大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

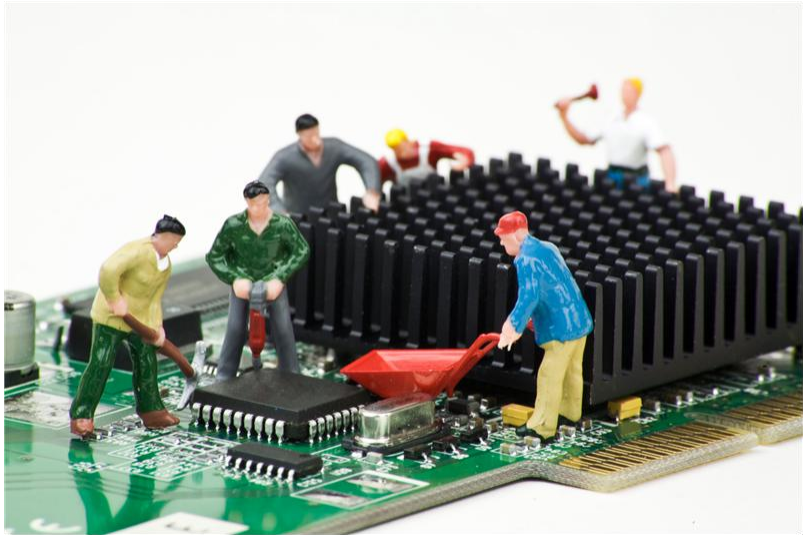
北京理工大学计算机学院　孙新

2019年1月

4、经典的机器学习方法

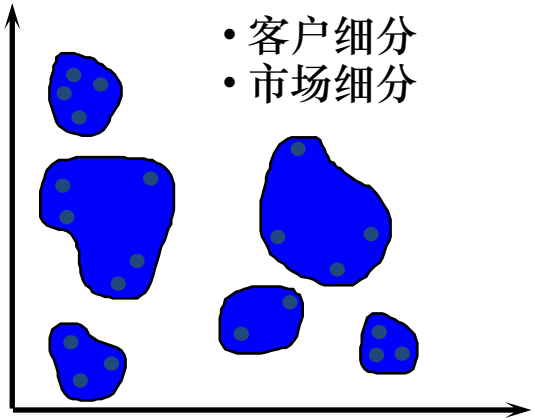
- 4.1 分类算法原理
- 4.2 决策树算法
- 4.3 K-近邻分类算法（KNN算法）
- 4.4 K-均值聚类算法（K-means算法）
- 4.5 Apriori关联规则算法

聚类是什么？



聚类(Clustering)

一个重要的非监督学习方法



聚类：将相似的对象组成多个类簇，以此来发现数据之间的关系

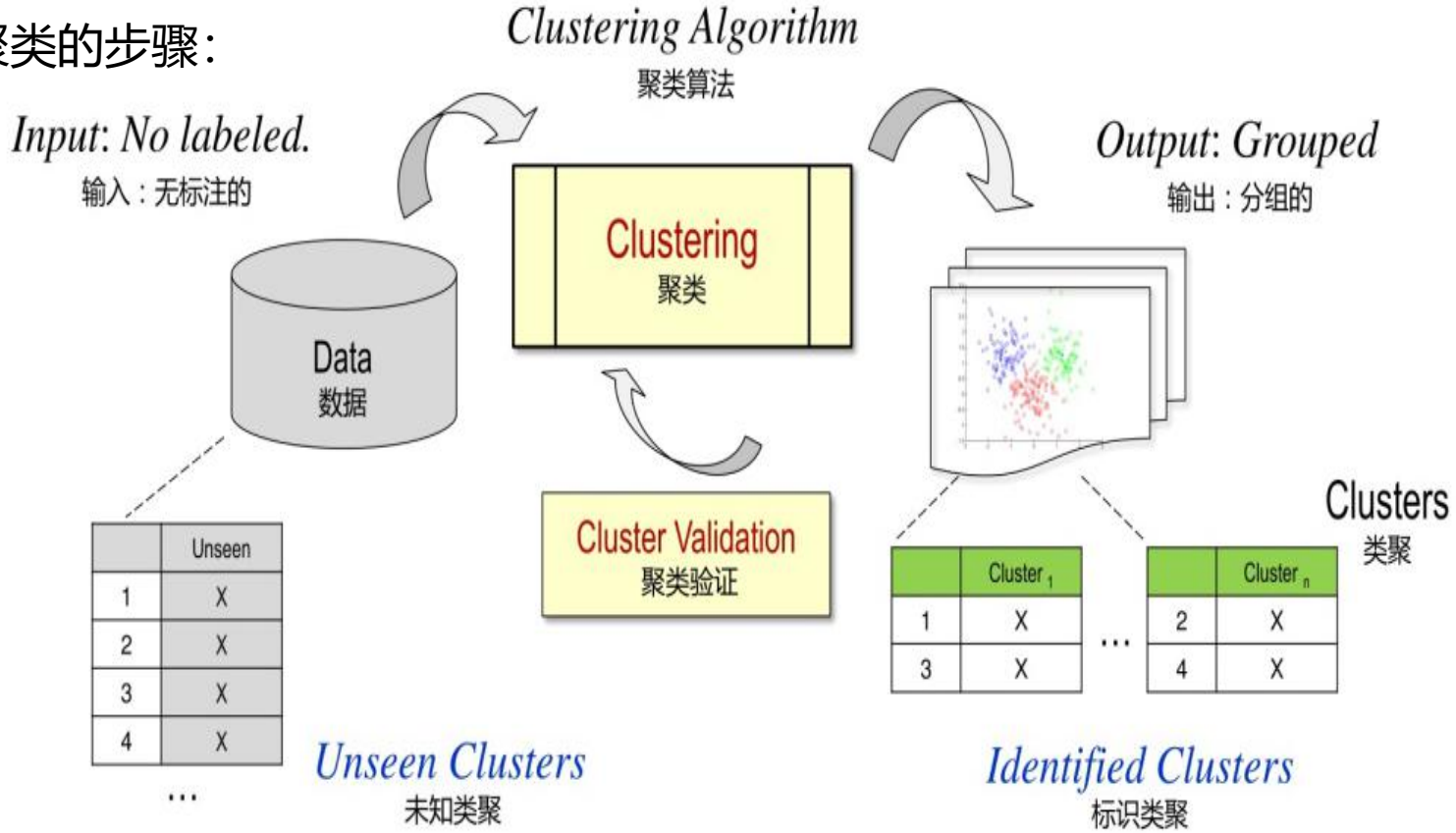
物以类聚，人以群分

聚类是什么？

objects	color(R_1)	shape(R_2)	size(R_3)
x_1	Red	Round	Small
x_2	Blue	Square	Large
x_3	Red	Triangular	Small
x_4	Blue	Triangular	Small
x_5	Yellow	Round	Small
x_6	Yellow	Square	Small
x_7	Red	Triangular	Large
x_8	Yellow	Triangular	Large

如何进行聚类？

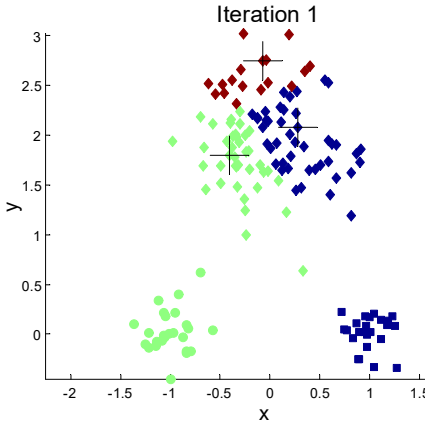
- 聚类的步骤：

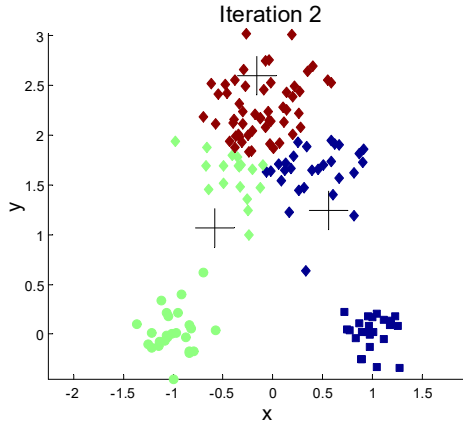


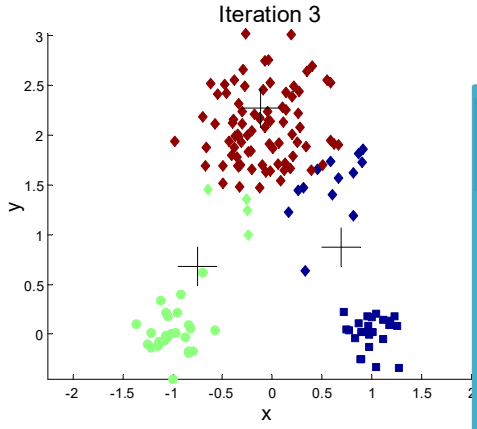
4.4 K-均值聚类算法（K-means算法）

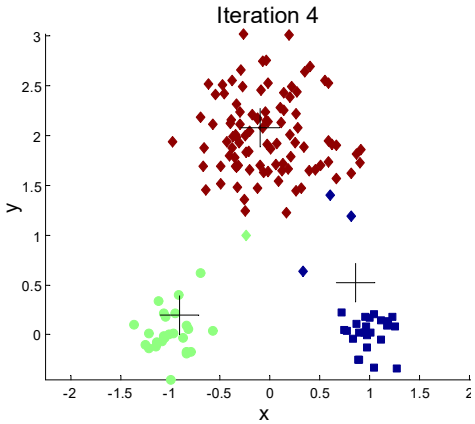
- 聚类是机器学习中非常重要的方法，是指将未标注的样本数据中相似的分为同一类。
- k-means是一种非监督方法，也被称为K-均值算法，它是被最广泛使用的、最为简单、高效的聚类算法。
- 核心思想：
  - 首先将各个聚类子集内的所有数据样本的均值作为该聚类的代表点，
  - 然后把每个数据点划分到最近的类别中，
  - 使得评价聚类性能的准则函数达到最优，从而使同一个类中的对象相似度较高，而不同类之间的对象的相似度较小。

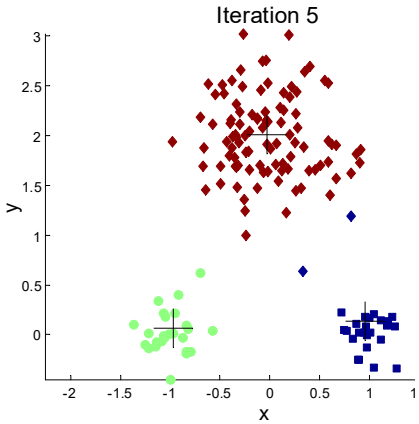
4.4 K-均值聚类算法（K-means算法）

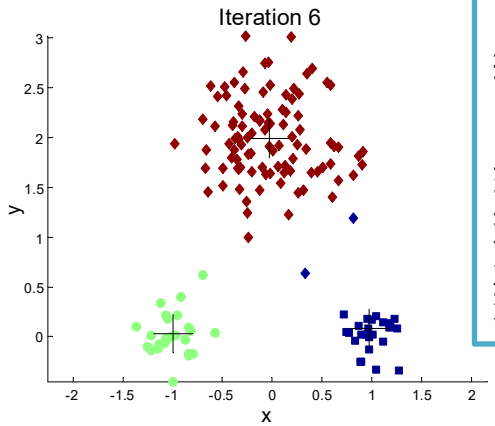












应用实例

利用K-means聚类算法，把原始数据聚成三个不同的簇的应用实例如左图示（K=3）。

基本思路：

（1）首先，随机选择k个数据点做为聚类中心；

（2）然后，计算其它点到这些聚类中心点的距离，通过对簇中距离平均值的计算，不断改变这些聚类中心的位置，直到这些聚类中心不再变化为止。

4.4 K-均值聚类算法（K-means算法）

k-means算法的基本步骤如下：

1 从数据集中随机取k个对象，作为k个簇的初始聚类中心。

2 计算剩下的对象到k个簇中心的相似度，将这些对象分别划分到**相似度最高**的簇。

3 根据聚类结果，更新k个**簇**的**中心**，计算方法是取簇中所有对象各自维度的算术平均数。

4 将数据集中全部元素按照新的中心重新聚类。

5 **达到算法停止条件，**转至步骤6；否则转至步骤3。

6 输出聚类结果。

4.4 K-均值聚类算法（K-means算法）

- 相似度计算：

  - 相似度计算方法比较多，欧几里得距离公式是常见的相似的计算公式

设讨论的对象集合为$U=\{\mathbf{x}_1,\cdots,\mathbf{x}_i,\cdots,\mathbf{x}_N\}$，聚类结果为 $CS=\{C_1,\cdots,C_k,\cdots,C_K\}$

其中，$\mathbf{x}=\{x^1,\cdots,x^d,\cdots,x^D\}$，$K$表示类的个数，$D$表示对象属性个数，

x^d 表示对象x 在第d维属性上的值。

论域U 中任意一个元素必须存在于至少一个类中。

对象 x_i 和 x_j之间的欧几里得距离记为 $d(x_i,\ x_j)$ 定义如下：

$$d(\mathbf{x}_i,\mathbf{x}_j)=\sqrt{\sum_{d=1}^{D}(x_i^d-x_j^d)^2}$$

K-means聚类算法

- 簇中心的计算方法：

聚类结果 $CS = \{C_1, \cdots, C_k, \cdots, C_K\}$ 中的类 C_k 的簇中心记为

$$\mathbf{v}_k = \{v^1, \cdots, v^d, \cdots, v^D\}$$

定义如下：

$$v^d = \frac{1}{|C_k|} \sum_{\mathbf{x}_i \in C_k} x_i^d$$

- 其中，$|C_k|$ 表示类 C_k 中对象的数目。

K-means聚类算法停止条件

k-means聚类算法的停止条件一般有以下几种：

1 设定迭代次数。

2 聚类中心不再变化 。

3 前后两次聚类结果的目标函数函数变化很小。

比如，定义误差的平方和（Sum of the Squared Error ,SSE）作为聚类质量的度量标准：$SSE = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mathbf{v}_k)^2$

设迭代次数为I，给定一个很小的正数δ，如果前后两次迭代结果$|SSE(I) - SSE(I+1)| < \delta$，算法结束；

否则，$I = I + 1$ 继续执行算法。