MM5425 商业分析

WEEK 8 LECTURE – CLUSTERING

WK 目录 Contents

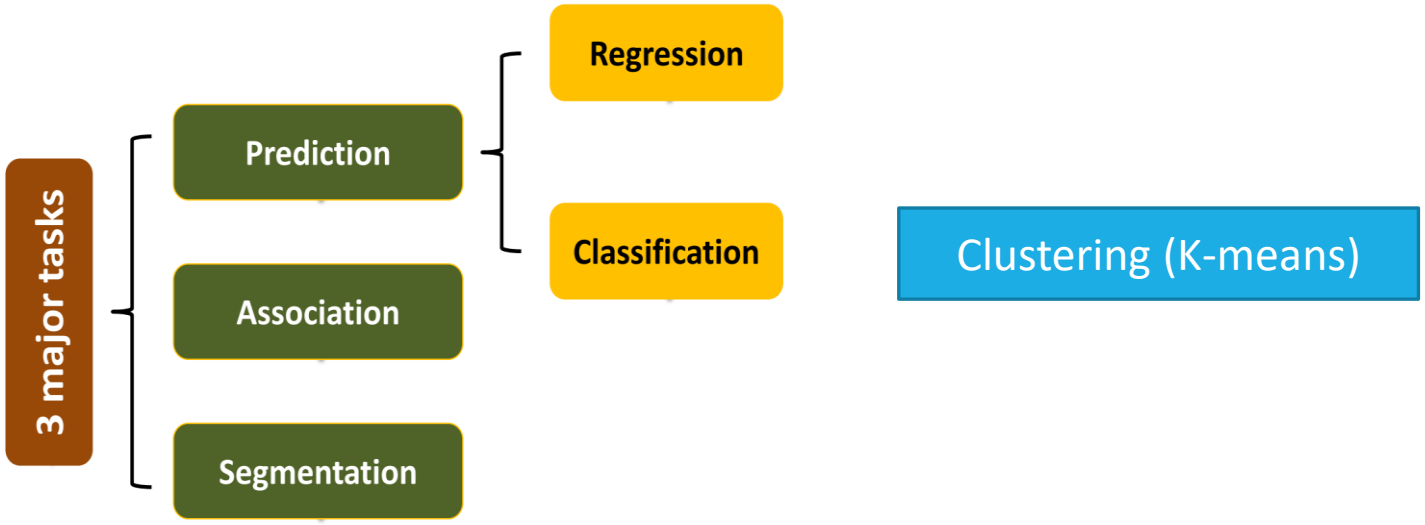
- K Nearest Neighbor (KNN) 分类模型

K 最近邻（KNN）是一种常用的分类算法，其基本思想是：对于一个新的样本点，找到训练集中距离它最近的 K 个样本，根据这 K 个邻居的类别，通过投票或加权投票的方式，决定新样本的类别

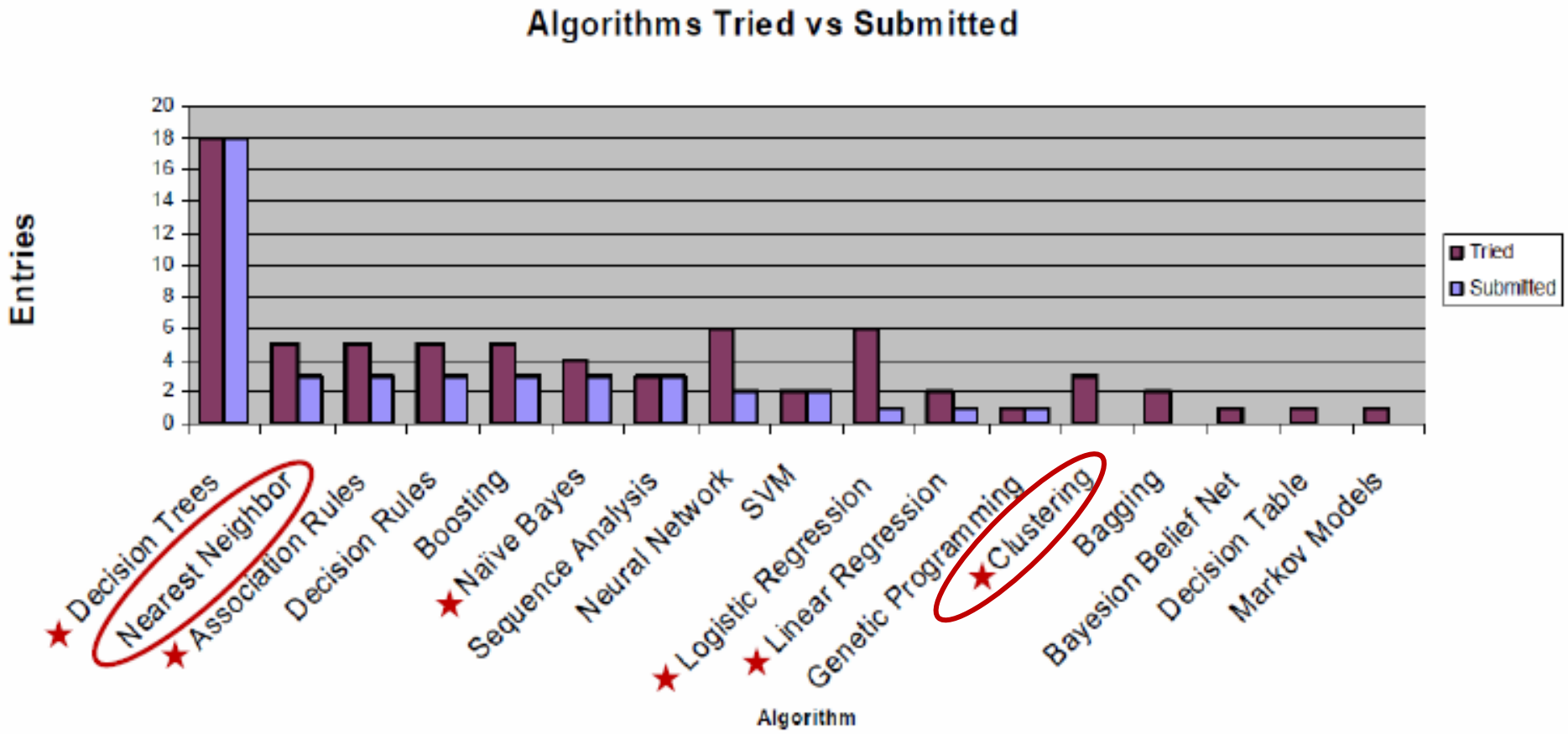
翻译如下：

聚类

- - 什么是聚类
- - 聚类中的相似性度量
- - K-means 算法
- - 模型评估



日常常用算法

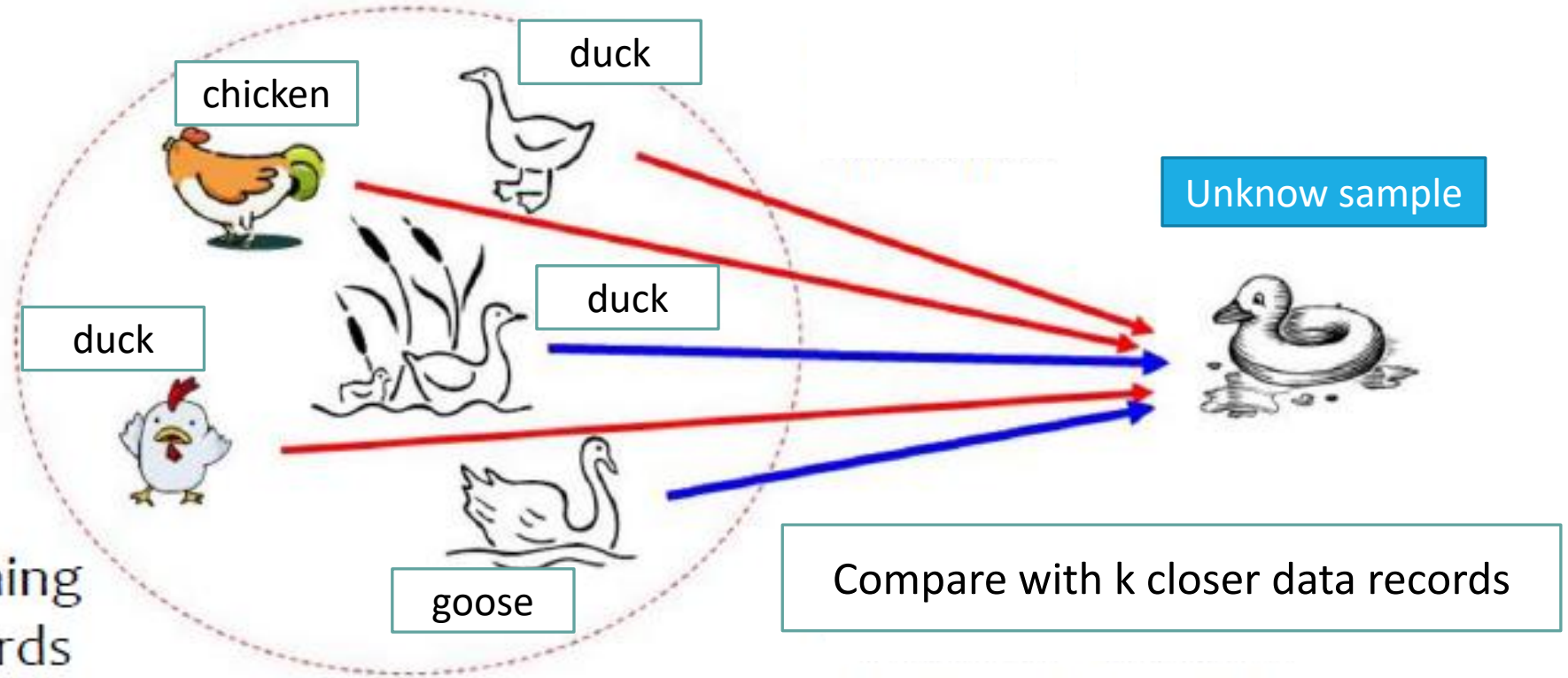

Algorithms Tried vs Submitted
Entries
20
18
16
14
12
10
8
6
4
2
0
Tried
Submitted
Decision Trees
Nearest Neighbor
Association Rules
Decision Rules
Boosting
Naïve Bayes
Sequence Analysis
Neural Network
SVM
Logistic Regression
Linear Regression
Genetic Programming
Clustering
Bagging
Bayesion Belief Net
Decision Table
Markov Models
Algorithm

K最近邻（简称KNN）

CLASSIFICATION/REGRESSION BASED ON SIMILARITY

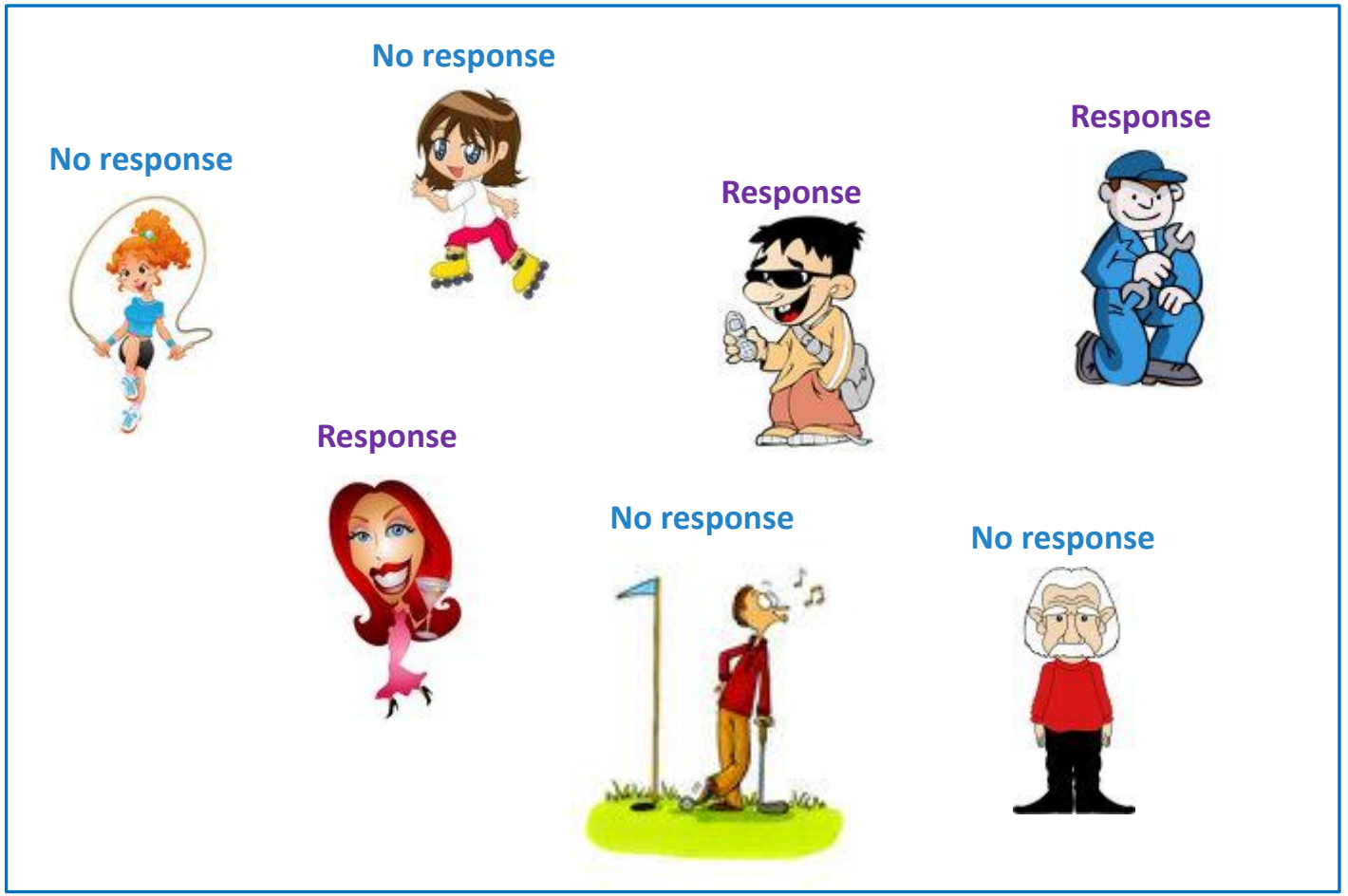
最近邻分类：基本思想

如果它走路像鸭子，叫声也像鸭子，那它很可能就是鸭子



例子：对促销活动的响应

新客户：是否响应???



我们所说的“相似”是什么意思?

我们用三个特征来描述一个客户：{年龄、年收入、每月购买次数}

	Age	Income	No. of purchase
Customer 1	18	150,000	10.5
Customer 2	23	250,000	8.2
Customer 3	49	700,000	2.5

哪一对客户更加相似?

距离度量：欧氏距离

欧氏距离 *dij* 是指在具有 k*k* 个属性的情况下，两个记录 i*i* 和 j*j* 之间的距离，其定义为：

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ik} - x_{jk})^2}$$

→ 客户1和客户3之间的欧氏距离：

$$d_{13} = \sqrt{(49 - 18)^2 + (700000 - 150000)^2 + (2.5 - 10.5)^2}$$

标准化 Standardization

- **欧氏距离对每个属性的尺度非常敏感。**

如果属性的取值范围差异很大，距离计算时，数值较大的属性会主导距离的结果，导致其他属性的影响被忽略。因此，在计算距离之前，必须对各个属性进行缩放（标准化或归一化），以防止距离度量被某一个属性主导，从而保证每个属性对距离的贡献是均衡的。

- **标准化 Standardization**
  - 将数据转换到统一的尺度（标准化）

$$X\_new = \frac{X - \mu}{\sigma}$$

	Age	Income	No. of purchase
Customer 1	18	150,000	10.5
Customer 2	23	250,000	8.2
Customer 3	49	700,000	2.5

Distance

	Age	Income	No. of purchase
Customer 1	18	150,000	10.5
Customer 2	23	250,000	8.2
Customer 3	49	700,000	2.5

	Age	Income	No. of purchase
Customer 1	-0.88305	-0.90575	1.020954
Customer 2	-0.51511	-0.48771	0.337014
Customer 3	1.398168	1.393466	-1.35797

$$d_{13} = \sqrt{(1.398168 - (-0.88305))^2 + (1.3934 - (-0.90575))^2 + (-1.35795 - 1.020954)^2}$$

K-近邻分类（KNN）

对一个给定样本进行分类的步骤如下：

1.计算距离

计算该样本与训练数据中所有样本之间的距离（如欧氏距离）。

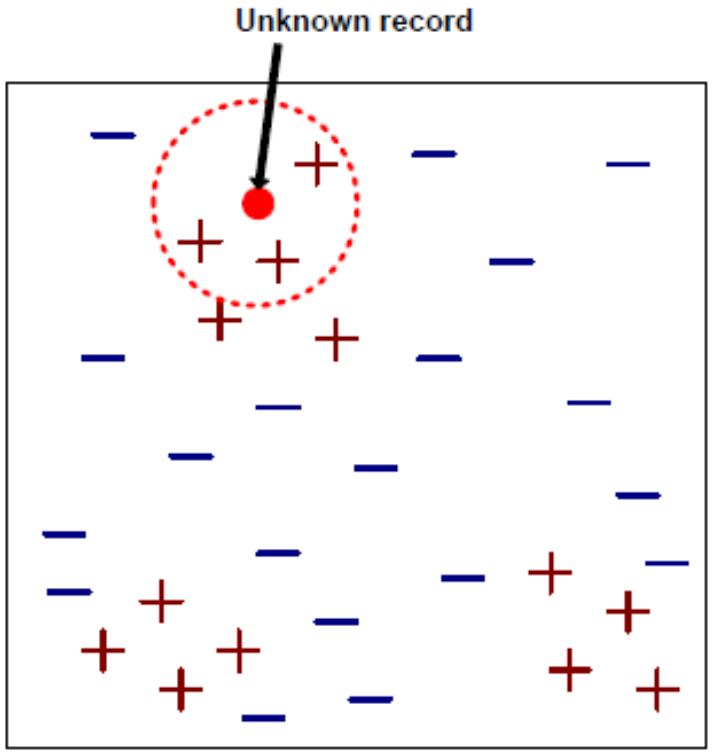
2.确定k个最近邻

找出距离最近的k个训练样本。

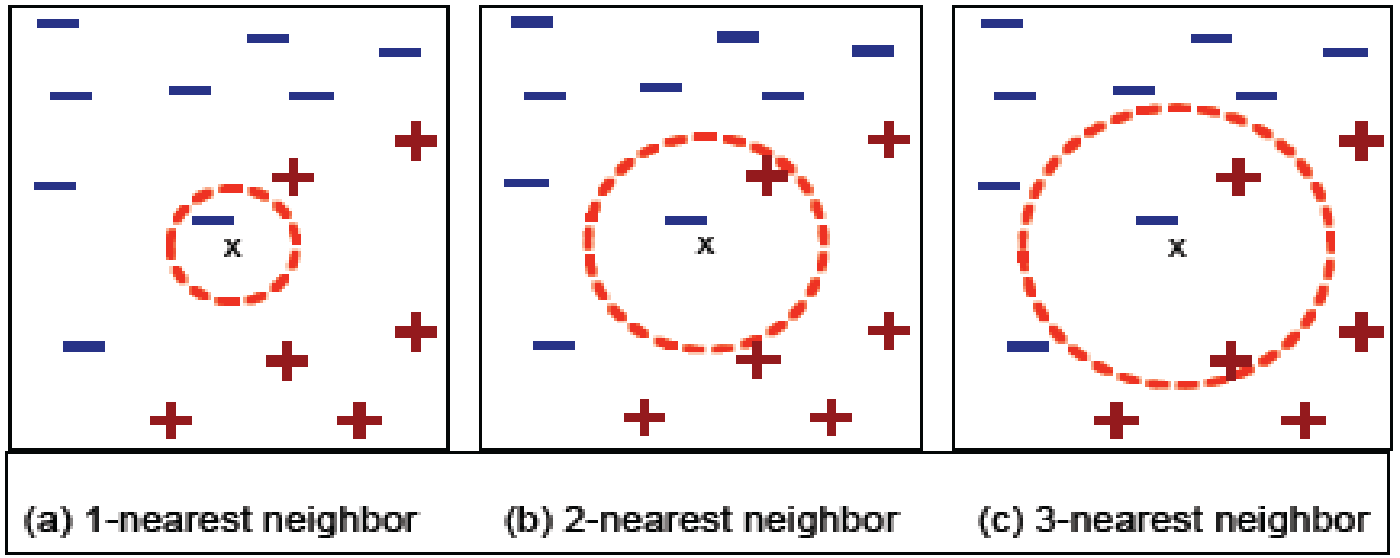
3.确定类别

根据这k个最近邻的类别标签，通过多数投票、比例或加权平均等方式，确定未知样本的类别标签。

这种方法可以有效地利用邻近样本的信息来进行分类，尤其适用于样本分布较为均匀的情况。



改变最近邻的K值



K-nearest neighbors of a record x are data points that have the k smallest distance to x

在KNN算法中，改变K的取值可能会改变样本的预测类别。

KNN算法中的两个主要问题：

1) 计算样本之间的距离

2) 如何选择K值——最近邻的数量？

如果K太小，对噪声点非常敏感；

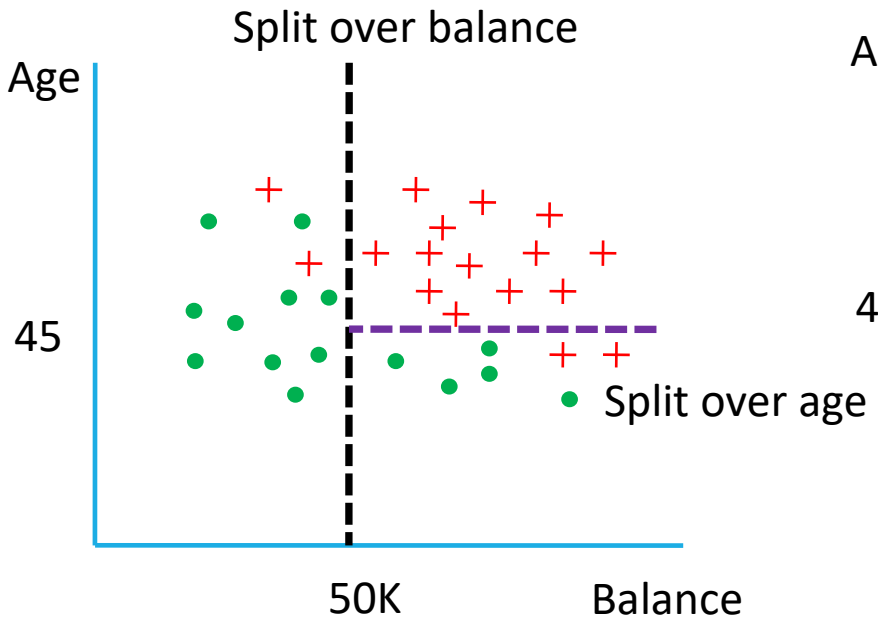
如果K太大，包含了太多可能无关的邻居。

可以考虑两个极端情况：

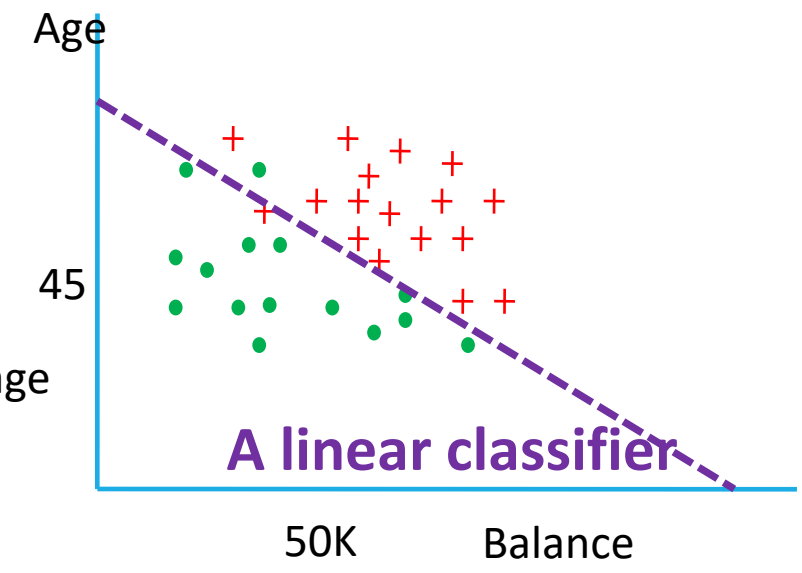
K=1 时，只考虑最近的一个邻居，容易受到异常值影响；

K=N 时，所有训练样本都被考虑在内，结果会倾向于数据中最多的类别。

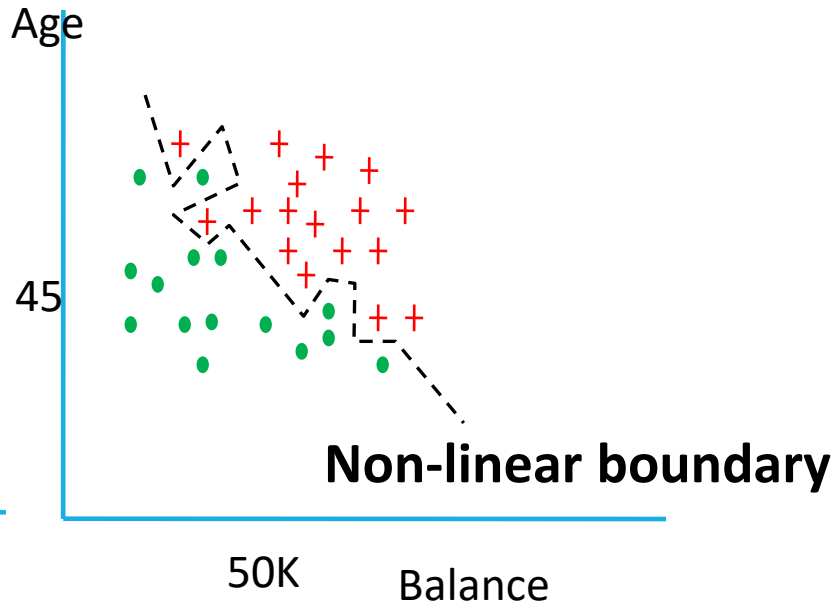
几何的解释 Geometric Interpretation



决策树分类器



逻辑回归分类器



K近邻分类器

KNN的优点/缺点

优点
- - 实现简单，易于使用
- - 可理解性强，预测结果容易解释
- - 通过对k个最近邻取平均，对噪声数据有一定鲁棒性（可控制过拟合）
- - 有一些很有吸引力的应用场景，如简单推荐、模式识别等

缺点
- - 对新样本分类时耗时较长
- - KNN不显式建立模型
- - 需要计算并比较新样本与所有训练样本的距离
- - 当样本数量很大时，计算代价极高

聚类 Clustering

DISCOVER THE HIDDEN PATTERN

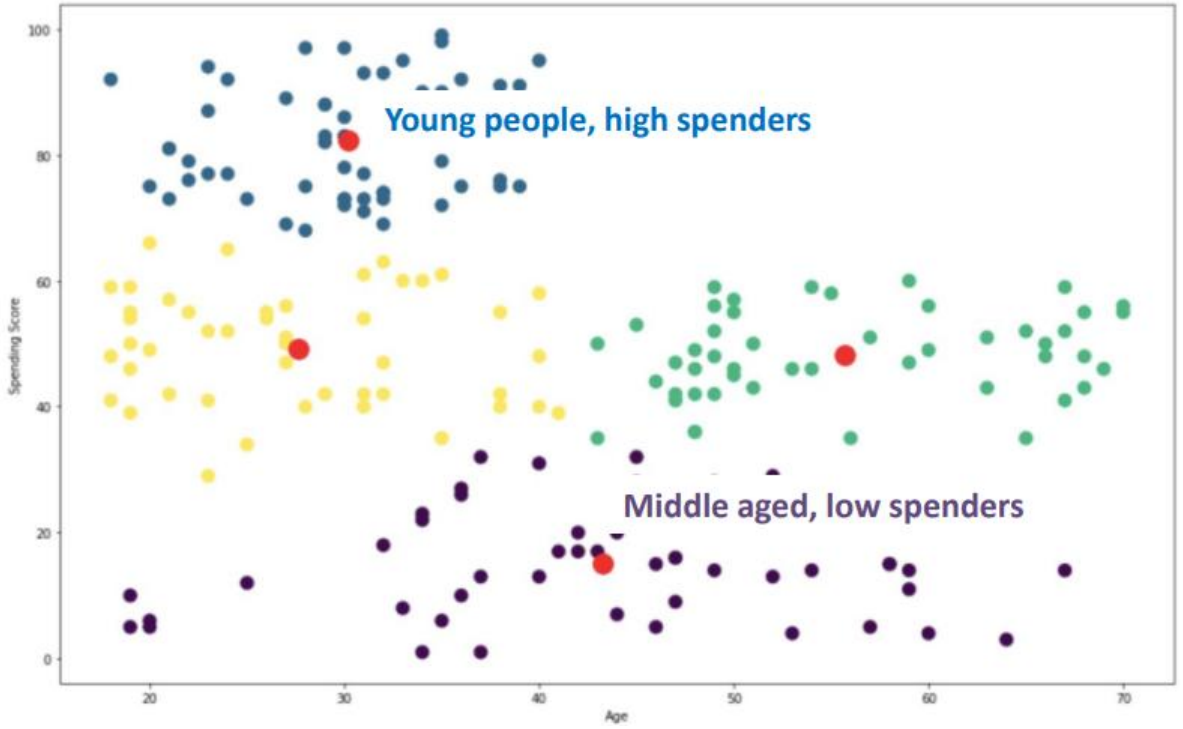
聚类：无监督技术

将数据对象分组，使得同一组（称为簇）中的对象在某种意义上彼此更为相似，而与其他组（簇）中的对象相对不那么相似。

什么是聚类?

将相似的数据对象归为一组

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40



商业案例 Business Examples

示例1：

将体型相似的人分为“小号”、“中号”和“大号”T恤组。

为每个人量身定制：成本太高

统一尺码：并不适合所有人

示例2：

在市场营销中，根据客户的相似性进行客户分群。

以便进行有针对性的营销

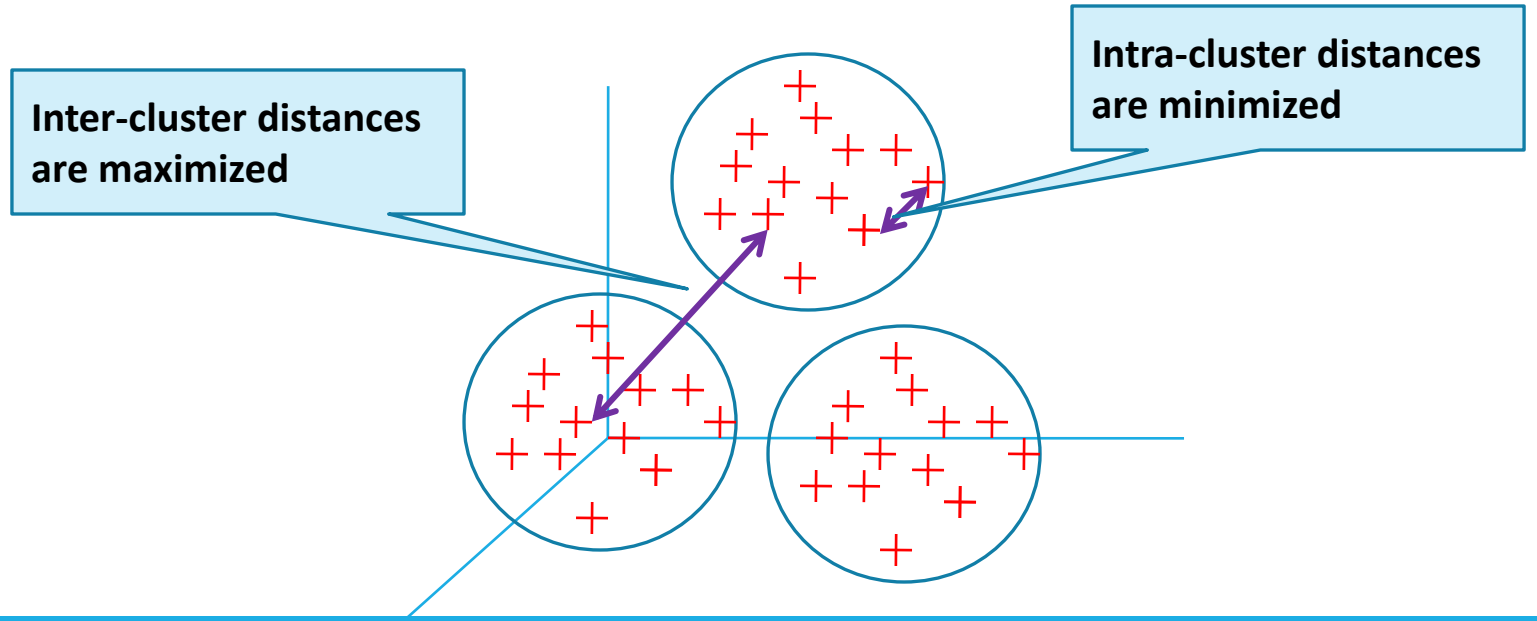
示例3：

给定一组文本文件，希望根据内容相似性进行组织和归类。

聚类: 主要思想 Main Idea

创建簇（clusters）的目标是：

- 最大化同一簇内记录之间的相似性
- 最大化不同簇之间记录的差异性

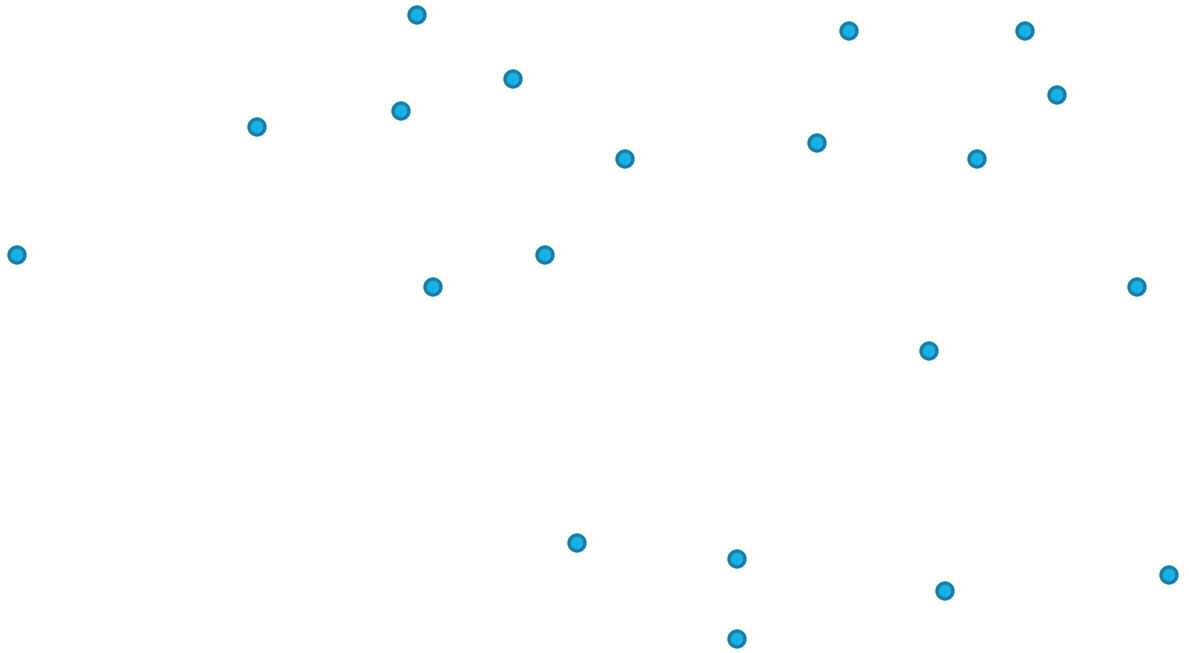


K均值聚类K-Means Clustering

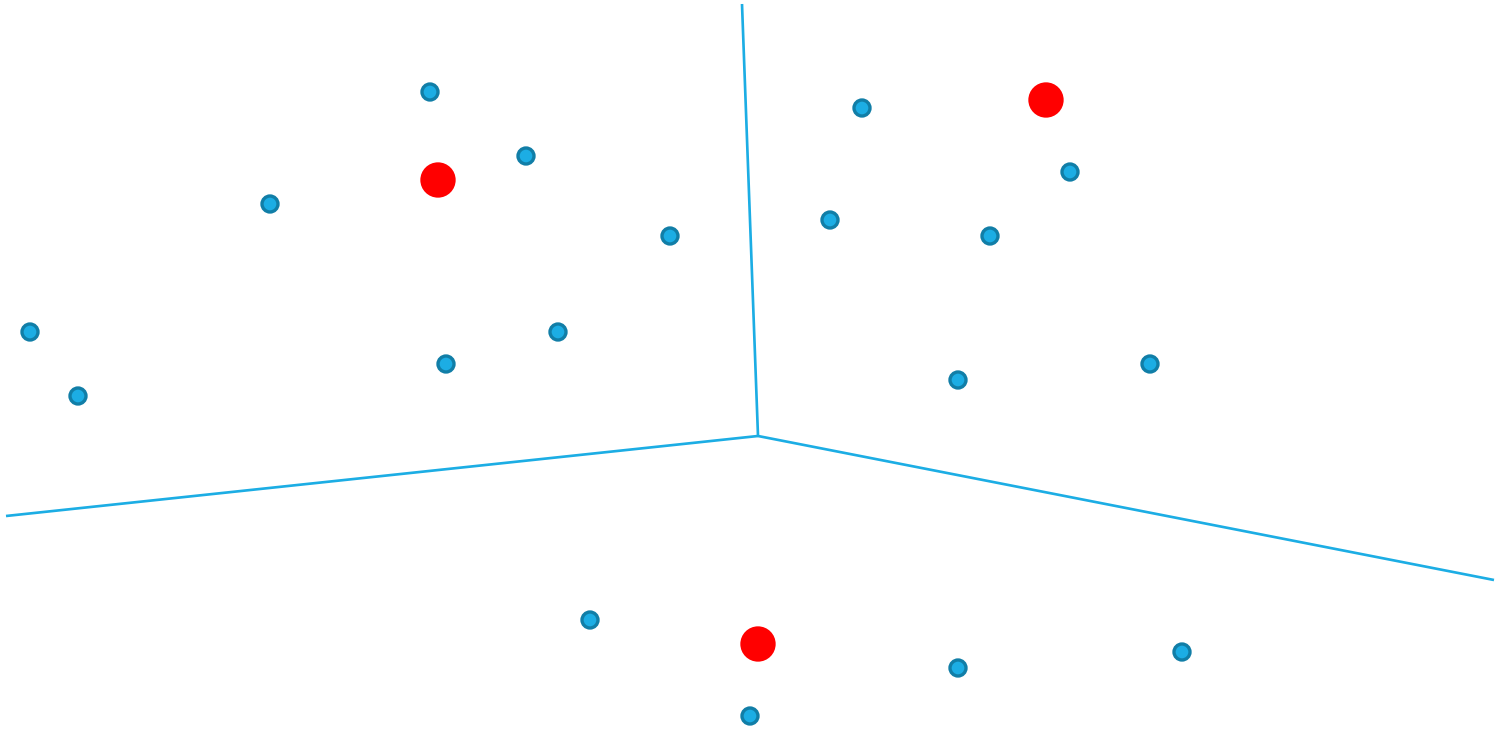
- 最常用且最简单的聚类方法：K均值（K-Means）
  - - 每个簇都对应一个质心（中心点）。
  - - 每个数据点被分配到距离其最近的质心所在的簇。
  - - 需要预先指定簇的数量k。

目标：最小化所有数据点到其所属k个中心的平方距离之和（SSD，Sum of Squared Distances）

一个例子: 3-Means (k=3)

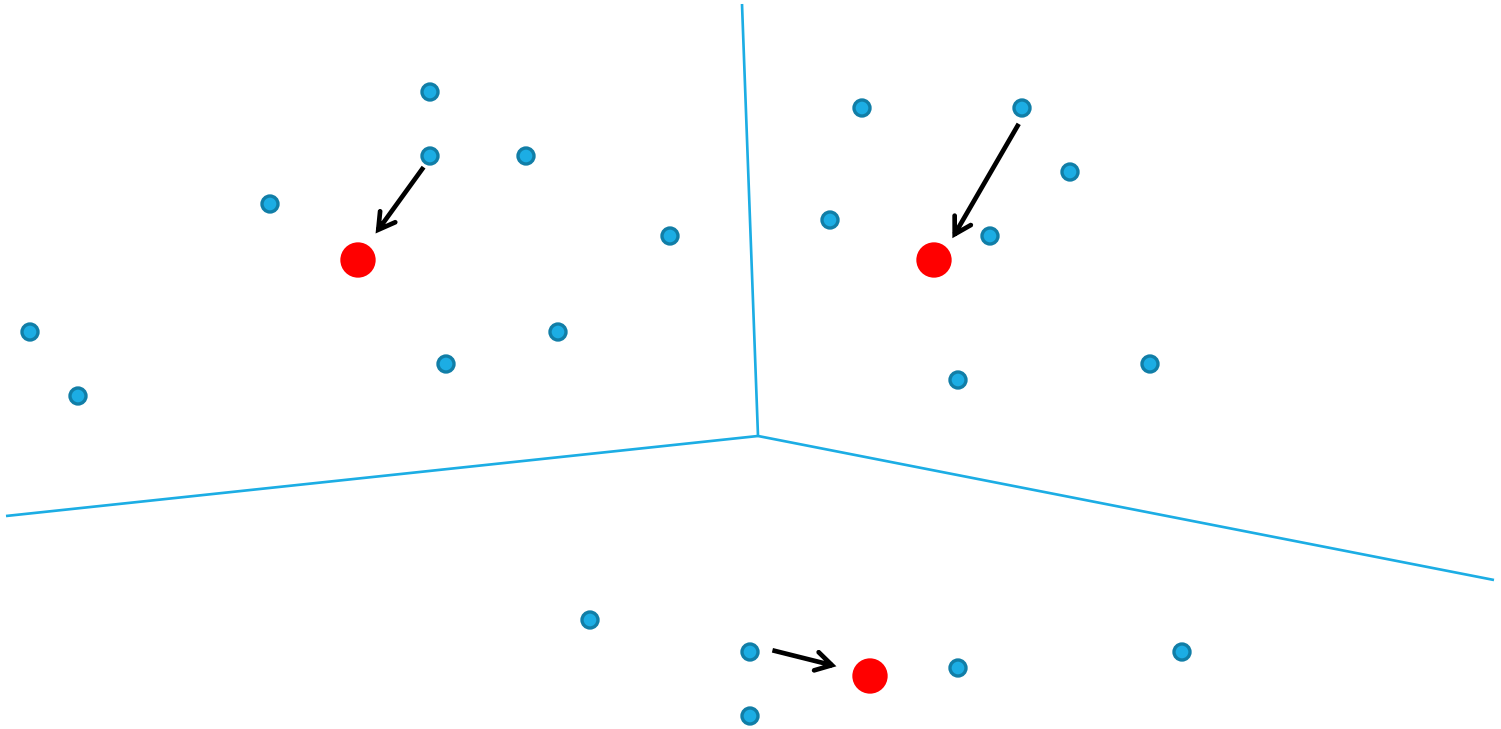


一个例子: 3-Means (*k*=3)



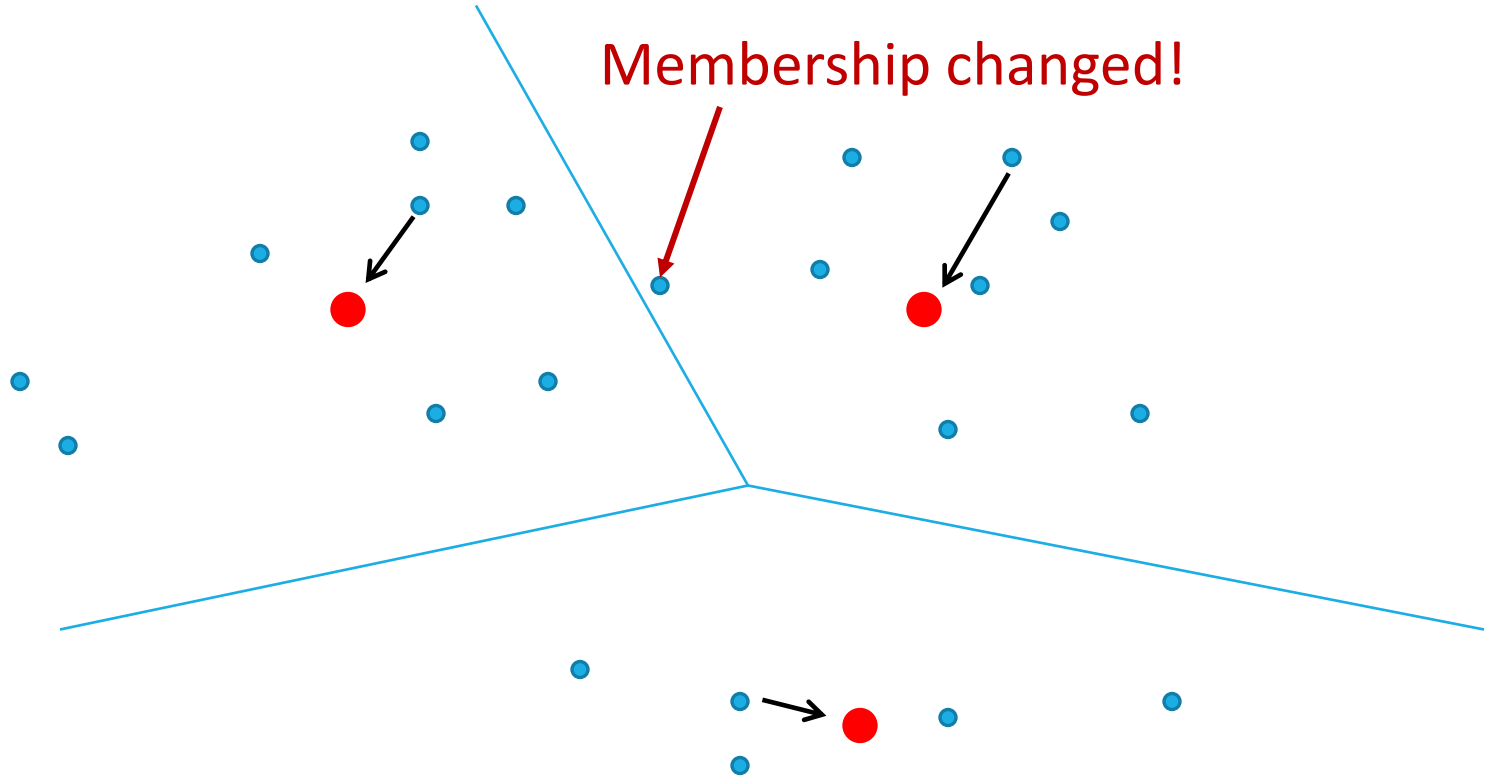
随机指定k个质心（中心点）
将每个样本分配到距离其最近的质心所在的簇

一个例子: 3-Means (*k*=3)



计算新的质心（注意：新的质心不一定是数据中的某个点，而是簇内所有点的均值）。

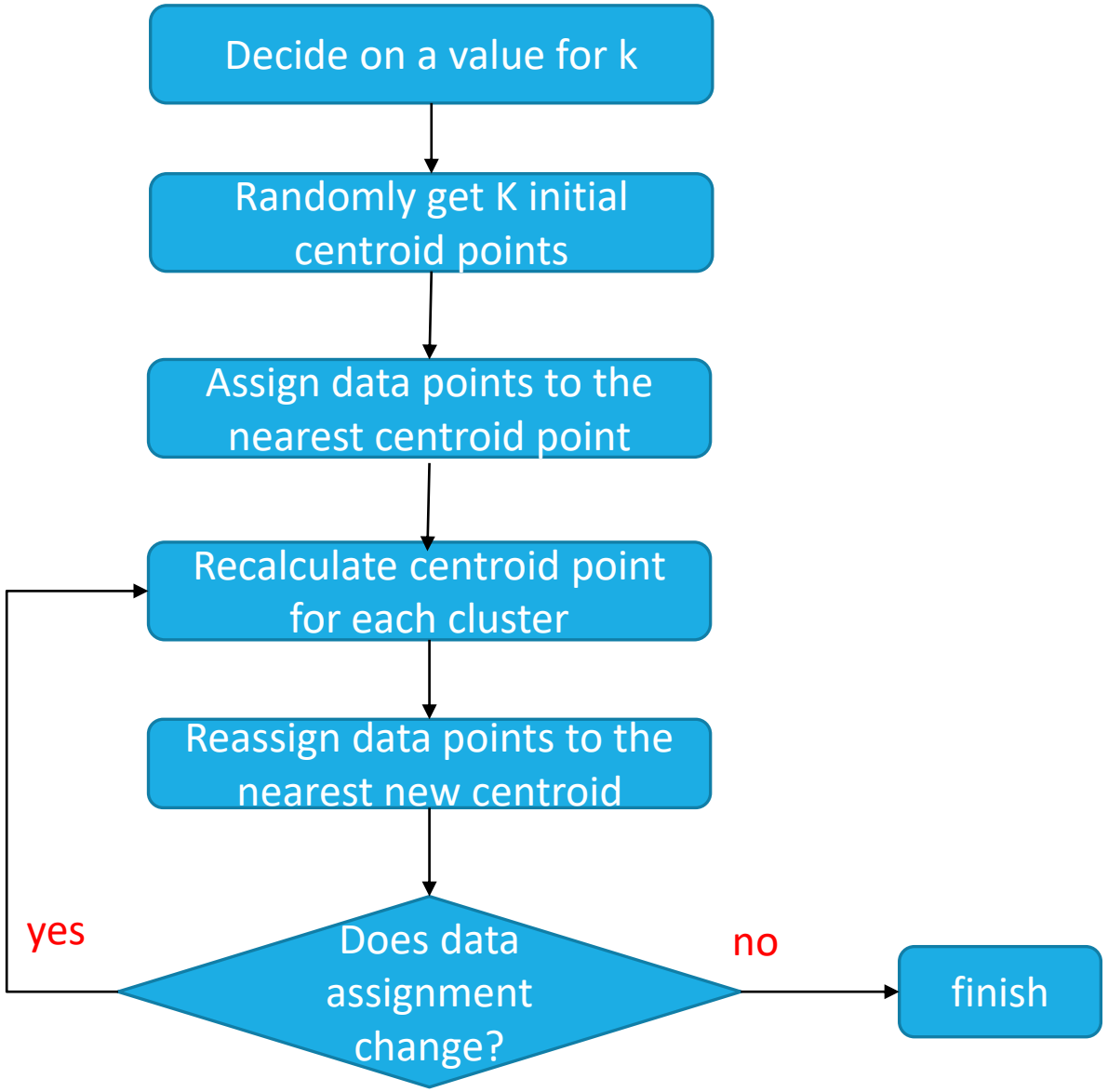
一个例子: 3-Means (*k*=3)



将每个样本分配到距离其最近的质心所在的簇

K-Means Algorithm

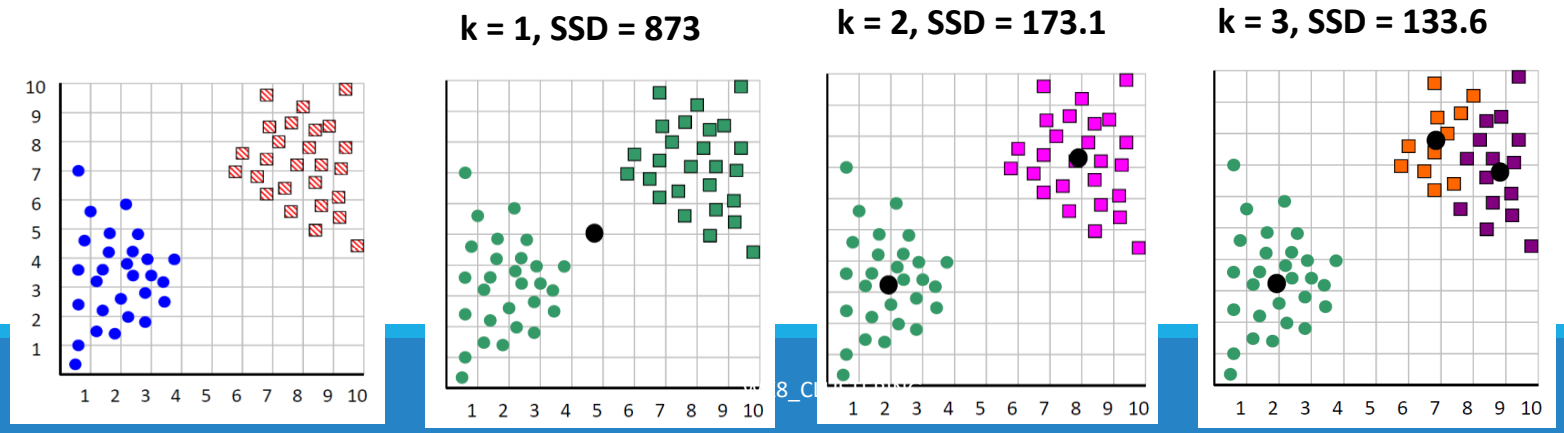
- 对初始中心的选择敏感
- 对噪声数据和异常值敏感
- 需要提前选择簇的数量k



如何选择簇的数量（k）？

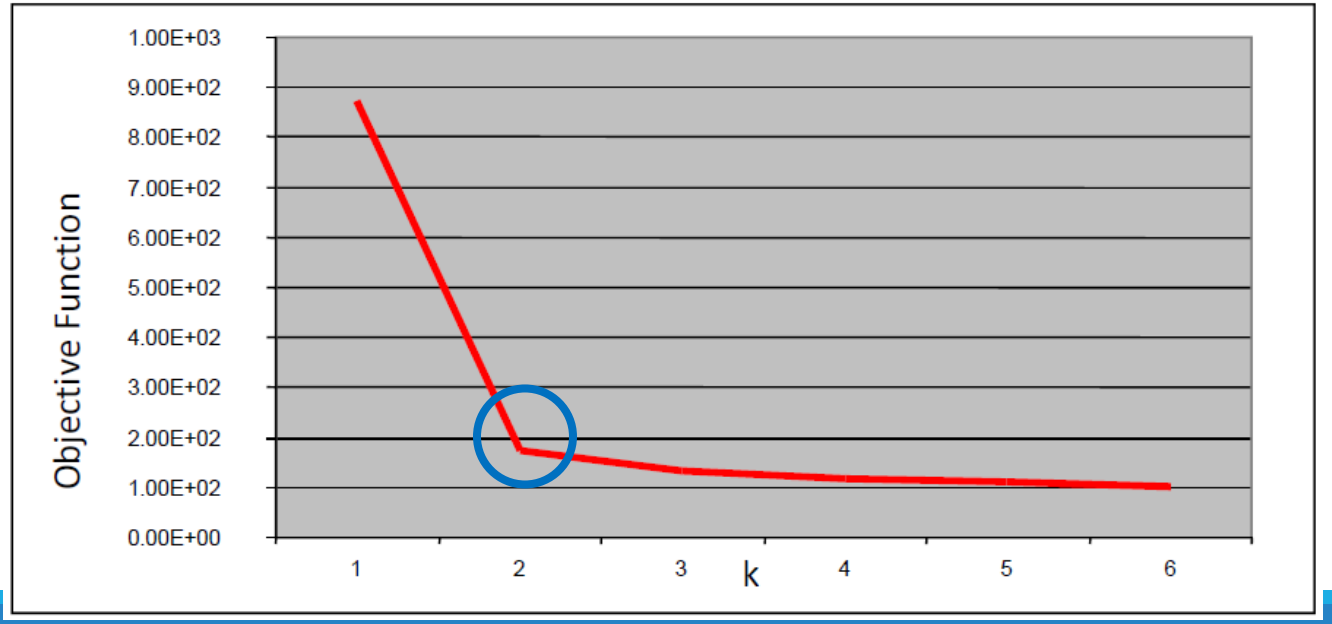
- 一般来说，选择最佳簇数（k）是一个尚未完全解决的问题。
  一种常用方法是：以平方距离和（SSD）作为目标函数，尝试不同的k值，选择最优的k。
- 具体做法如下：
  1. 对每个候选的k值，运行K均值算法，计算聚类后的SSD（所有点到各自质心的距离平方和）。
  2. 绘制k值与SSD的关系图，通常SSD会随着k的增加而减小。
  3. 选择SSD下降速度明显变缓的“肘部”位置作为最佳k值（即肘部法则）。

  这种方法可以帮助你在聚类效果和模型复杂度之间做出权衡，但并不能保证一定找到最优的k。实际应用时，常结合领域知识和其他评估指标综合判断。



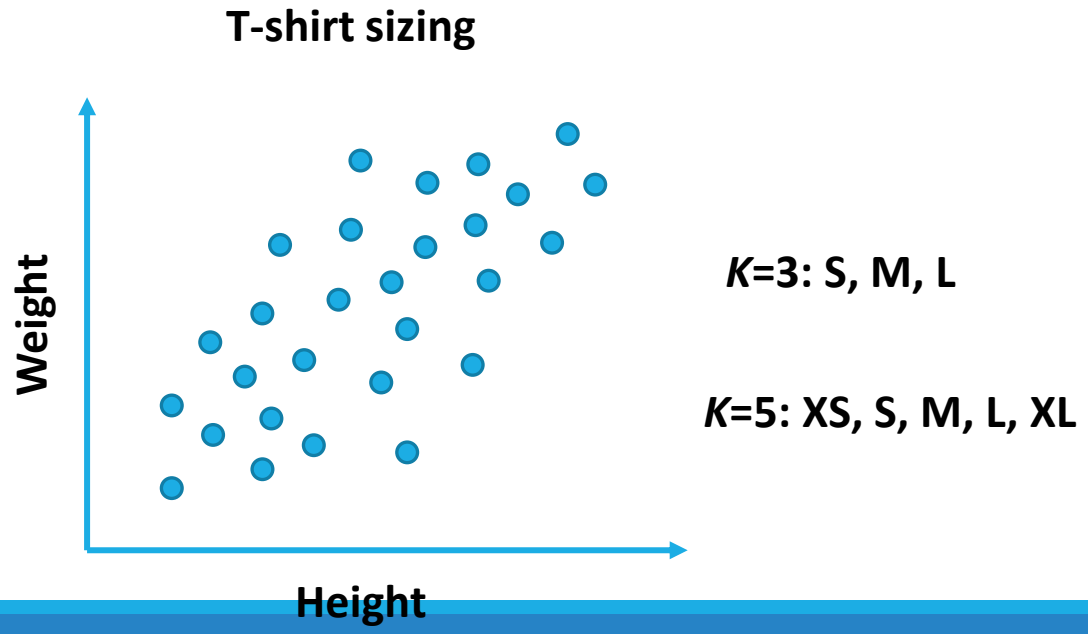
如何选择簇的数量（k）？——肘部法则（Elbow Method）

- 绘制不同k值下的SSD（平方距离和）曲线。
- 选择SSD出现大幅下降、之后下降变缓的位置对应的k值。
- 这种方法被称为“肘部法则”（elbow method）。



实际应用：如何选择K值

- 簇内成员应具有高度相似性
- 结合领域知识进行评估
- 根据K均值聚类在后续实际用途上的表现来评价聚类效果



Reference：
Business analytics: The science of data-driven decision making / U. Dinesh Kumar, New Delhi Wiley India, 2022