聚类分析

机器学习的两类问题

监督学习

分类问题

- 决策树、逻辑回归、支持向量机等

回归问题

- 线性回归

无监督学习

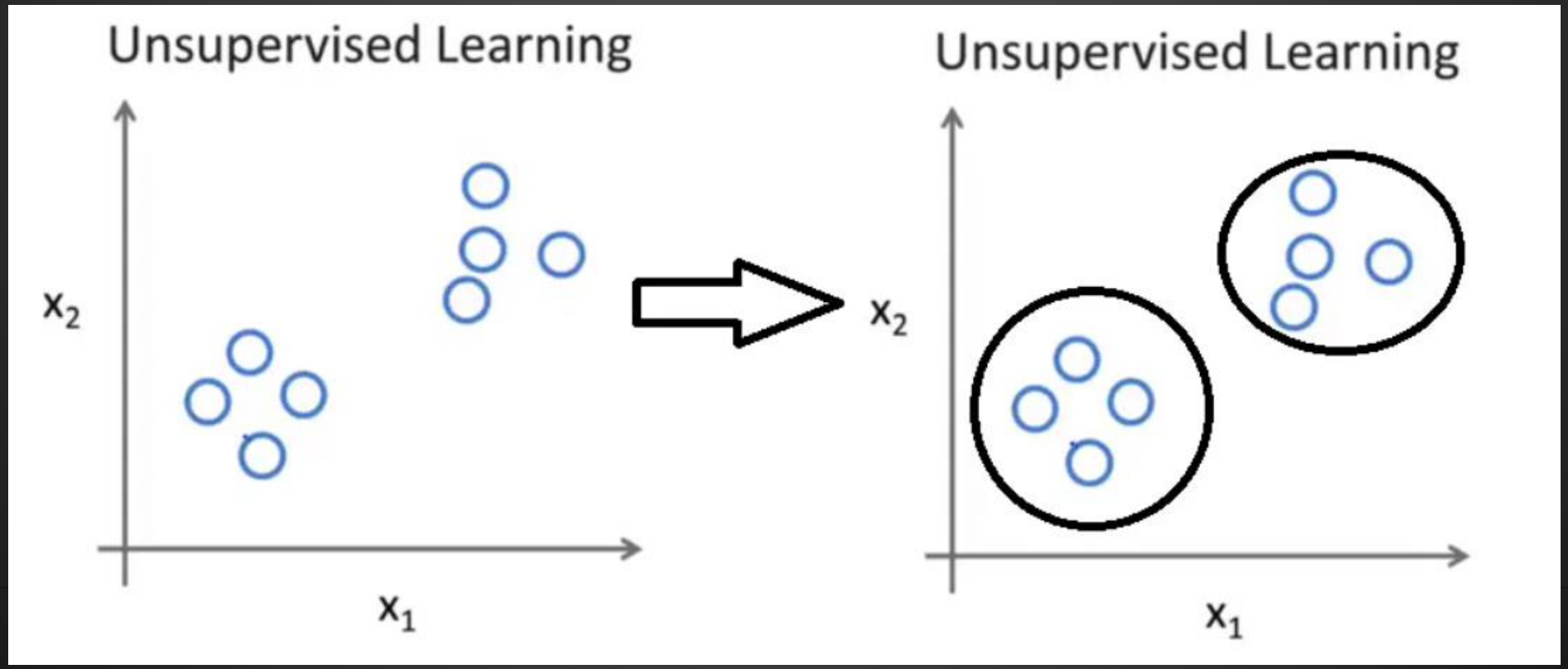
聚类分析

- 层次聚类法、k均值聚类、DBSCAN

关联分析

无监督学习

- 无监督学习，unsupervised learning，主要用于聚类、降维，例如K均值聚类法、主成分分析（PCA）等。
- 样本数据只有样本的特征（X），而没有类别（y）（标签）



聚类分析基本原理

- 例如，用户分群管理，已知的样本数据只有用户的特征，如首次消费时间、消费金额、消费频次等，需要根据用户的特征将用户分为不同的类别（类别事先未知）

用户编码	首次消费时间	最近一次消费时间	消费总金额	消费总次数
1	2016/6/25	2016/6/26	20000	3
2	2016/6/26	2016/6/26	50000	3
3	2016/6/13	2016/6/19	108000	4
4	2016/6/16	2016/6/16	30000	3
5	2016/6/27	2016/6/27	100000	3
6	2016/6/22	2016/6/22	15000	3

聚类分析基本原理

- 再比如，下表是同一批客户对经常光顾的五座商厦在购物环境和服务质量两方面的平均评分。现希望根据这批数据将五座商厦分类。

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

层次聚类法

计算步骤：

1、开始每个样本自成一类，有几个样本就有几个类。

2、计算各个类之间的距离，每次合并距离最近的两个类。

3、重复步骤2，直至所有样本都合并为一个类。

距离计算方式：欧式距离

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

案例：商厦分类

- 下表是同一批客户对经常光顾的五座商厦在购物环境和服务质量两方面的平均评分。现希望根据这批数据将五座商厦分类。

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

层次聚类法计算过程

例如，计算样本A和B之间的距离

$$D(A,B)=\sqrt{(73-66)^2+(68-64)^2}=8.06$$

按照类似的方式，可以将所有样本之间的距离都计算出来。

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

$$D_0=\begin{bmatrix} & A & B & C & D & E \\ A & 0 & & & & \\ B & 8.06 & 0 & & & \\ C & 17.80 & 25.46 & 0 & & \\ D & 26.91 & 34.66 & 9.22 & 0 & \\ E & 30.41 & 38.21 & 12.81 & 3.61 & 0 \end{bmatrix}$$

D,E之间的距离最小，因此将D，E合并为一个新类，记为CL4，CL4={D,E}

层次聚类法计算过程

接着计算各类之间的距离矩阵

$$D_1 = \begin{bmatrix} & A & B & C & CL4=\{D,E\} \\ A & 0 & & & \\ B & 8.06 & 0 & & \\ C & 17.80 & 25.46 & 0 & \\ CL4=\{D,E\} & 26.91 & 34.66 & 9.22 & 0 \end{bmatrix}$$

A,B之间的距离最小，因为将A，B合并为一个新类，记为CL3，CL3={A,B}

层次聚类法计算过程

接着计算各类之间的距离矩阵

$$D_2 = \begin{bmatrix} & CL3 & C & CL4=\{D,E\} \\ CL3 & 0 & & \\ C & 17.80 & 0 & \\ CL4=\{D,E\} & 26.91 & 9.22 & 0 \end{bmatrix}$$

C,CL4之间的距离最小，因为将C，CL4合并为一个新类，记为CL2，CL2={C,CL4}

层次聚类法计算过程

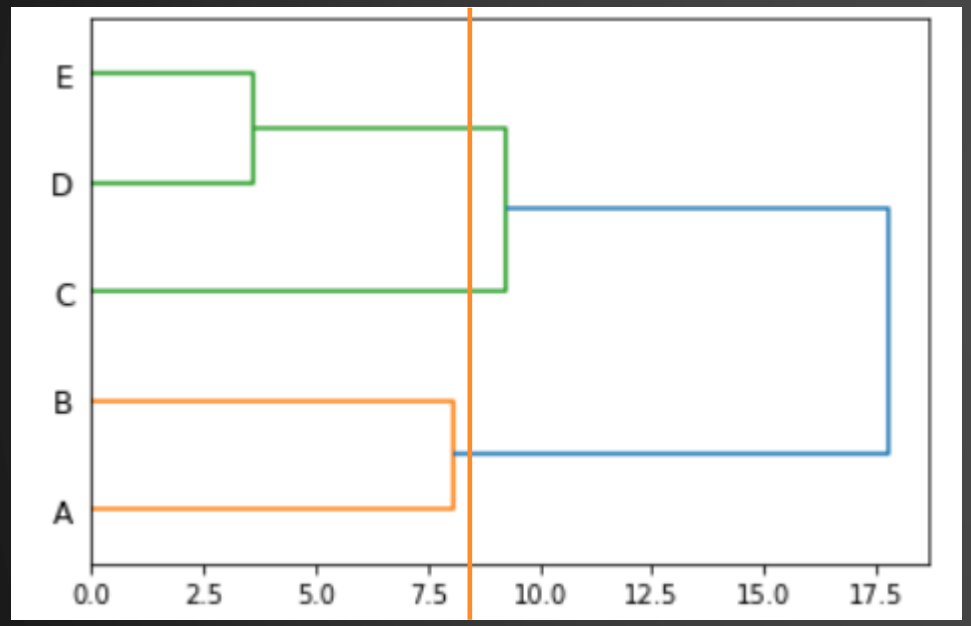
接着计算各类之间的距离矩阵

$$D_2 = \begin{bmatrix} & CL3 & CL2 \\ CL3 & 0 & \\ CL2 & 17.80 & 0 \end{bmatrix}$$

最后，将CL3和CL2合并为CL1。至此，并类完成。

层次聚类法

将上述聚类过程用一个图来表示，即谱系聚类图，如下图所示。



编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

层次聚类法：Python实现

- 数据：5座商厦的评分数据
- 实现层次聚类法的第一种方法：scipy
- 实现层次聚类法的第二种方法：sklearn

K均值聚类法

K均值聚类英文名称是k-means clustering，所以叫作K均值聚类法。

计算步骤

1、确定聚类类别数、初始聚类中心

2、通过计算每一个样本与聚类中心之间的距离，将样本归到距离最近的类中

3、重新计算每个类的聚类中心，重复这样的过程，直到每个样本都被归入类中

K均值聚类法计算过程

- 确定聚类类别数：假设分为3类
- 选取初始聚类中心：B、C、E

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

K均值聚类法计算过程

计算出其余样本与所有聚类中心之间的距离

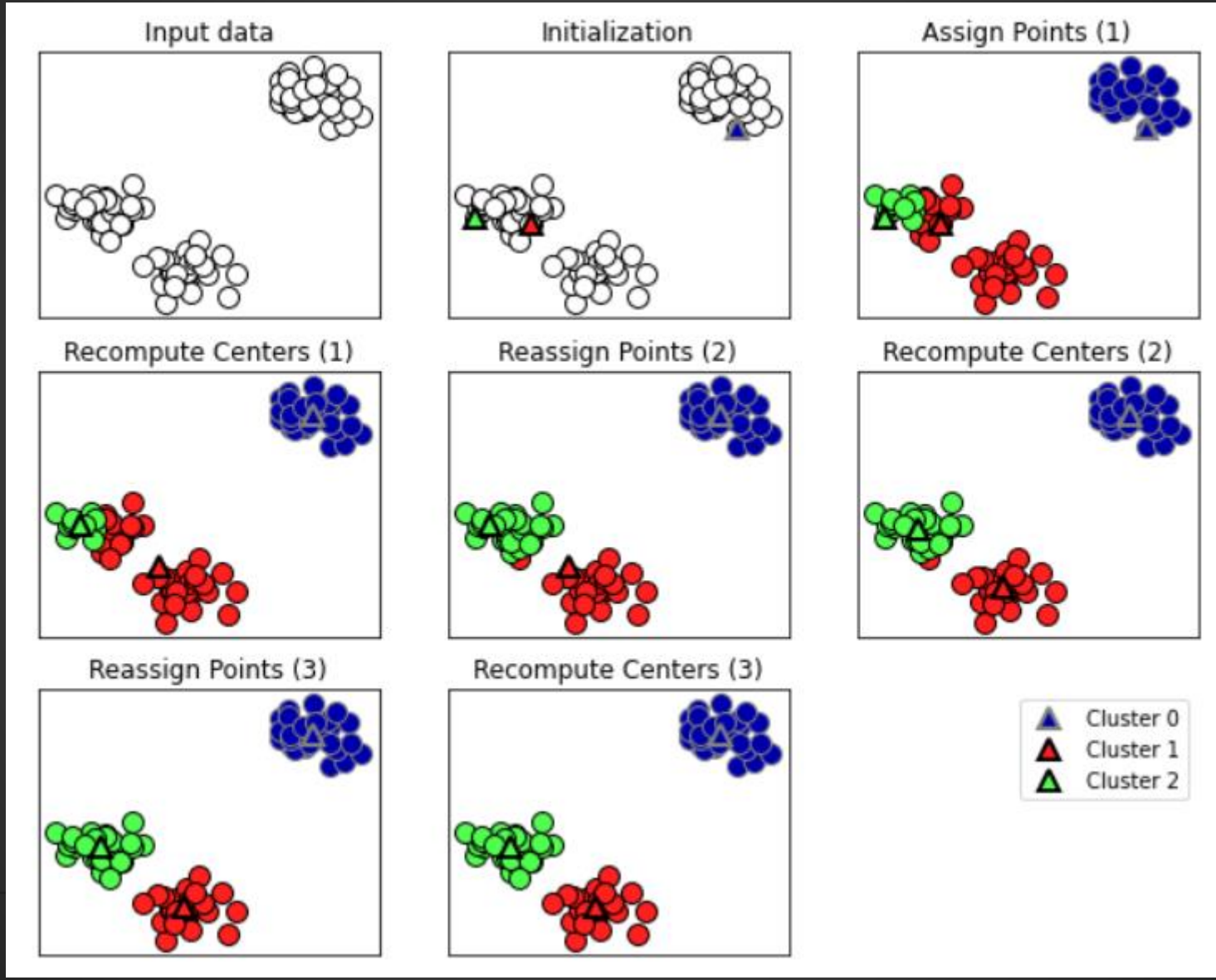
$$D_0 = \begin{bmatrix} & A & B & C & D & E \\ A & 0 & & & & \\ B & 8.06 & 0 & & & \\ C & 17.80 & 25.46 & 0 & & \\ D & 26.91 & 34.66 & 9.22 & 0 & \\ E & 30.41 & 38.21 & 12.81 & 3.61 & 0 \end{bmatrix}$$

编号	购物环境	服务质量
A	73	68
B	66	64
C	84	82
D	91	88
E	94	90

- 样本A：离B最近，将其归入样本B所在的类
- 样本D：离C最近，将其归入样本C所在的类
- 至此，得到三类，C1={A，B}，C2={C}，C3={D，E}

说明：上述K均值聚类法计算过程并不严格，只是为了让大家快速理解K均值聚类的计算过程！

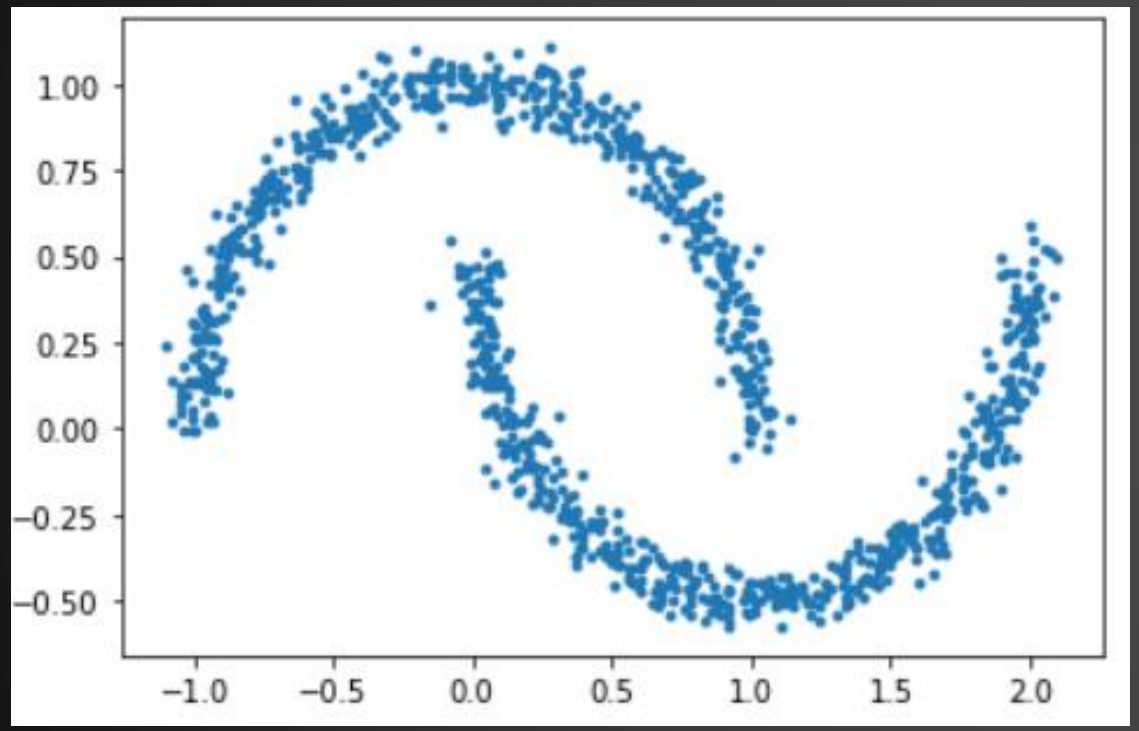
K均值聚类法流程示意图

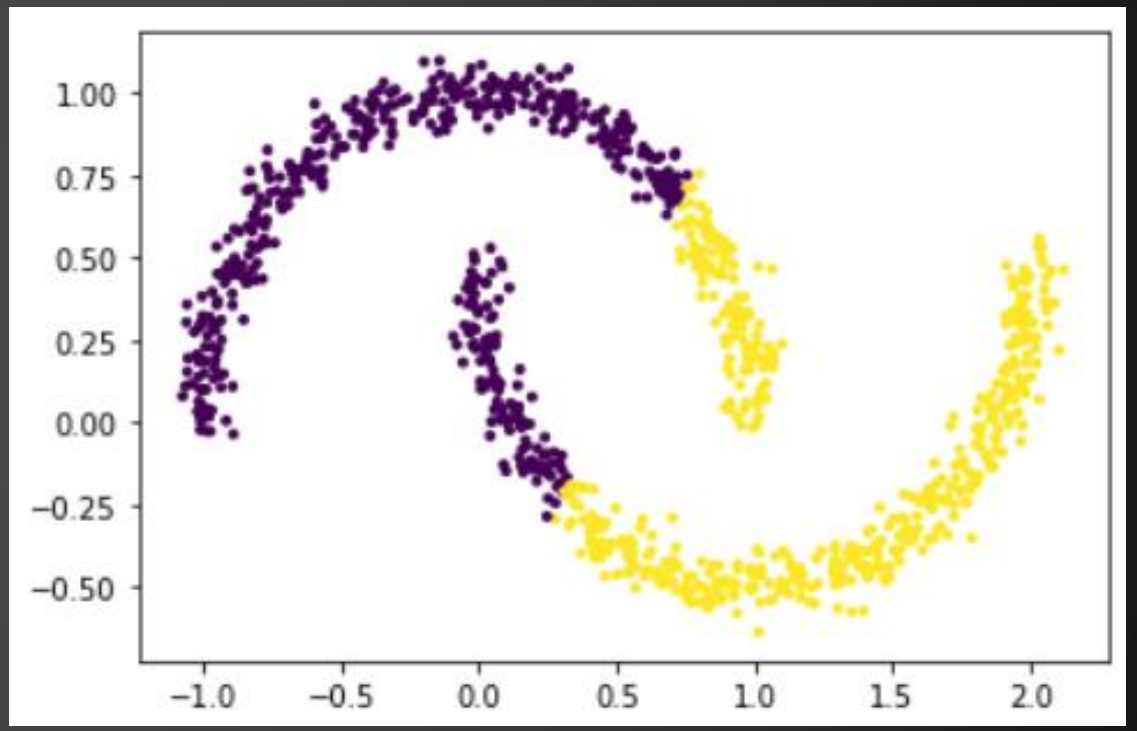


K均值聚类法：Python实现

- 数据：5座商厦的评分数据
- K均值聚类建模
- 获得样本的类别标签
- 将类别标签赋回原记录

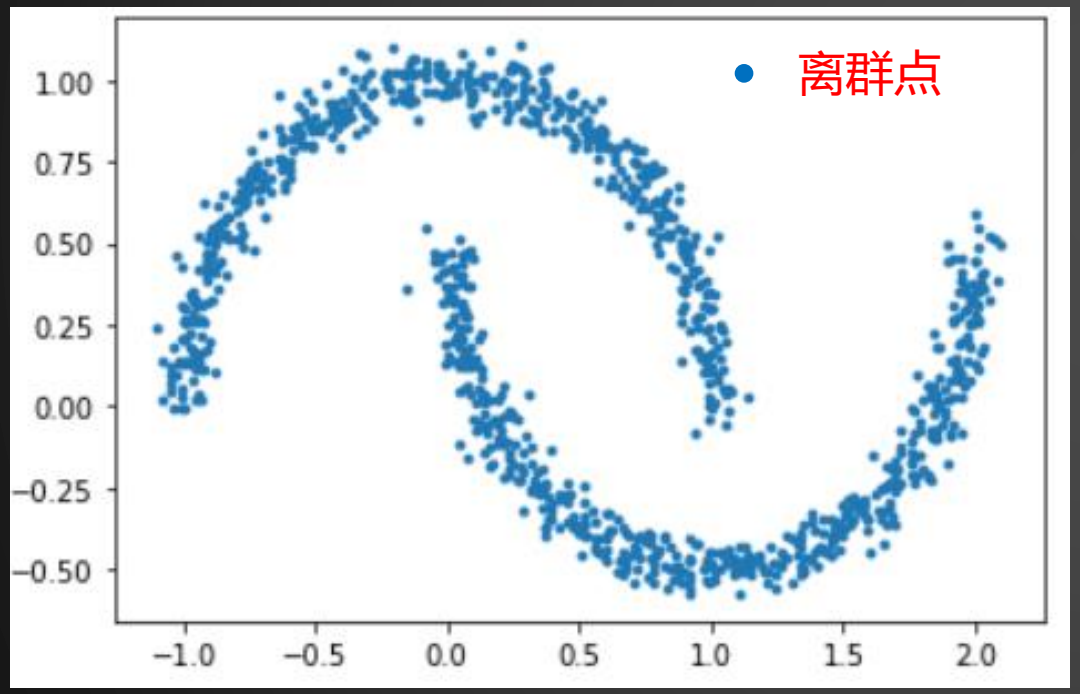
K均值聚类法的局限性

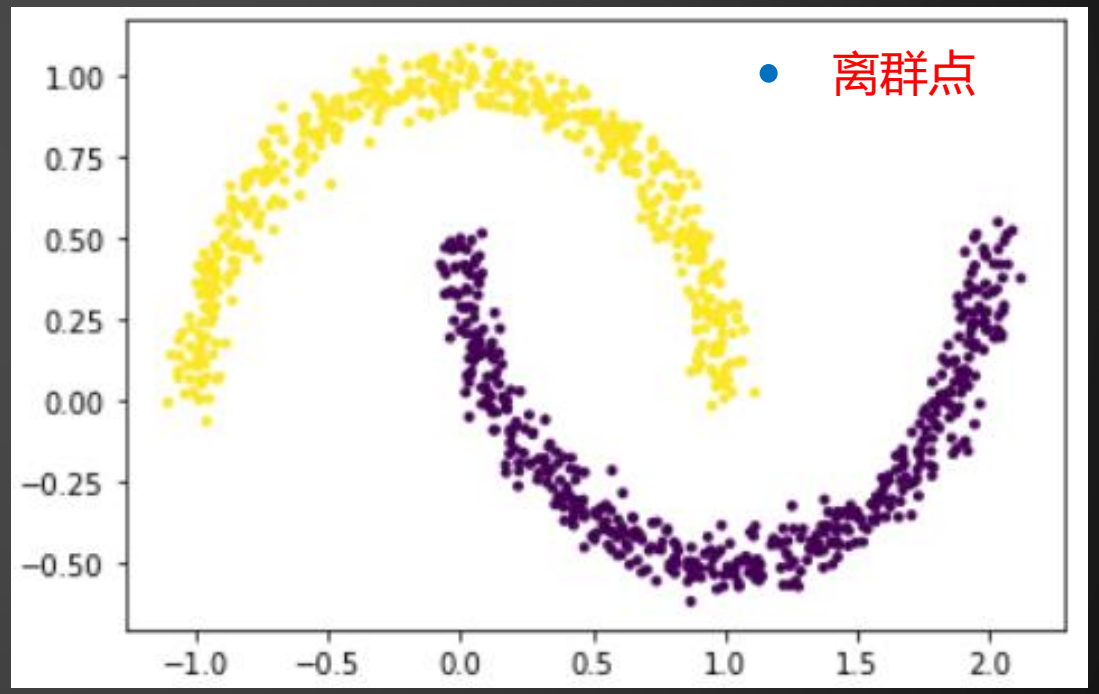

1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

DBSCAN

Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法





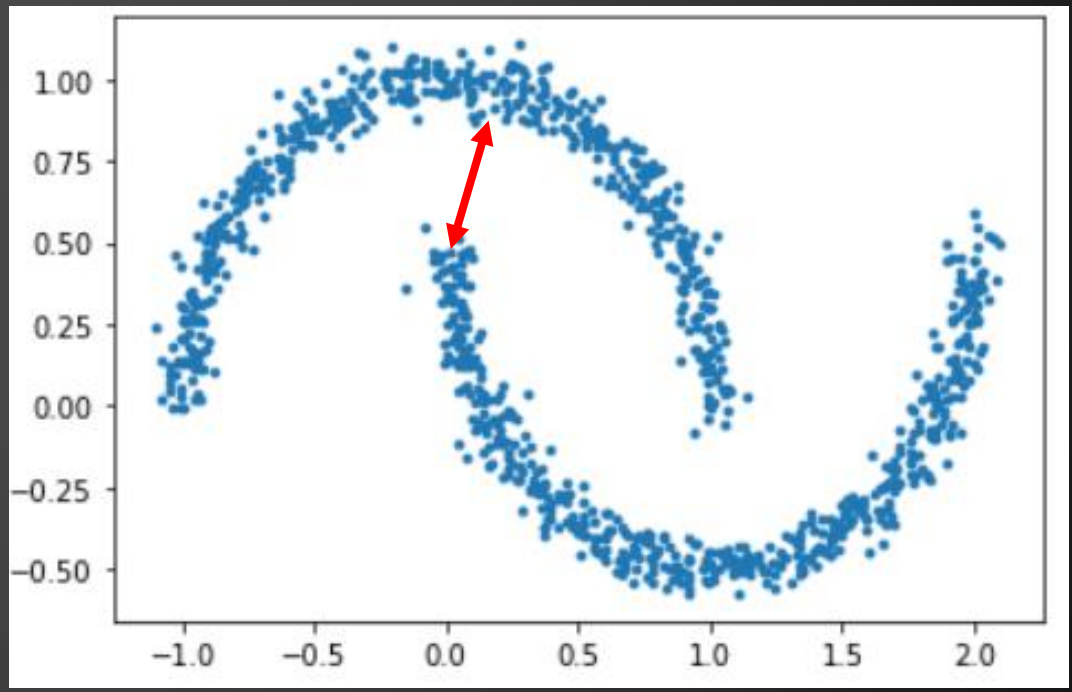
两个核心参数

半径：样本点的距离阈值

- 大了，圈住的就多了，簇的个数就少了；反之，簇的个数就多了
- 关键点：找突变点

密度：表示圈住的样本点个数的最小值

- 圈住的点的个数，相当于密度



sklearn中的参数：

- 半径：eps
- 密度：min_samples

DBSCAN评估：轮廓系数

对于某个样本来说，该样本与其所在簇内其他样本的平均距离，记为a_i，该样本与其他簇内样本的平均距离b_i，这个样本的轮廓系数为：

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- 所以，聚类分析的轮廓系数为所有样本的轮廓系数的均值：

$$s = \frac{1}{m}\sum_{i=1}^{m} s_i$$

- 轮廓系数取值范围为[−1,1]，取值越接近1则说明聚类性能越好，相反，取值越接近-1则说明聚类性能越差。

DBSCAN案例：啤酒数据聚类分析

- 案例介绍：有一些关于啤酒的样本数据，一共有20个啤酒品牌
- 四个特征：卡路里含量、纳含量、酒精含量以及售卖价格。
- 任务目标：将啤酒分为不同的类别

name	calories	sodium	alcohol	cost
Budweiser	144	15	4.7	0.43
Schlitz	151	19	4.9	0.43
Lowenbrau	157	15	0.9	0.48
Kronenbourg	170	7	5.2	0.73
Heineken	152	11	5.0	0.77