日期: /

一. K means. 计算.

1.1 方法:

1. 输入训练数据和聚类数目K;
2. 执行下面二者之一:
   - 随机将数据分为K个类C_1, ……, C_K, 计算每个类的中心 c_i, $i = 1, \ldots, K$;
   - 指定K个类的中心 c_i, , $i = 1, \ldots, K$, 将所有数据点划分到离其最近的类中心所在的类
3. 计算每个数据点到其所属类的中心的平方距离

   无了 $$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|_2^2, \quad c_i = \frac{1}{n_i} \sum_{x \in C_i} x$$
4. 重新将每个数据点划分到离其最近的类中心所在的类, 使得 SSE 减少. 完成后重新计算各类的中心 c_i, $i = 1, \ldots, K$;
5. 重复 3 和 4, 直到没有样本点需要调整（SSE不能再减少）;

1.2 例题:

用k-均值算法将右表中的8个点聚为三个簇，假设第一次迭代选择序号1、序号4和序号7当作初始点，请给出第一次执行后的三个聚类中心以及最后的三个簇

序号	属性1	属性2
1	2	10
2	2	5
3	8	4
4	5	8
5	7	5
6	6	4
7	1	2
8	4	9

解: 第一次迭代过程如下:

(1) 初始类中心以向量形式表示为

(2, 10) (5, 8) (1, 2)

(2) 计算各点到3个类中心的距离, 如下表所示.

日期: /	类1 (2,10)	类2 (5,8)	类3 (1,2)	类别
1	(2,10)	0	$\sqrt{9+4}=\sqrt{13}$	$\sqrt{1+64}=\sqrt{65}$	1
2	(2,5)	$\sqrt{25}$	$\sqrt{9+9}=\sqrt{18}$	$\sqrt{1+9}=\sqrt{10}$	3
3	(8,4)	$\sqrt{36+36}=\sqrt{72}$	$\sqrt{9+16}=\sqrt{25}$	$\sqrt{49+4}=\sqrt{53}$	2
4	(5,8)	$\sqrt{9+4}=\sqrt{13}$	0	$\sqrt{16+36}=\sqrt{52}$	2
5	(7,5)	$\sqrt{49+25}=\sqrt{50}$	$\sqrt{4+9}=\sqrt{13}$	$\sqrt{36+9}=\sqrt{45}$	2
6	(6,4)	$\sqrt{16+36}=\sqrt{52}$	$\sqrt{1+16}=\sqrt{17}$	$\sqrt{25+4}=\sqrt{29}$	2
7	(1,2)	$\sqrt{1+64}=\sqrt{65}$	$\sqrt{16+36}=\sqrt{52}$	0	3
8	(4,9)	$\sqrt{4+1}=\sqrt{5}$	$\sqrt{1+1}=\sqrt{2}$	$\sqrt{9+49}=\sqrt{58}$	2

类1： 1：(2,10) => 类中心 (2, 10)

类2： 3：(8,4)，4：(5,8)，5：(7,5)，6：(6,4)

8：(4,9) => 类中心 (6, 6)

类3： 2：(2,5)，7：(1,2) => 类中心 (1.5, 3.5)

	1 (2,10)	2 (6,6)	3 (1.5,3.5)	类别
1	(2,10)	0	$\sqrt{16+16}=\sqrt{32}$	$\sqrt{\frac{1}{4}+\frac{169}{4}}=\sqrt{42.5}$	1
2	(2,5)	$\sqrt{25}$	$\sqrt{16+1}=\sqrt{17}$	$\sqrt{\frac{1}{4}+\frac{5}{4}}=\sqrt{2.5}$	3
3	(8,4)	$\sqrt{36+36}=\sqrt{72}$	$\sqrt{4+4}=\sqrt{8}$	$\sqrt{\frac{169}{4}+\frac{1}{4}}=\sqrt{\frac{170}{4}}$	2
4	(5,8)	$\sqrt{13}$	$\sqrt{1+4}=\sqrt{5}$	$\sqrt{\frac{13}{4}+\frac{121}{4}}=\sqrt{\frac{134}{4}}$	2
5	(7,5)	$\sqrt{50}$	$\sqrt{1+1}=\sqrt{2}$	$\sqrt{\frac{25}{4}+\frac{121}{4}}$ $\sqrt{\frac{146}{4}}$	2
6	(6,4)	$\sqrt{52}$	$\sqrt{0+4}=\sqrt{4}$	$\sqrt{\frac{1}{4}+\frac{169}{4}}=\sqrt{\frac{170}{4}}$	2
7	(1,2)	$\sqrt{65}$	$\sqrt{25+16}=\sqrt{41}$	$\sqrt{\frac{25}{4}+\frac{1}{4}}=\sqrt{\frac{164}{4}}$	3
8	(4,9)	$\sqrt{5}$	$\sqrt{4+9}=\sqrt{13}$	$\sqrt{\frac{121}{4}+\frac{25}{4}}=\sqrt{\frac{146}{4}}$	1

日期: /

类1： 1：(2,10)，8：(4,5) => 类中心 (3, 5.5)

类2： 3：(8,4)，4：(5,8)，5：(7,5)，6：(16,4)

=> 类中心 $(\frac{20}{4}, \frac{21}{4})$

类3： 2：(2,5)，7：(1,2) => 类中心 (1.5, 3.5)

		1 $(3, \frac{19}{2})$	2 $(\frac{13}{2}, \frac{21}{4})$	3 (1.5, 3.5)	类别
1	(2,10)	$\sqrt{1+\frac{1}{4}}=\sqrt{\frac{5}{4}}$	$\sqrt{\frac{5}{4}}+\cdots$	$\sqrt{\frac{1}{4}+\frac{169}{4}}=\sqrt{42.5}$	1
2	(2,5)	$\sqrt{1+\frac{5}{4}}=\sqrt{\frac{13}{4}}$	$\sqrt{\frac{81}{4}+\frac{1}{16}}$	$\sqrt{\frac{1}{4}+\frac{5}{4}}=\sqrt{2.5}$	3
3	(8,4)	$\sqrt{25+\frac{1}{4}}$	$\sqrt{\frac{5}{4}+\frac{121}{16}}$	$\sqrt{\frac{169}{4}+\frac{1}{4}}=\sqrt{\frac{170}{4}}$	2
4	(5,8)	$\sqrt{4+\frac{9}{4}}$	$\sqrt{\frac{9}{4}+\frac{121}{16}}$	$\sqrt{\frac{169}{4}+\frac{9}{4}}=\sqrt{\frac{130}{4}}$	1
5	(7,5)	$\sqrt{16+\frac{25}{4}}$	$\sqrt{\frac{49}{4}+\frac{1}{16}}$	$\sqrt{\frac{121}{4}+\frac{5}{4}}=\sqrt{\frac{126}{4}}$	2
6	(16,4)	$\sqrt{9+\frac{121}{4}}$	$\sqrt{\frac{1}{4}+\frac{121}{16}}$	$\sqrt{\frac{169}{4}+\frac{1}{4}}=\sqrt{\frac{82}{4}}$	2
7	(1,2)	$\sqrt{4+\frac{225}{4}}$	$\sqrt{\frac{121}{4}+\frac{169}{16}}$	$\sqrt{\frac{5}{4}+\frac{1}{4}}=\sqrt{\frac{17}{4}}$	3
8	(4,9)	$\sqrt{1+\frac{49}{4}}$	$\sqrt{\frac{81}{4}+1}$	$\sqrt{\frac{121}{4}+\frac{25}{4}}=\sqrt{\frac{146}{4}}$	1

类1： 1：(2,10)，8：(4,5)，4 (5,8)

类2： 3：(8,4)，5：(7,5)，6：(16,4)

类3： 2：(2,5)，7：(1,2) => 类中心 (1.5, 3.5)

此时到了终点.

答案 (1,4,8) (3,5,6) (2,7)

日期: /

二、层次聚类：（单链，全链，组平均）

2.1

口基于距离(相异度)的层次聚类

➢1. 单链层次聚类

- 步骤：① 找出所有点距离最小的两个点，第一个合并；
  ② 按照“小中取小”的原则依次合并剩余点，直至合并完所有点。

➢2. 全链层次聚类

- 步骤：① 找出所有点距离最小的两个点，第一个合并；
  ② 按照“大中取小”的原则依次合并剩余的点，直至所有点合并完成。

口基于相似度矩阵的层次聚类

➢1. 单链层次聚类

- 步骤：① 找出所有点相似度最大的两个点，第一个合并；
  ② 按照“大中取大”的原则进行合并剩余的点，直至所有点合并完成为止。

➢2. 全链层次聚类

- 步骤：① 找出所有点相似度最大的两个点，第一个合并；
  ② 按照“小中取大”的原则进行合并剩余的点，直至所有点合并完成为止。

2.2 举例

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

6x6 矩阵，记录两两城市间的距离

日期: /

1. 单链:

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

合并TO/MI

	BA	FI	NA	RM	MI/TO
BA	0	662	255	412	877
FI	662	0	468	268	295
NA	255	468	0	219	754
RM	412	268	219	0	564
MI/TO	877	295	754	564	0

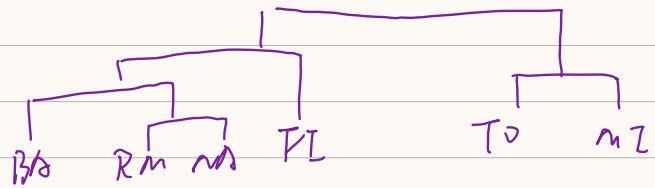
合并RM

	BA	FI	NA/RM	MI/TO
BA	0	662	255	877
FI	662	0	268	295
NA/RM	255	268	0	564
MI/TO	877	295	564	0

合并BA

	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
TO/MI	564	295	0

	BA/NA/RM/FI	TO/MI
B~	0	295
TO/MI	295	0



日期: /

2.3 举例

图7.3 6个点的xy坐标

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

单链接
全链接

	p1	p2	p3	p4	p5	p6
p1	0	0.2343	0.2158	0.3676	0.3417	0.2353
p2	0.2343	0	0.1431	0.1941	0.1431	0.2495
p3	0.2158	0.1431	0	0.1587	0.2846	0.1019
p4	0.3676	0.1941	0.1587	0	0.2842	0.2195
p5	0.3417	0.1431	0.2846	0.2842	0	0.3860
p6	0.2353	0.2495	0.1019	0.2195	0.3860	0

① 单链接：3和6合并

	p1	p2	p3/p6	p4	p5
p1	0	0.2343	0.2158	0.3676	0.3417
p2	0.2343	0	0.1431	0.1941	0.1431
p3/p6	0.2158	0.1431	0	0.1587	0.2846
p4	0.3676	0.1941	0.1587	0	0.2842
p5	0.3417	0.1431	0.2846	0.2842	0

日期: /

2,5 合并

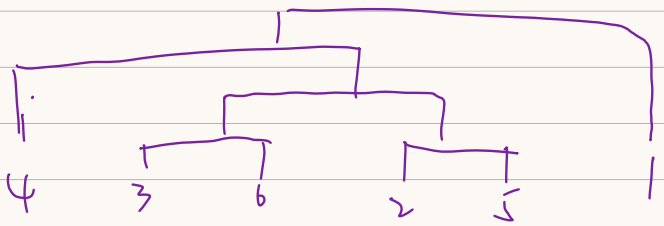
	p1	p2/p5	p3/p6	p4
p1	0	0.2343	0.2158	0.3676
p2/p5	0.2343	0	0.1431	0.1941
p3/p6	0.2158	0.1431	0	0.1587
p4	0.3676	0.1941	0.1587	0

2,5,3,6 合并

	p1	p2/p5/p3/p6	p4
p1	0	0.2158	0.3676
p2/p5/p3/p6	0.2158	0	0.1587
p4	0.3676	0.1587	0

4 5 2,5,3,6 合并

	p1	p2/p5/p3/p6/p4
p1	0	0.2158
p4/p2/p5/p3/p6	0.2158	0



日期: /

② 全链接：3和6合并

	p1	p2	p3/p6	p4	p5
p1	0	0.2343	0.2353	0.3676	0.3417
p2	0.2343	0	0.2435	0.1941	0.1431
p3/p6	0.2153	0.2435	0	0.2125	0.3860
p4	0.3676	0.1941	0.2125	0	0.2842
p5	0.3417	0.1431	0.3860	0.2842	0

合并 2,5

	p1	p2/p5	p3/p6	p4
p1	0	0.3417	0.2353	0.3676
p2/p5	0.3417	0	0.3860	0.2845
p3/p6	0.2153	0.3860	0	0.2125
p4	0.3676	0.2845	0.2125	0

合并 4，3，6

	p1	p2/p5	p3/p6/p4
p1	0	0.3417	0.3676
p2/p5	0.3417	0	0.3860
p4/p3/p6	0.3676	0.3860	0

1，合并 1，2，5：

	p1/p2/p5	p3/p6/p4
p5/p1/p2	0	0.3860
p4/p3/p6	0.3860	0

日期: /