

1

4个集合的矩阵

row	S1	S2	S3	S4
0	1	0	0	1
1	0	0	1	0
2	0	1	0	1
3	1	0	1	1
4	0	0	1	0

使用4个哈希函数模拟排列转换

- $h_1(x) = x + 1 \bmod 5$
- $h_2(x) = 3x + 1 \bmod 5$
- $h_3(x) = 2x + 4 \bmod 5$
- $h_4(x) = 3x - 1 \bmod 5$

row	$h_1(x)$	$h_2(x)$	$h_3(x)$	$h_4(x)$
0	1	1	4	4
1	2	4	1	2
2	3	2	3	0
3	4	0	0	3
4	0	3	2	1

遍历排列顺序，确定最小哈希值

	S1	S2	S3	S4
$h_1(x)$	1	3	0	1
$h_2(x)$	0	2	0	0
$h_3(x)$	0	3	0	0
$h_4(x)$	3	0	1	0

2

(a) 两个簇上点之间的最短距离，两个点分别来自不同的簇。

- 1.合并 (11,4) 和 (12,3)

2. 合并 (10,5) 和第一步得到的簇

(b) 簇上点对之间的平均距离 (质心之间的距离), 两个点分别来自不同的簇。

1. 合并 (11,4) 和 (12,3), 新簇质心为 (11.5,3.5)

2. 合并 (4,8) 和 (4,10), 新簇质心为 (4,9)

### 3

#### 计算所有点与 (3, 4) 的欧几里得距离

公式为:

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

对每个点计算距离:

- (2,2):  $\sqrt{[(2-3)^2 + (2-4)^2]} = \sqrt{5} \approx 2.24$
- (4,8):  $\sqrt{[(4-3)^2 + (8-4)^2]} = \sqrt{17} \approx 4.12$
- (4,10):  $\sqrt{[(4-3)^2 + (10-4)^2]} = \sqrt{37} \approx 6.08$
- (5,2):  $\sqrt{[(5-3)^2 + (2-4)^2]} = \sqrt{8} \approx 2.83$
- (6,8):  $\sqrt{[(6-3)^2 + (8-4)^2]} = \sqrt{25} = 5.00$
- (7,10):  $\sqrt{[(7-3)^2 + (10-4)^2]} = \sqrt{52} \approx 7.21$
- (9,3):  $\sqrt{[(9-3)^2 + (3-4)^2]} = \sqrt{37} \approx 6.08$
- (10,5):  $\sqrt{[(10-3)^2 + (5-4)^2]} = \sqrt{50} \approx 7.07$
- (11,4):  $\sqrt{[(11-3)^2 + 0^2]} = \sqrt{64} = 8.00$
- (12,3):  $\sqrt{[(12-3)^2 + (3-4)^2]} = \sqrt{82} \approx 9.06$
- (12,6):  $\sqrt{[(12-3)^2 + (6-4)^2]} = \sqrt{85} \approx 9.22$  ✓

与 (3,4) 最远的是 (12,6), 作为第二个初始点

#### 第三个点要尽量远离前两个点: (3,4) 和 (12,6)

方法是对所有剩余点, 计算它们与前两个点的最小距离, 并选择最小距离最大的那个点。

点	距离 (3,4)	距离 (12,6)	最小距离
(2,2)	2.24	10.20	2.24
(4,8)	4.12	8.25	4.12
(4,10)	6.08	8.54	6.08
(5,2)	2.83	8.25	2.83
(6,8)	5.00	6.71	5.00
(7,10)	7.21	6.71	6.71 ✓
(9,3)	6.08	3.61	3.61
(10,5)	7.07	2.24	2.24

点	距离 (3,4)	距离 (12,6)	最小距离
(11,4)	8.00	2.24	2.24
(12,3)	9.06	3.00	3.00

最小距离最大的是 (7,10)，选它作为第三个初始点

✅ 最终选择的三个点是：

- (3, 4)
- (12, 6)
- (7, 10)

## 4

(a) 利用 BFR 算法中的方式计算所有簇的表示

簇	N	SUM(x,y)	SUMSQ(x,y)
● 橙色簇	3	(10, 8)	(38, 24)
● 蓝色簇	4	(21, 36)	(117, 328)
● 灰色簇	5	(54, 21)	(590, 95)

(b) 计算每个簇在两个维度中每个维上的方差和标准差

$$\begin{aligned} \text{平均值: } \mu_i &= \frac{\text{SUM}_i}{N} \\ \text{方差: } \sigma_i^2 &= \frac{\text{SUMSQ}_i}{N} - \mu_i^2 \\ \text{标准差: } \sigma_i &= \sqrt{\sigma_i^2} \end{aligned}$$

簇	$\mu(x,y)$	$\sigma^2(x,y)$	$\sigma(x,y)$
● 橙色簇	(3.33,2.67)	(1.56, 0.89)	(1.25, 0.94)
● 蓝色簇	(5.25,9.00)	(1.69, 1.00)	(1.30, 1.00)
● 灰色簇	(10.8,4.2)	(1.36, 1.36)	(1.17, 1.17)