Homework 3:
聚类算法对比实验

大数据原理与技术 (Spring 2025)

22336226 王泓沣
Lectured by: Changdong Wang
Sun Yat-sen University

1 问题描述

- 在鸢尾花数据集上实现 K-means 和 DBSCAN 算法
- 调整超参数（如簇数、邻域半径）观察聚类结果变化
- 用准确率、轮廓系数和 Calinski-Harabasz 指数评估性能

2 Method

2.1 K-means

1. 初始化 K 个聚类中心：随机从数据中选 K 个点
2. 计算每个样本与各质心的距离，将样本分配给距离其最近的质心所在的簇
3. 更新质心：对每个簇的样本取平均值，得到新的质心
4. 重复分配和更新质心的步骤，直到收敛或达到最大迭代次数

2.2 DBSCAN

DBSCAN 依赖两个重要参数：`eps`：邻域半径;`min_samples`：核心点需要的最少邻居数

1. 对未访问的点，寻找其 `eps` 邻域内的所有点。
2. 若该邻域内的点数 `min_samples`，则将该点标记为“核心点”，接着对其邻域点进行密度扩展（合并到同一聚类）
3. 若邻域内的点数 $<$ `min_samples`，则标记为“噪声”点或“边界”点（在后续过程如果它恰好是别的核心点的邻域，则会并入对应聚类）。

3 Evaluation

3.1 准确率 Accuracy

对每个簇，找出在该簇里出现最多的真实类别作为该簇的“代表”标签。再看所有样本里，真正落在这个簇的样本中，有多少样本是该“代表”类别。

$$Acc = \frac{true\_label}{cluster\_label}$$

3.2 轮廓系数 Silhouette Coefficient

对每个样本 i，找到同簇内其他样本，求平均距离 = a(i)，同时找到最近的不同簇的样本群，计算与它们的平均距离 = b(i)，计算每个样本的 silhouette 值，最后取平均。

$$s(i) = \begin{cases} \frac{a(i)-b(i)}{max\{a(i),b(i)\}}, & max\{a(i),b(i)\} > 0 \\ 0, & else \end{cases}$$

3.3 Calinski-Harabasz 指数

先算出整体数据的全局均值 M, 对每个簇 C_i，计算簇均值 M_i，之后计算簇间散度 SSB 和簇内散度 SSW，最后计算 CH 指数

$$SSB = \sum_{i=1}^{k} |C_i| \cdot \|M_i - M\|^2$$

$$SSW = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - M_i\|^2$$

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)}$$

4 Result

K	Accuracy	Silhouette	Calinski-Harabasz
2	0.666667	0.686735	306.279535
3	0.773333	0.529627	182.687821
4	0.773333	0.359678	128.923359
5	0.840000	0.365143	111.034052

Table 1: K-means 调参结果

eps	n_clusters	Accuracy	Silhouette	Calinski-Harabasz
0.3	3	0.353333	0.776834	105.127502
0.4	4	0.766667	0.502528	149.157570
0.5	2	0.620000	0.735356	327.774365
0.6	2	0.633333	0.722973	324.875295
0.7	2	0.666667	0.694110	310.480854

Table 2: DBSCAN 调参结果

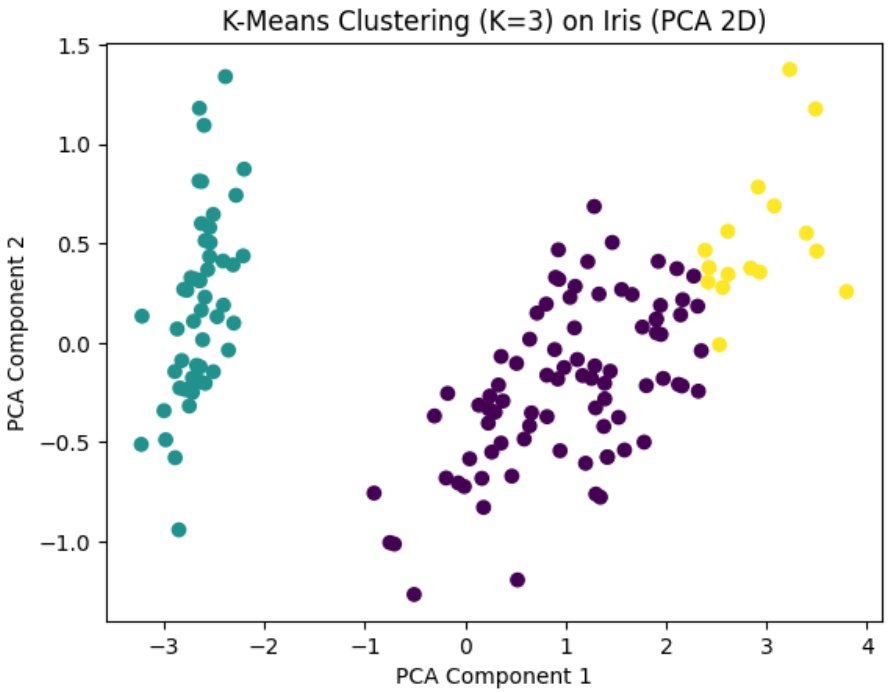

Figure 1: K-means 聚类可视化结果

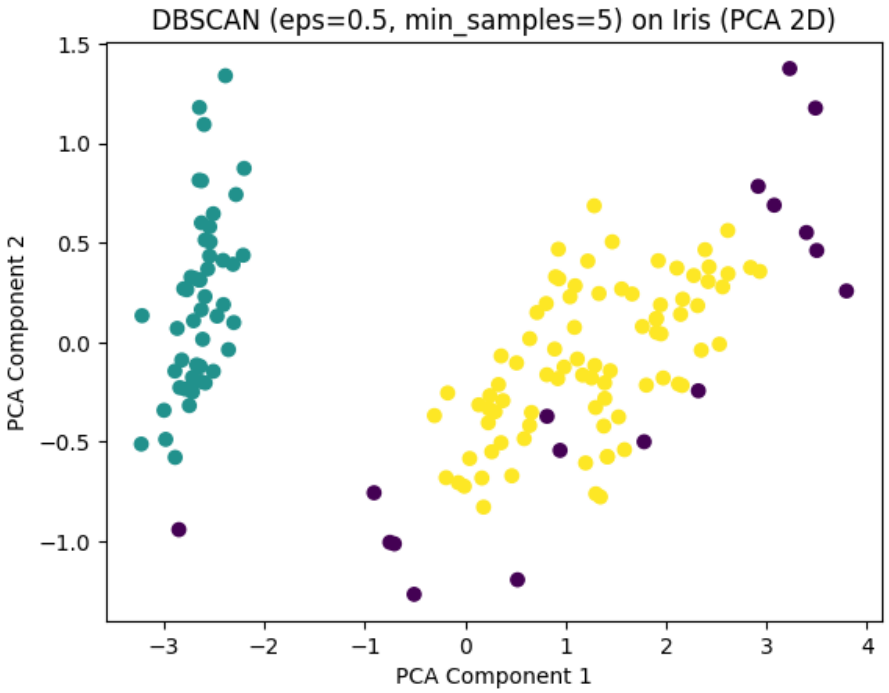

Figure 2: DBSCAN 聚类可视化结果可视化