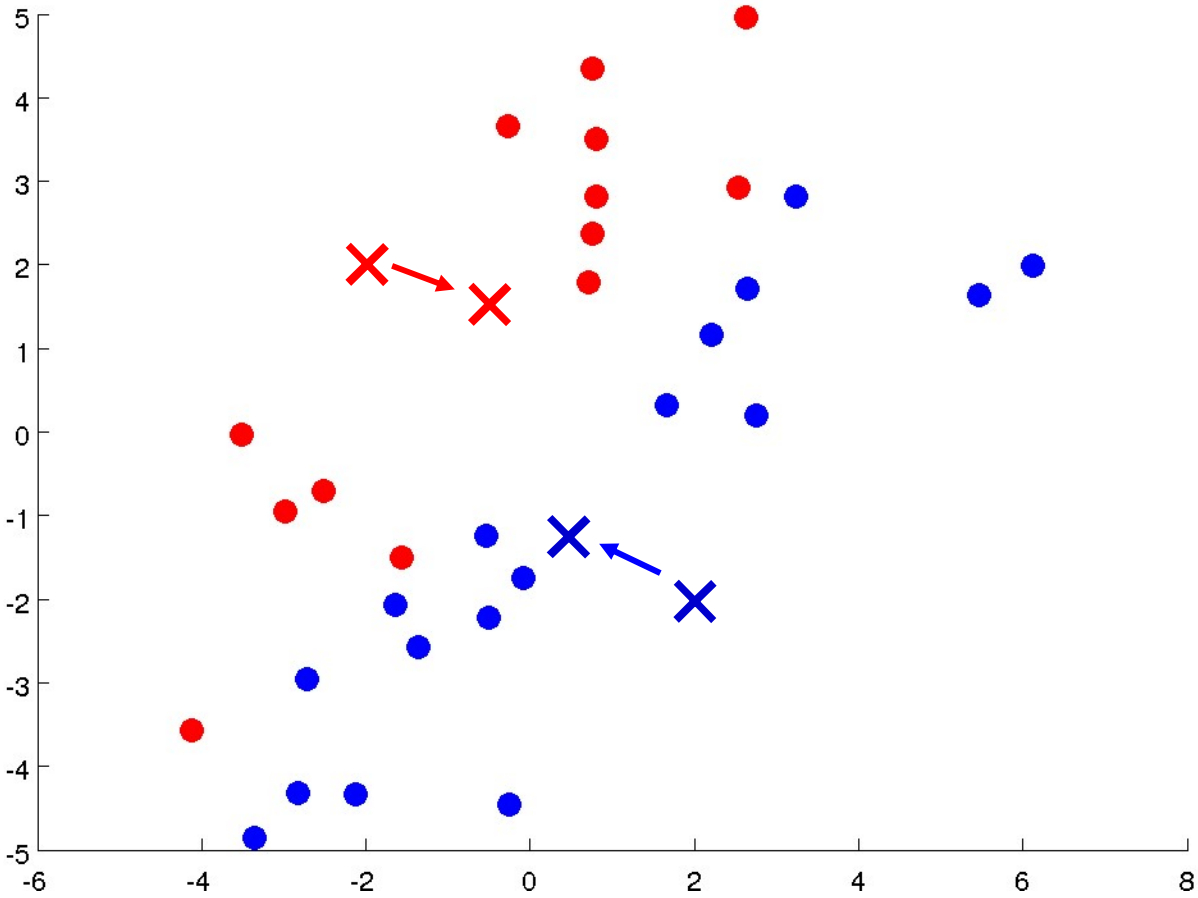
随机选择2个点（聚类中心）

K-均值算法



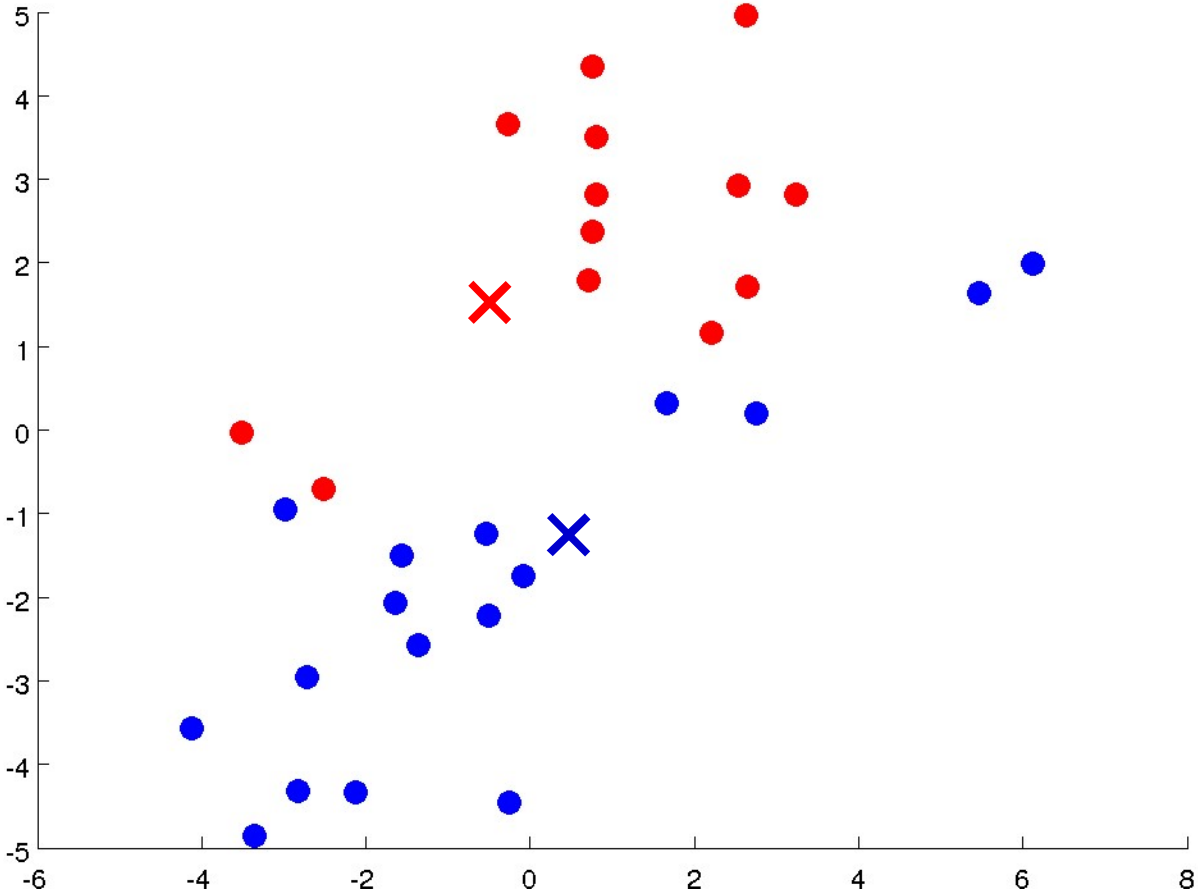
按照距离这2个中心点的距离，将其与距离最近的中心点关联起来，聚成一类

K-均值算法



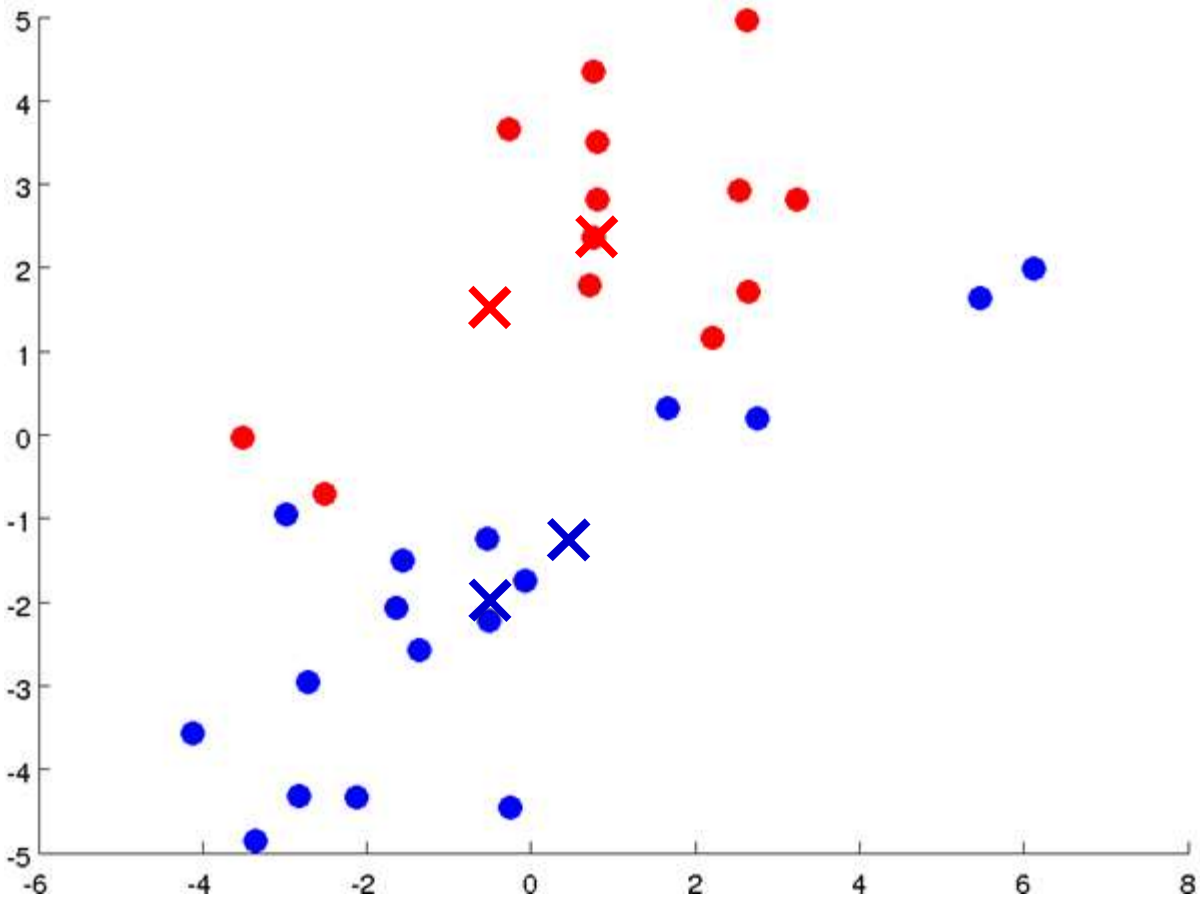
计算每一个簇的平均值，将该簇聚类中心点移动到平均值的位置

K-均值算法

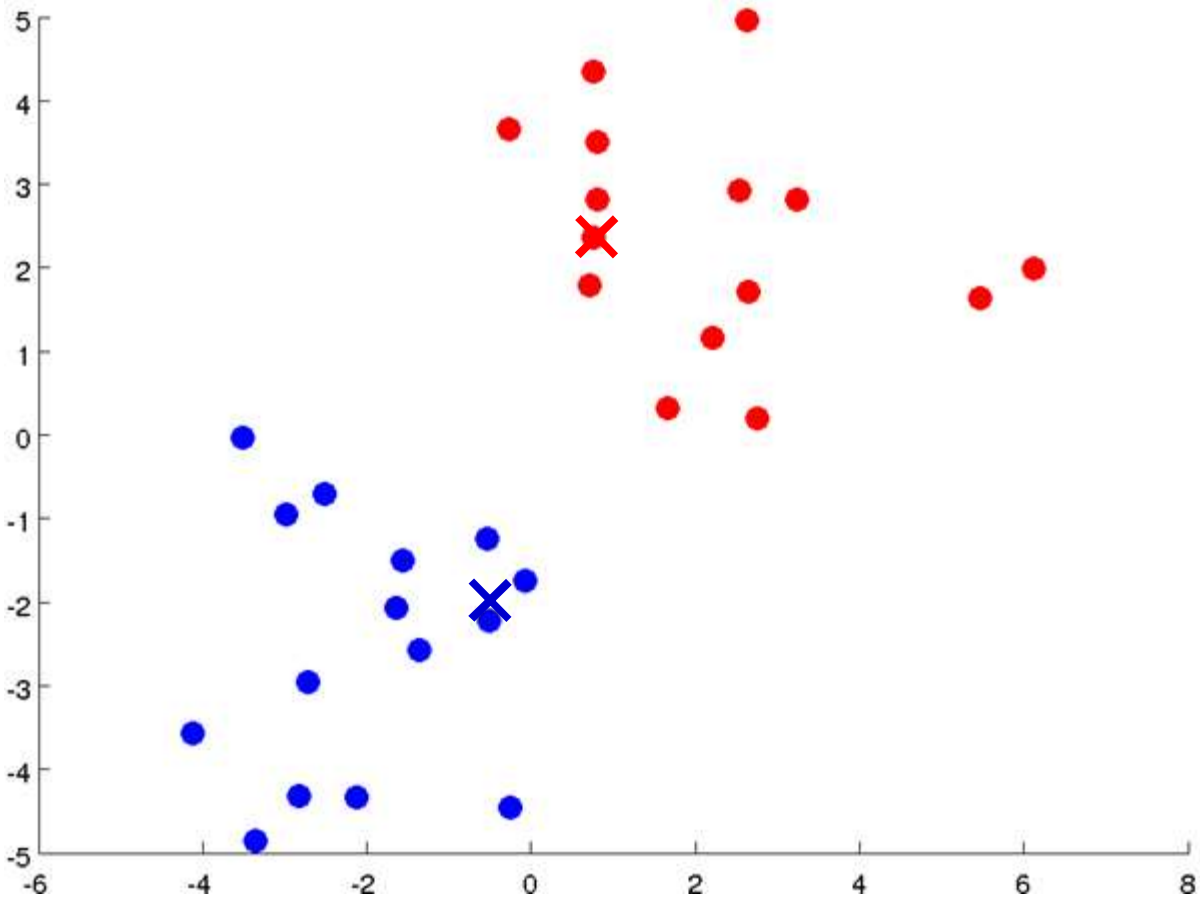


再按照各样本距离簇中心点的距离，将其与距离最近的中心点关联起来

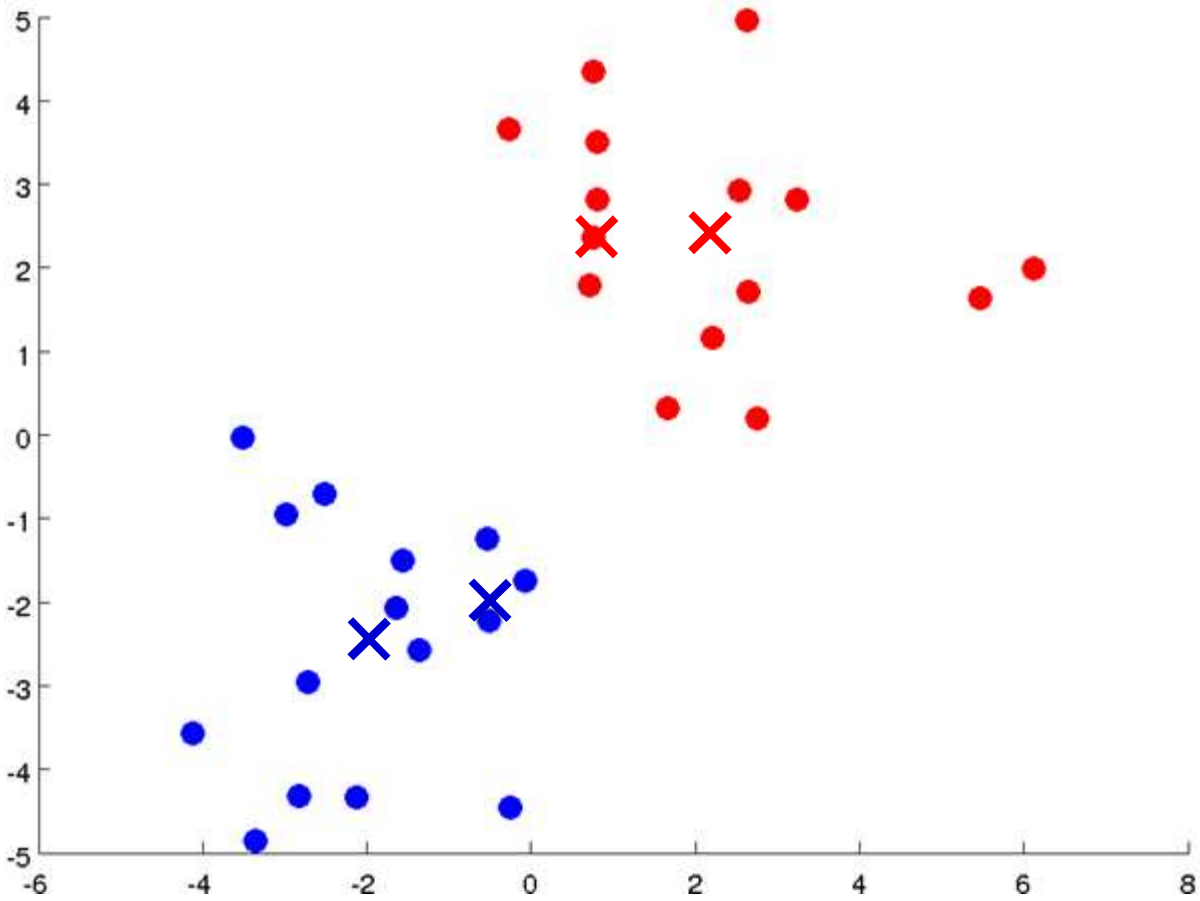
K-均值算法



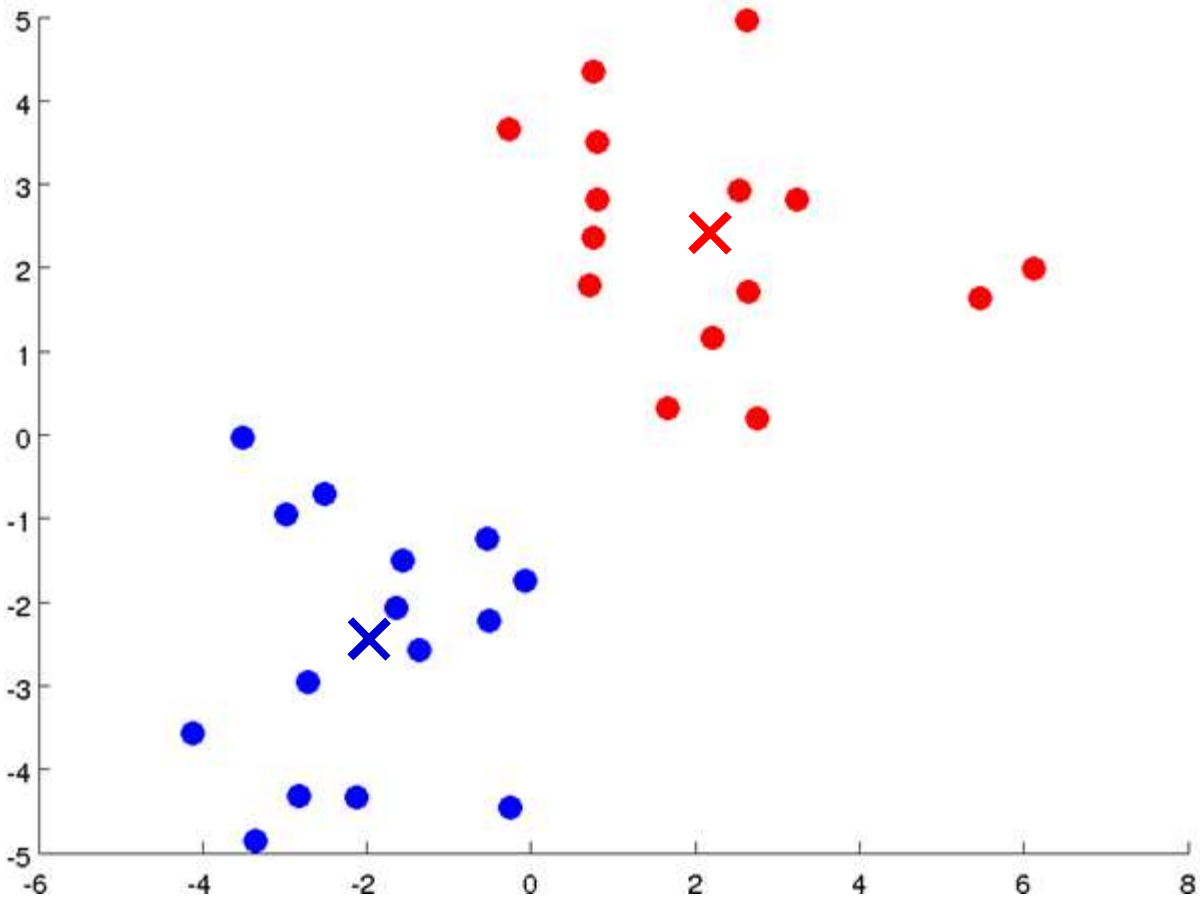
K-均值算法



K-均值算法



K-均值算法



K-均值算法

算法输入：

- 数据集：$\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$
- K (聚类数)

K-均值算法

样本与聚类中心之间的距离来定义样本和某聚类之间的相似度度量，距离越大，相似度越小，距离越小，相似度越大。

距离的计算：

$$x^{(i)} = \left(x_1^{(i)},\ x_2^{(i)}, x_3^{(i)}, \cdots,\ x_n^{(i)}\right)$$
$$x^{(j)} = (x_1^{(j)},\ x_2^{(j)}, x_3^{(j)}, \cdots,\ x_n^{(j)})$$

最常用使用欧式距离计算方法：

$$dist\left(x^{(i)},\ x^{(j)}\right) = \left|\left|x^{(i)} - x^{(j)}\right|\right|^2 = \sqrt{\sum_{u=1}^{n} |x_u^{(i)} - x_u^{(j)}|^2}$$

K-均值算法

K-均值算法的基础：最小化误差平方和。

$$J(c,\mu)=\sum_{i=1}^{k}\left\|x^{(i)}-\mu_{c^{(i)}}\right\|^2$$

$\mu_{c(i)}$表示第i个聚类的均值

代价函数最小！
各簇内的样本越相似，其与该簇均值间的误差平方越小。

K-均值算法

算法流程：

STEP 1： 随机选取 K个聚类质心点

STEP 2：重复下面过程直到收敛

{

对于每一个样例 i，计算其应该属于的类：

$$c(i) := arg\,\min_j \left\| x^{(i)} - \mu_j \right\|^2$$

对于每一个类 j，重新计算该类的质心：

$$\mu_j := \frac{\sum_{i=1}^{m} l\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} l\{c^{(i)} = j\}}$$

}

K-均值算法

算法伪代码：

****

创建k个点作为初始的质心点（随机选择）

当任意一个点的簇分配结果发生改变时

对数据集中的每一个数据点

对每一个质心

计算质心与数据点的距离

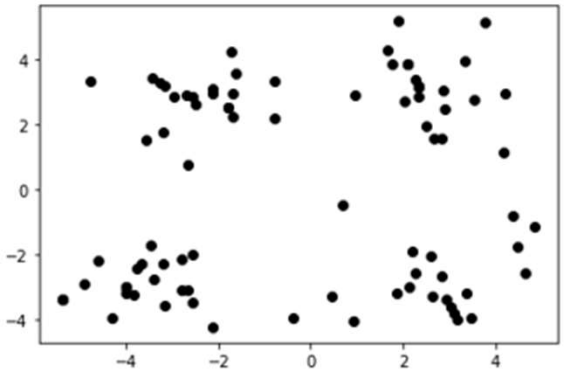
将数据点分配到距离最近的簇

对每一个簇，计算簇中所有点的均值，并将均值作为质心

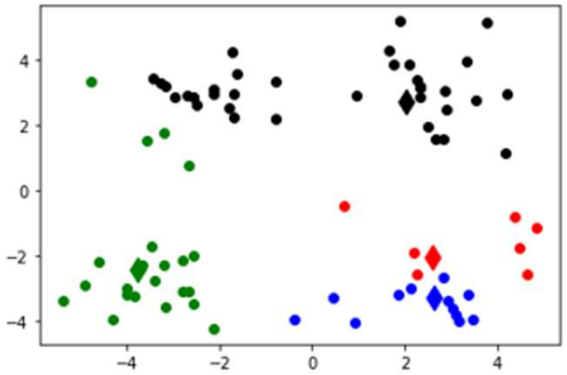
****

K-均值算法

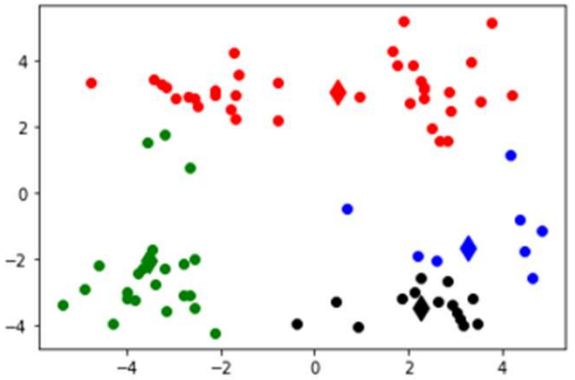
算法过程示例



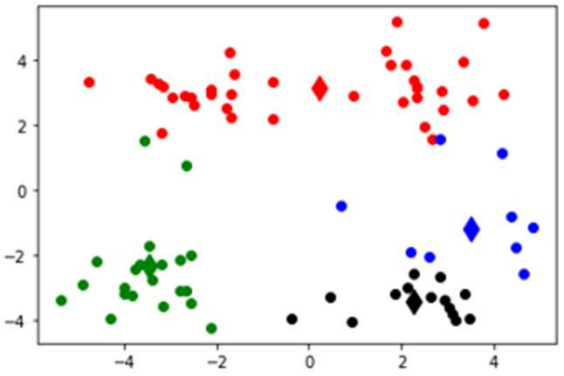
（1）输入数据集



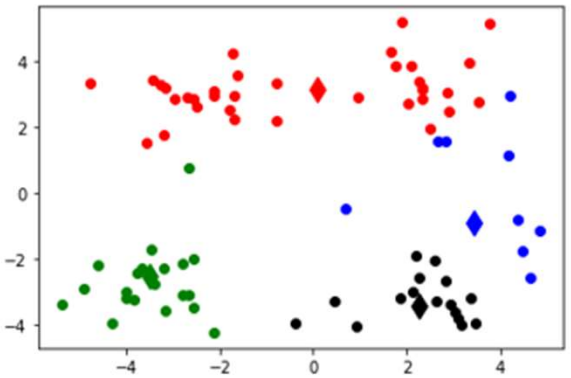
（2）初始化



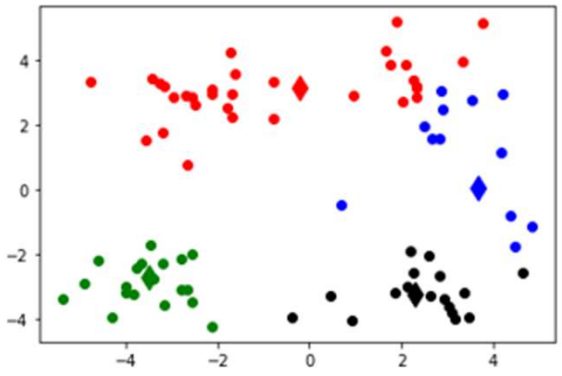
（3）迭代计算 1



（4）迭代计算 2



（5）迭代计算 3

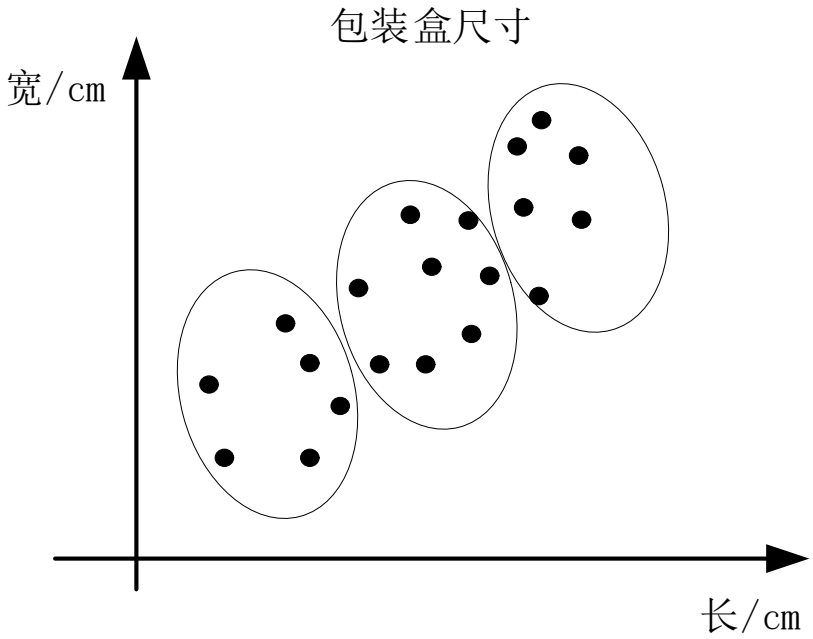


（6）迭代计算 4

K-均值算法

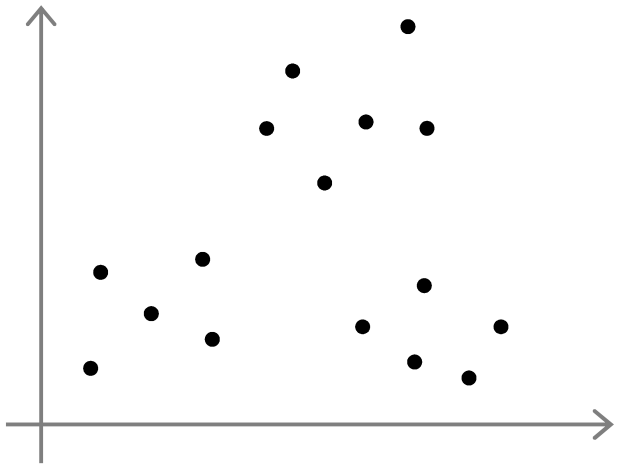
有明显界限的数据集

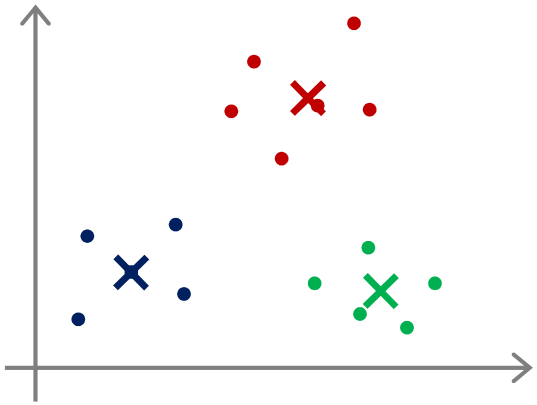
没有明显界限的数据集

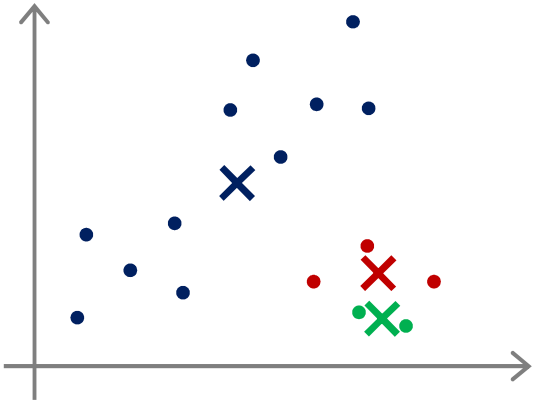


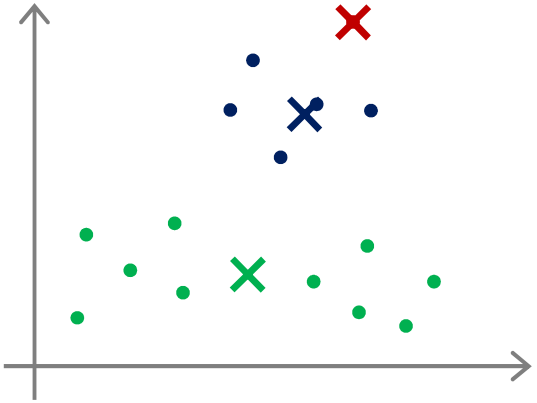
K-均值算法

局部最佳的问题









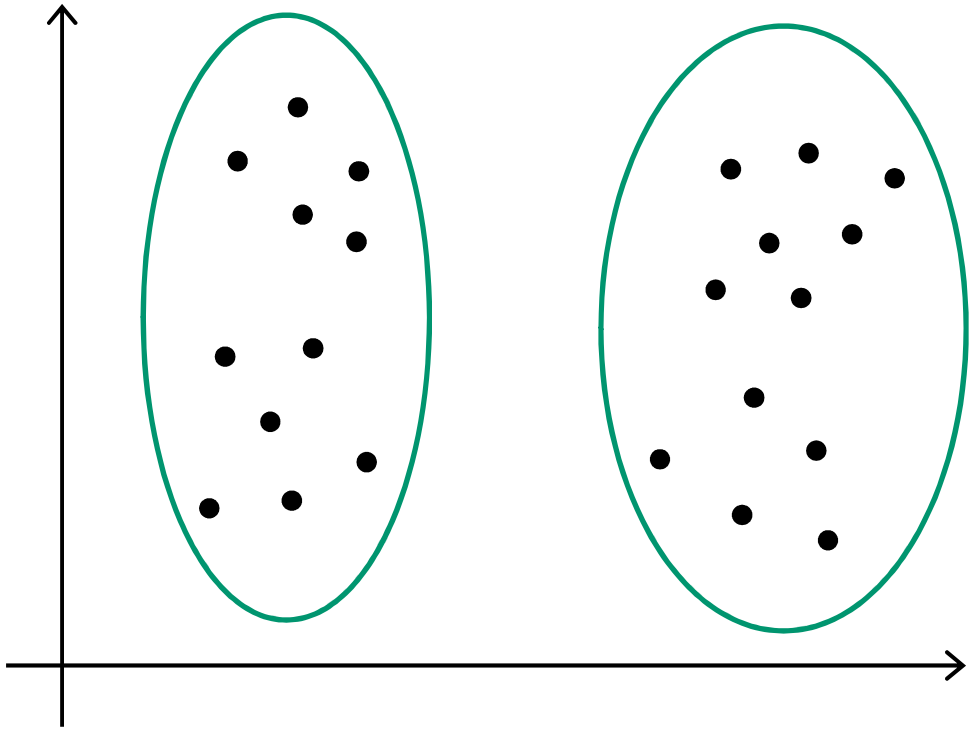
随机初始化

对于局部最优问题，多次运行K均值算法，每一次都重新进行随机初始化，最后再比较多次运行K均值的结果，选择代价函数最小的结果。

在K较小的时候（2~10）是可行的。如果K较大，这么做也可能不会有明显地改善。

K-均值算法

聚类数的选择?



分2类、3类、4类?

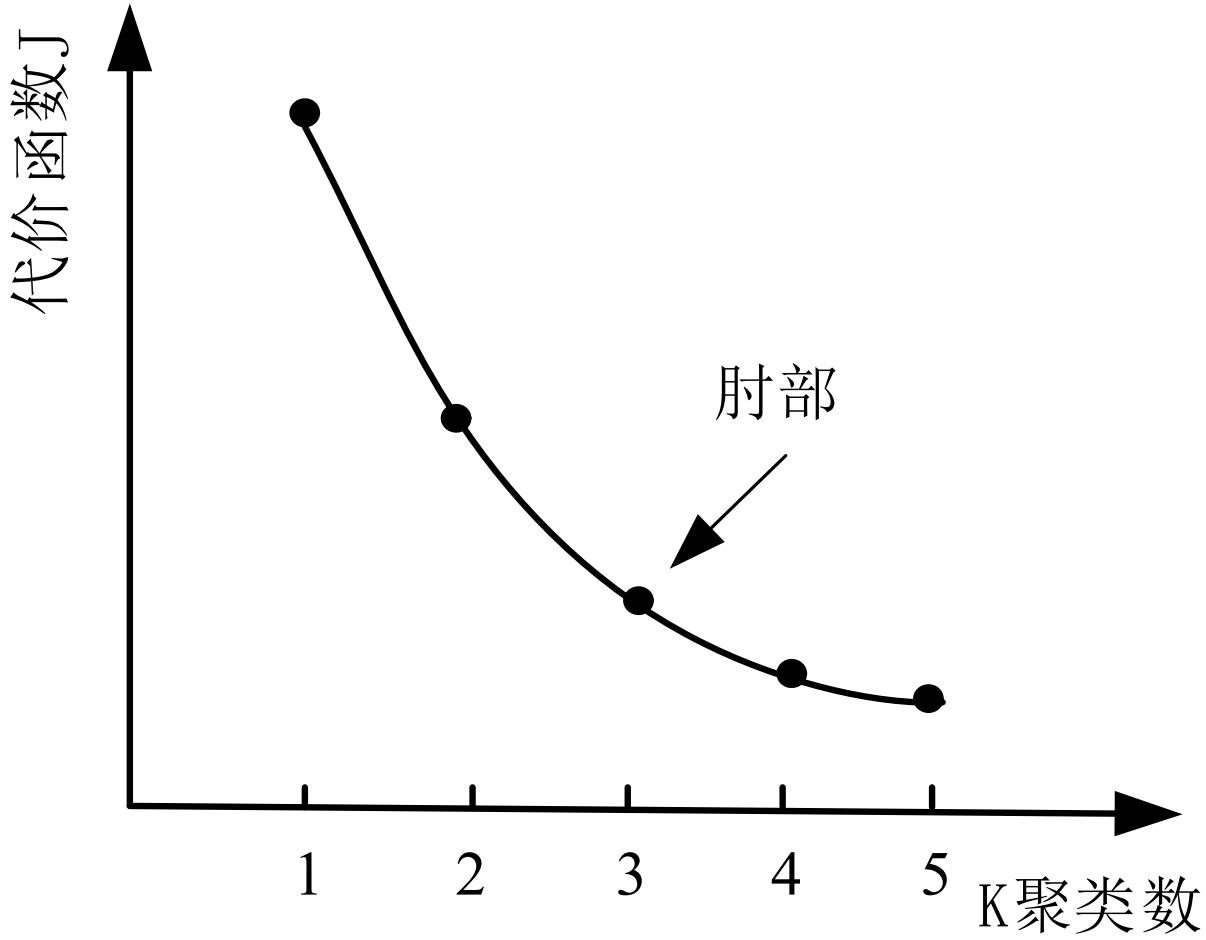
选择聚类数

通常需要根据不同的问题，人工进行选择。

考虑运用**K**均值算法聚类的动机，选择能最好实现该目的的聚类数。

K-均值算法

选择K值



目录

0. 介绍
1. 航天器零部件货运装箱背景
2. 聚类理论
3. K均值算法
4. 问题示例
5. 总结与作业
6. 知识扩展

航天器零部件包装标准化问题示例

？

航天器零部件包装标准化问题示例

仅讨论零部件的长、宽、高三个属性,通过聚类操作，选取3类包装箱外形尺寸.

采用K均值算法可将相似度较高的对象归类至同一簇中。其中，K代表簇的个数。K取值为3：

（1）首次分配聚类中心数据点

（2）更新数据点分配

样本编号	长/mm	宽/mm	厚/mm
1	163	57	2
2	314	291	287
3	281	145	90
4	246	149	26
5	137	118	101
6	201	36	26
7	157	26	4
8	454	428	285
9	432	97	62
10	430	117	104
…	…	…	…
49	377	149	90
50	430	415	236

聚类的Python实现

使用scikit-learn机器学习库可更为简单的实现K均值算法。

- 将KMeans类实例化，并设置好聚类中心数K

```
kmeans = KMeans( n_clusters = 3 )
```

- 对kmeans调用fit方法

```
kmeans.fit( dataset )
```

fit方法调用后，将为dataset中的每个数据样本分配一个簇标签，该标签可在kmeans.labels_属性中进行查看，如下：

```
print( “Cluster ID:\n{}”.format(kmeans.labels_))
```

- 调用predict方法为新数据点分配簇标签

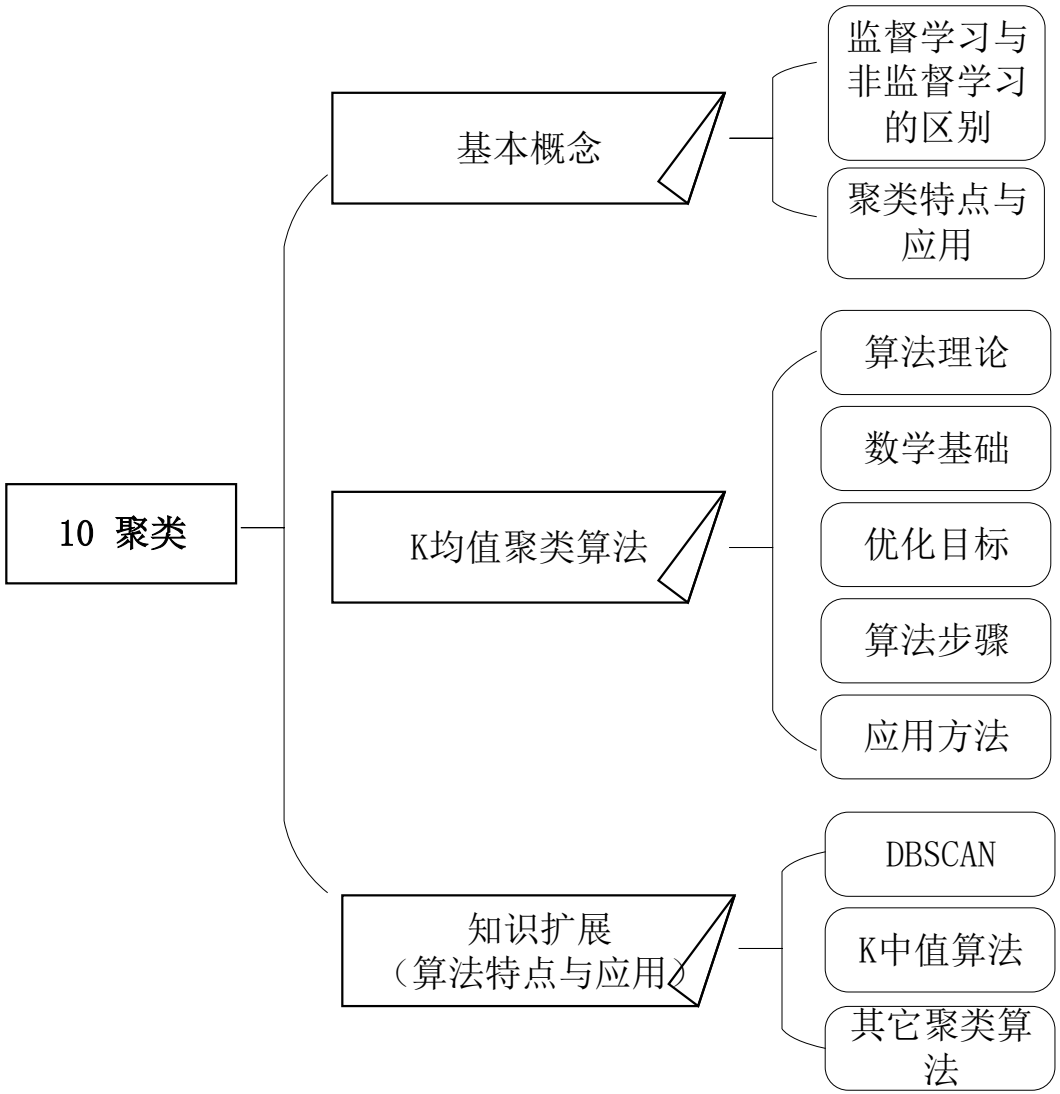
```
kmeans.predict(dataset)
```

目录

0. 介绍
1. 航天器零部件货运装箱背景
2. 聚类理论
3. K均值算法
4. 问题示例
5. 总结与作业
6. 知识扩展

总结与作业


人工智能
机器学习
深度学习


10 聚类
基本概念
监督学习与非监督学习的区别
聚类特点与应用
K均值聚类算法
算法理论
数学基础
优化目标
算法步骤
应用方法
知识扩展（算法特点与应用）
DBSCAN
K中值算法
其它聚类算法

总结与作业

1）采用Python语言编写K均值算法中的关键函数

①initCentroids随机初始化聚类中心函数

②minDistance计算数据集与各聚类中心距离，并给数据集中每个样本赋给一个簇标记

③getVar计算各个数据样本与聚类中心的距离累加和

2）根据K均值算法流程，将数据源k_means_data.txt中的数据，按据聚类数K=6，输出聚类中心点坐标和聚类后的图形

目录

0. 介绍
1. 航天器零部件货运装箱背景
2. 聚类理论
3. K均值算法
4. 问题示例
5. 总结与作业
6. 知识扩展

基于密度的聚类算法（DBSCAN）

算法思想：

只要样本点的密度大于阈值，就将该样本点加到最近的簇中。

优点：

- 不需确定簇的个数；
- 可找出不属于任何簇的点；
- 可以划分比较复杂的簇

基于密度的聚类算法（DBSCAN）

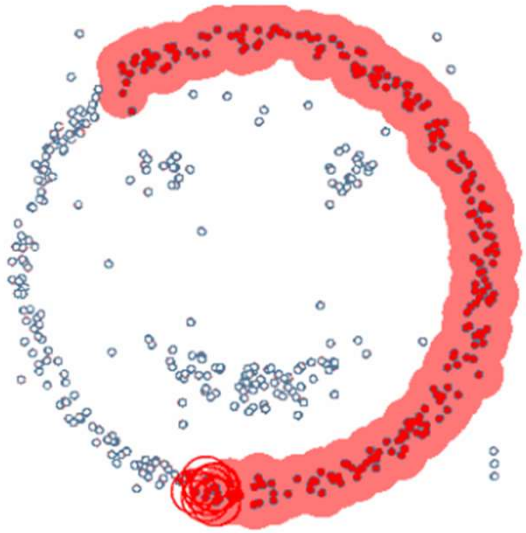
算法步骤：

①找出各样本的ε邻域，如果得到的邻域内点的数量小于阈值m，则这个点被标记为噪声；如果达到阈值，则此点成为核心对象，由此确定核心对象集合；

② 从核心对象集合中随机选取一个核心对象，找出由它密度可达的所有样本，并分配一个新的簇标签；

③ 将步骤②中找到的聚类簇中包含的核心对象从核心对象集合中去除；

④ 重复步骤②、步骤③直到核心对象集合为空。



其他聚类算法

- **K中值算法**
- 均值漂移算法