参考答案

备注：目前参考答案内容以英文为主，仅供参考。

第一题

初始时，签名矩阵为:

	S1	S2	S3	S4
h1	∞	∞	∞	∞
h2	∞	∞	∞	∞
h3	∞	∞	∞	∞
h4	∞	∞	∞	∞

首先，考虑原始矩阵中第 0 行。h1(0)=1;h2(0)=1,h3(0)=4;h4(0)=4。此外，只有 S1 和 S4 列的值为 1，因此签名矩阵中这两列的值需要修改，修改后的签名矩阵为：

	S1	S2	S3	S4
h1	1	∞	∞	1
h2	1	∞	∞	1
h3	4	∞	∞	4
h4	4	∞	∞	4

接下来，考虑原始矩阵中第 1 行。h1(1)=2；h2(1)=4;h3(1)=1;h4(1)=2. 此外，有 S3 列的值为 1，因此签名矩阵中这一列的值需要修改，修改后的签名矩阵为:

	S1	S2	S3	S4
h1	1	∞	2	1

h2	1	∞	4	1
h3	4	∞	1	4
h4	4	∞	2	4

接下来，考虑原始矩阵中第 2 行。h1(2)=3；h2(2)=2;h3(2)=3;h4(2)=0. 此外，只有 S2 和 S4 列的值为 1，因此签名矩阵中这两列的值需要对比修改为更小的值，修改后的签名矩阵为:

	S1	S2	S3	S4
h1	1	3	2	1
h2	1	2	4	1
h3	4	3	1	3
h4	4	0	2	0

接下来，考虑原始矩阵中第 3 行。h1(3)=4；h2(3)=0;h3(3)=0;h4(3)=3. 此外，有 S1，S3 和 S4 列的值为 1，因此签名矩阵中这三列的值需要对比修改为更小的值，修改后的签名矩阵为:

	S1	S2	S3	S4
h1	1	3	2	1
h2	0	2	0	0
h3	0	3	0	0
h4	3	0	2	0

接下来，考虑原始矩阵中第 4 行。h1(4)=0；h2(4)=3;h3(4)=2;h4(4)=1. 此外，有 S3 列的值为 1，因此签名矩阵中这列的值需要对比修改为更小的值，修改后的最终签名矩阵为:

	S1	S2	S3	S4
h1	1	3	0	1
h2	0	2	0	0
h3	0	3	0	0

h4	3	0	1	0

第二题

（a）根据图的信息，先计算任意两个数据之间的欧氏距离得到一个相似度矩阵(proximity matrix)如下【备注是对称矩阵，因此只写了其中一部分信息】

	(2,2)	(3,4)	(4,8)	(4,10)	(5,2)	(6,8)	(7,10)	(9,3)	(10,5)	(11,4)	(12,3)	(12,6)
(2,2)	0											
(3,4)	$\sqrt{5}$	0										
(4,8)	$\sqrt{40}$	$\sqrt{17}$	0									
(4,10)	$\sqrt{68}$	$\sqrt{37}$	$\sqrt{4}$	0								
(5,2)	$\sqrt{6}$	$\sqrt{16}$	$\sqrt{37}$	$\sqrt{65}$	0							
(6,8)	$\sqrt{52}$	$\sqrt{25}$	$\sqrt{4}$	$\sqrt{16}$	$\sqrt{37}$	0						
(7,10)	$\sqrt{89}$	$\sqrt{52}$	$\sqrt{13}$	$\sqrt{9}$	$\sqrt{68}$	$\sqrt{5}$	0					
(9,3)	$\sqrt{50}$	$\sqrt{37}$	$\sqrt{50}$	$\sqrt{74}$	$\sqrt{17}$	$\sqrt{34}$	$\sqrt{53}$	0				
(10,5)	$\sqrt{73}$	$\sqrt{50}$	$\sqrt{45}$	$\sqrt{61}$	$\sqrt{34}$	$\sqrt{25}$	$\sqrt{34}$	$\sqrt{5}$	0			
(11,4)	$\sqrt{85}$	$\sqrt{64}$	$\sqrt{65}$	$\sqrt{85}$	$\sqrt{40}$	$\sqrt{41}$	$\sqrt{52}$	$\sqrt{5}$	$\sqrt{2}$	0		
(12,3)	$\sqrt{101}$	$\sqrt{82}$	$\sqrt{89}$	$\sqrt{113}$	$\sqrt{50}$	$\sqrt{61}$	$\sqrt{74}$	$\sqrt{9}$	$\sqrt{16}$	$\sqrt{2}$	0	
(12,6)	$\sqrt{116}$	$\sqrt{85}$	$\sqrt{68}$	$\sqrt{80}$	$\sqrt{65}$	$\sqrt{40}$	$\sqrt{41}$	$\sqrt{18}$	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{9}$	0

一开始每个点构成一个簇，其簇质心就是点本身。现在簇间距离定义为两个簇上点之间的最短距离，两个点分别来自不同的簇。在所有的点对之间的距离最近（10，5）和（11,4），或（11,4）和（12,3），都是$\sqrt{2}$。因此随机选择一组合并。例如（11,4）和（12,3），将他们当做一个新的簇。那么接下来更新簇到簇之间的距离，这儿定义为两个簇上点对之间的最短距离，且这两个点来自不同的簇。因此更新相似度矩阵如下：

	(2,2)	(3,4)	(4,8)	(4,10)	(5,2)	(6,8)	(7,10)	(9,3)	(10,5)	(11,4),(12,3)	(12,6)
(2,2)	0										
(3,4)	$\sqrt{5}$	0									
(4,8)	$\sqrt{40}$	$\sqrt{17}$	0								
(4,10)	$\sqrt{68}$	$\sqrt{37}$	$\sqrt{4}$	0							
(5,2)	$\sqrt{6}$	$\sqrt{16}$	$\sqrt{37}$	$\sqrt{65}$	0						
(6,8)	$\sqrt{52}$	$\sqrt{25}$	$\sqrt{4}$	$\sqrt{16}$	$\sqrt{37}$	0					
(7,10)	$\sqrt{89}$	$\sqrt{52}$	$\sqrt{13}$	$\sqrt{9}$	$\sqrt{68}$	$\sqrt{5}$	0				
(9,3)	$\sqrt{50}$	$\sqrt{37}$	$\sqrt{50}$	$\sqrt{74}$	$\sqrt{17}$	$\sqrt{34}$	$\sqrt{53}$	0			
(10,5)	$\sqrt{73}$	$\sqrt{50}$	$\sqrt{45}$	$\sqrt{61}$	$\sqrt{34}$	$\sqrt{25}$	$\sqrt{34}$	$\sqrt{5}$	0		
(11,4),(12,3)	$\sqrt{85}$	$\sqrt{64}$	$\sqrt{65}$	$\sqrt{85}$	$\sqrt{40}$	$\sqrt{41}$	$\sqrt{52}$	$\sqrt{5}$	$\sqrt{2}$	0	
(12,6)	$\sqrt{116}$	$\sqrt{85}$	$\sqrt{68}$	$\sqrt{80}$	$\sqrt{65}$	$\sqrt{40}$	$\sqrt{41}$	$\sqrt{18}$	$\sqrt{5}$	$\sqrt{5}$	0

那么接下来再去检查簇之间的最短距离，发现（10，5）聚类刚才的簇（11,4），（12,3)距离最近，因此将该点聚类到一起。因此更新相似度矩阵如下：

	(2,2)	(3,4)	(4,8)	(4,10)	(5,2)	(6,8)	(7,10)	(9,3)	(11,4),(12,3),(10,5)	(12,6)
(2,2)	0									
(3,4)	$\sqrt{5}$	0								
(4,8)	$\sqrt{40}$	$\sqrt{17}$	0							
(4,10)	$\sqrt{68}$	$\sqrt{37}$	$\sqrt{4}$	0						
(5,2)	$\sqrt{6}$	$\sqrt{16}$	$\sqrt{37}$	$\sqrt{65}$	0					
(6,8)	$\sqrt{52}$	$\sqrt{25}$	$\sqrt{4}$	$\sqrt{16}$	$\sqrt{37}$	0				
(7,10)	$\sqrt{89}$	$\sqrt{52}$	$\sqrt{13}$	$\sqrt{9}$	$\sqrt{68}$	$\sqrt{5}$	0			
(9,3)	$\sqrt{50}$	$\sqrt{37}$	$\sqrt{50}$	$\sqrt{74}$	$\sqrt{17}$	$\sqrt{34}$	$\sqrt{53}$	0		

(11,4),(12,3),(10,5)	$\sqrt{73}$	$\sqrt{50}$	$\sqrt{45}$	$\sqrt{61}$	$\sqrt{34}$	$\sqrt{25}$	$\sqrt{34}$	$\sqrt{5}$	0	
(12,6)	$\sqrt{116}$	$\sqrt{85}$	$\sqrt{68}$	$\sqrt{80}$	$\sqrt{65}$	$\sqrt{40}$	$\sqrt{41}$	$\sqrt{18}$	$\sqrt{5}$	0

那么接下来再去检查簇之间的最短距离，发现（4，10）聚类点(4,8)距离最近，因此将这两个点聚类到一起。因此更新相似度矩阵如下：

	(2,2)	(3,4)	(4,8),(4,10)	(5,2)	(6,8)	(7,10)	(9,3)	(11,4),(12,3),(10,5)	(12,6)
(2,2)	0								
(3,4)	$\sqrt{5}$	0							
(4,10),(4,8)	$\sqrt{40}$	$\sqrt{17}$	0						
(5,2)	$\sqrt{6}$	$\sqrt{16}$	$\sqrt{37}$	0					
(6,8)	$\sqrt{52}$	$\sqrt{25}$	$\sqrt{4}$	$\sqrt{37}$	0				
(7,10)	$\sqrt{89}$	$\sqrt{52}$	$\sqrt{13}$	$\sqrt{68}$	$\sqrt{5}$	0			
(9,3)	$\sqrt{50}$	$\sqrt{37}$	$\sqrt{50}$	$\sqrt{17}$	$\sqrt{34}$	$\sqrt{53}$	0		
(11,4),(12,3),(10,5)	$\sqrt{73}$	$\sqrt{50}$	$\sqrt{45}$	$\sqrt{34}$	$\sqrt{25}$	$\sqrt{34}$	$\sqrt{5}$	0	
(12,6)	$\sqrt{116}$	$\sqrt{85}$	$\sqrt{68}$	$\sqrt{65}$	$\sqrt{40}$	$\sqrt{41}$	$\sqrt{18}$	$\sqrt{5}$	0

如此循环。

（b）初始时最先也是先选两个最近的点合并。例如（11,4）和（12,3），将他们当做一个新的簇。接下来，采用簇上点对的平均距离时，需要更详细地计算不同簇之间的所有点对的距离再求平均值作为簇间距离。其他步骤类似(a)。

第三题

在图中除了（3,4）点以外的其他 11 个点到（3,4）的距离分别是：

	(3,4)
(2,2)	$\sqrt{(3-2)^2+(4-2)^2}=\sqrt{5}$
(4,8)	$\sqrt{(4-3)^2+(8-4)^2}=\sqrt{17}$
(4,10)	$\sqrt{(4-3)^2+(10-4)^2}=\sqrt{37}$
(5,2)	$\sqrt{(5-3)^2+(4-2)^2}=\sqrt{16}$
(6,8)	$\sqrt{(6-3)^2+(8-4)^2}=\sqrt{25}$
(7,10)	$\sqrt{(7-3)^2+(10-4)^2}=\sqrt{52}$
(9,3)	$\sqrt{(9-3)^2+(4-3)^2}=\sqrt{37}$
(10,5)	$\sqrt{(10-3)^2+(5-4)^2}=\sqrt{50}$
(11,4)	$\sqrt{(11-3)^2+(4-4)^2}=\sqrt{64}$
(12,3)	$\sqrt{(12-3)^2+(4-3)^2}=\sqrt{82}$
(12,6)	$\sqrt{(12-3)^2+(6-4)^2}=\sqrt{85}$

根据表格可知，距离（3,4）最远的点是（12,6），因此第二个初始点为（12,6）。
在剩下的 10 个点中，离（3,4）或（12,6）的距离分别是：

	(3,4)	(12,6)	该点的最终得分

			（两个距离中取最小的那个值）
(2,2)	$\sqrt{5}$	$\sqrt{(12-2)^2+(6-2)^2}=\sqrt{116}$	$\sqrt{5}$
(4,8)	$\sqrt{17}$	$\sqrt{(12-4)^2+(8-6)^2}=\sqrt{68}$	$\sqrt{17}$
(4,10)	$\sqrt{37}$	$\sqrt{(12-4)^2+(10-6)^2}=\sqrt{80}$	$\sqrt{37}$
(5,2)	$\sqrt{16}$	$\sqrt{(12-5)^2+(6-2)^2}=\sqrt{65}$	$\sqrt{16}$
(6,8)	$\sqrt{25}$	$\sqrt{(12-6)^2+(8-6)^2}=\sqrt{40}$	$\sqrt{25}$
(7,10)	$\sqrt{52}$	$\sqrt{(12-7)^2+(10-6)^2}=\sqrt{41}$	$\sqrt{41}$
(9,3)	$\sqrt{37}$	$\sqrt{(12-9)^2+(6-3)^2}=\sqrt{18}$	$\sqrt{18}$
(10,5)	$\sqrt{50}$	$\sqrt{(12-10)^2+(6-5)^2}=\sqrt{5}$	$\sqrt{5}$
(11,4)	$\sqrt{64}$	$\sqrt{(12-11)^2+(6-4)^2}=\sqrt{5}$	$\sqrt{5}$
(12,3)	$\sqrt{82}$	$\sqrt{(12-12)^2+(6-3)^2}=\sqrt{9}$	$\sqrt{9}$

从该表格可知，到点（3,4）或（12,6）的最短距离最大的点是（7,10）。
综上，我们选择（3,4），（12,6），（7,10）作为三个初始点。

第四题

（a）根据图，图中存在三个簇：
Cluster 1: (4,10), (7,10), (4,8), (6,8)
Cluster 2: (12,6), (10,5), (11,4), (3,4), (9,3)
Cluster 3: (2,2), (5,2), (12,3)

Cluster 1:
N = 4
SUM_x = 21, SUM_y = 36
SUMSQ_x = 117, SUMSQ_y = 328

Cluster 2:
N = 5
SUM_x = 54, SUM_y = 21
SUMSQ_x = 590, SUMSQ_y = 95

Cluster 3:
N = 3
SUM_x = 10, SUM_y = 8
SUMSQ_x = 38, SUMSQ_y = 24

（b）Cluster 1:
Variance_x = 27/16, Variance_y = 1
Standard Deviation_x = 3 / 4*√3, Standard Deviation_y = 1

Cluster 2:
Variance_x = 34/25, Variance_y = 34/25
Standard Deviation_x = √34 / 5, Standard Deviation_y = √34 / 5

Cluster 3:
Variance_x = 14/9, Variance_y = 8/9
Standard Deviation_x = √14 / 3, Standard Deviation_y = 2/3 √2