Northeastern University, Data Mining Techniques - CS6220
Fall 2017

Solutions to Homework 2, Part 2

Nakul Camasamudram

October 24, 2017

1. K-Means

Solution:

- **Iteration 1:**
  The initial centroids are C_1 = (2, 10) C_2 = (1, 2) C_3 = (5, 8). Let D(C_i) represent the **euclidean** distance between the respective point and the i^{th} centroid.

  The *E*-step:

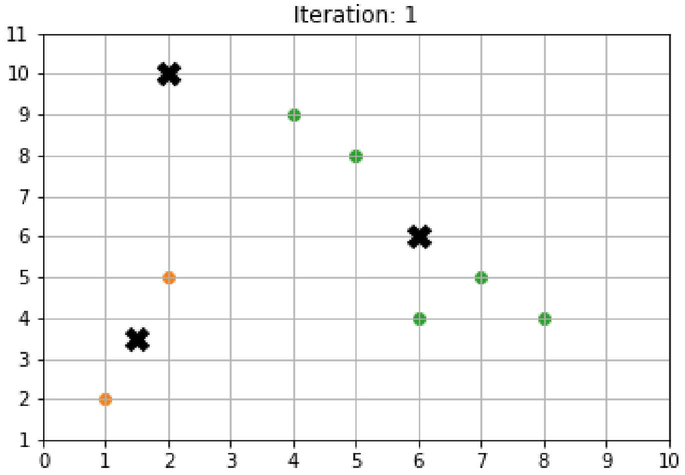
Data Point	**D(C_1)**	**D(C_2)**	**D(C_3)**	**Optimal centroid**
(4,9)	2.23606797749979	7.615773105863909	**1.4142135623730951**	C_3
(2,10)	**0.0**	8.06225774829855	3.605551275463989	C_1
(1,2)	8.06225774829855	**0.0**	7.211102550927978	C_2
(2,5)	5.0	**3.1622776601683795**	4.242640687119285	C_2
(6,4)	7.211102550927978	5.385164807134504	**4.123105625617661**	C_3
(8,4)	8.48528137423857	7.280109889280518	**5.0**	C_3
(7, 5)	7.0710678118654755	6.708203932499369	**3.605551275463989**	C_3
(5, 8)	3.605551275463989	7.211102550927978	**0.0**	C_3

  The *M*-step:

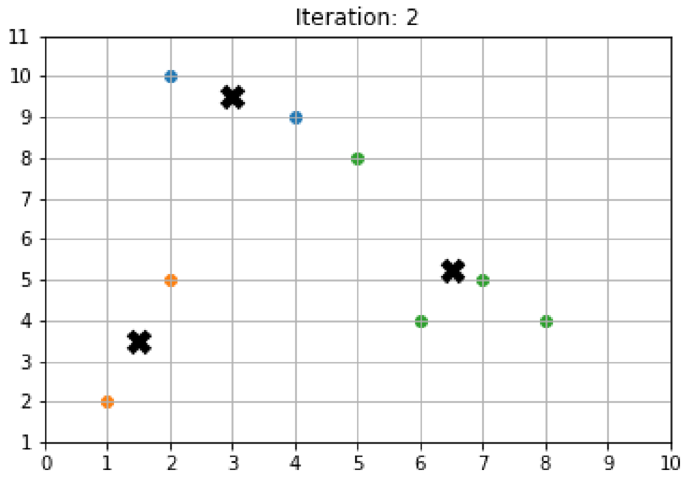
  C_1 = mean[(2, 10)] = (2.0, 10.0)

  C_2 = mean[(1, 2), (2, 5)] = (1.5, 3.5)

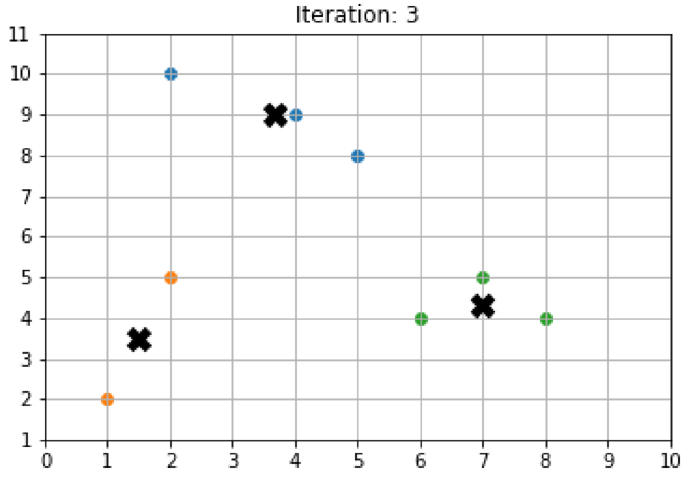
  C_3 = mean[(4, 9), (6, 4), (8, 4), (7, 5), (5, 8)] = (6.0, 6.0)


Iteration: 1

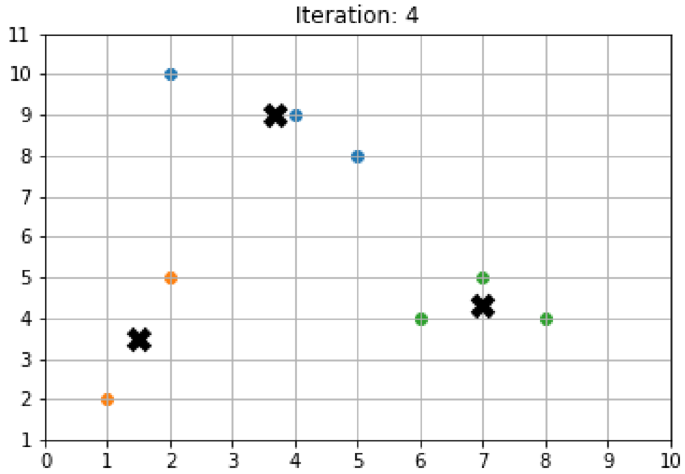
- **Iteration 2:**



- **Iteration 3:**



- **Iteration 4:**


Iteration: 4
11
10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10

2. Agglomerative Hierarchical

Solution:

1. Using MIN as an inter-cluster measure

Iteration 1:

	p_1	p_2	p_3	p_4	p_5	p_6
p_1						
p_2	0.2421					
p_3	0.2159	0.1523				
p_4	0.3677	0.1965	0.1581			
p_5	0.3418	0.1334	0.2846	0.2842		
p_6	0.2354	0.2530	**0.1020**	0.2195	0.3860	

Cluster 1: {3, 6}

- d({1},{3,6}) = min(d({1},{3}), d({1},{6})) = d({1},{3})
- d({2},{3,6}) = min(d({2},{3}), d({2},{6})) = d({2},{3})
- d({4},{3,6}) = min(d({4},{3}), d({4},{6})) = d({4},{3})
- d({5},{3,6}) = min(d({5},{3}), d({5},{6})) = d({5},{3})

Iteration 2:

	p_1	p_2	p_3	p_4	p_5	p_6
p_1						
p_2	0.2421					
p_3	0.2159	0.1523				
p_4	0.3677	0.1965	0.1581			
p_5	0.3418	**0.1334**	0.2846	0.2842		
p_6	~~0.2354~~	~~0.2530~~	~~0.1020~~	~~0.2195~~	~~0.3860~~	

Cluster 1: {3, 6}
Cluster 2: {2, 5}

- d({1},{2,5}) = min(d({1},{2}), d({1},{5})) = d({1},{2})
- d({4},{2,5}) = min(d({4},{2}), d({4},{5})) = d({4},{2})

- d({3,6},{2,5}) = min(d({3},{2}), d({3},{5}), d({6},{2}), d({6},{5})) = d({3},{2})

Iteration 3:

	p_1	p_2	p_3	p_4	p_5	p_6
p_1						
p_2	0.2421					
p_3	0.2159	**0.1523**				
p_4	0.3677	0.1965	0.1581			
p_5	~~0.3418~~	~~0.1334~~	~~0.2846~~	~~0.2842~~		
p_6	~~0.2354~~	~~0.2530~~	~~0.1020~~	~~0.2195~~	~~0.3860~~	

Cluster 1: {3, 6}
Cluster 2: {2, 5}
Cluster 3: {{2, 5}, {3, 6}}

- d{{2, 5, 3, 6}, {1}} = min(d{2, 1}, d{5, 1}, d{3, 1}, d{6, 1}) = d{3, 1}
- d{{2, 5, 3, 6}, {4}} = min(d{2, 4}, d{5, 4}, d{3, 4}, d{6, 4}) = d{3, 4}

Iteration 4:

	p_1	p_2	p_3	p_4	p_5	p_6
p_1						
p_2	~~0.2421~~					
p_3	0.2159	~~0.1523~~				
p_4	0.3677	~~0.1965~~	**0.1581**			
p_5	~~0.3418~~	~~0.1334~~	~~0.2846~~	~~0.2842~~		
p_6	~~0.2354~~	~~0.2530~~	~~0.1020~~	~~0.2195~~	~~0.3860~~	

Cluster 1: {3, 6}
Cluster 2: {2, 5}
Cluster 3: {{2, 5}, {3, 6}}
Cluster 4: {{2, 5, 3, 6}, {4}}

2. Using MAX as an inter-cluster measure

Iteration 1:

	p_1	p_2	p_3	p_4	p_5	p_6
p_1						
p_2	0.2421					
p_3	0.2159	0.1523				
p_4	0.3677	0.1965	0.1581			
p_5	0.3418	0.1334	0.2846	0.2842		
p_6	0.2354	0.2530	**0.1020**	0.2195	0.3860	

Cluster 1: {3, 6}

- d({1},{3,6}) = max(d({1},{3}), d({1},{6})) = d({1},{6})
- d({2},{3,6}) = max(d({2},{3}), d({2},{6})) = d({2},{6})
- d({4},{3,6}) = max(d({4},{3}), d({4},{6})) = d({4},{6})
- d({5},{3,6}) = max(d({5},{3}), d({5},{6})) = d({5},{6})

Iteration 2:

	p_1	p_2	p_3	p_4	p_5	p_6
p_1						
p_2	0.2421					
p_3	~~0.2159~~	~~0.1523~~				
p_4	0.3677	0.1965	~~0.1581~~			
p_5	0.3418	**0.1334**	~~0.2846~~	0.2842		
p_6	0.2354	0.2530	~~0.1020~~	0.2195	0.3860	

Cluster 1: {3, 6}
Cluster 2: {2, 5}

- d({1},{2,5}) = max(d({1},{2}), d({1},{5})) = d({1},{5})
- d({4},{2,5}) = max(d({4},{2}), d({4},{5})) = d({4},{5})
- d({3, 6},{2,5}) = max(d({3},{2}), d({3},{5}), d({6},{2}), d({6},{5})) = d({6},{5})

Iteration 3:

	p_1	p_2	p_3	p_4	p_5	p_6
p_1						
p_2	~~0.2421~~					
p_3	~~0.2159~~	~~0.1523~~				
p_4	0.3677	~~0.1965~~	~~0.1581~~			
p_5	0.3418	~~0.1334~~	~~0.2846~~	0.2842		
p_6	0.2354	~~0.2530~~	~~0.1020~~	**0.2195**	0.3860	

Cluster 1: {3, 6}
Cluster 2: {2, 5}
Cluster 3: {{3, 6}, {4}}

- d({1},{3,4,6}) = max(d({1},{3}), d({1},{4}), d({1},{6})) = d({1},{4})
- d({2,5},{3,4,6}) = max(d({2},{3}), d({2},{4}), d({2},{6}), d({5},{3}), d({5},{4}), d({5},{6})) = d({5},{6})

Iteration 4:

	p_1	p_2	p_3	p_4	p_5	p_6
p_1						
p_2	~~0.2421~~					
p_3	~~0.2159~~	~~0.1523~~				
p_4	0.3677	~~0.1965~~	~~0.1581~~			
p_5	**0.3418**	~~0.1334~~	~~0.2846~~	~~0.2842~~		
p_6	~~0.2354~~	~~0.2530~~	~~0.1020~~	~~0.2195~~	0.3860	

Cluster 1: {3, 6}
Cluster 2: {2, 5}
Cluster 3: {{3, 6}, {4}}
Cluster 4: {{2, 5}, {1}}

- d({1},{3,4,6}) = max(d({1},{3}), d({1},{4}), d({1},{6})) = d({1},{4})
- d({2,5},{3,4,6}) = max(d({2},{3}), d({2},{4}), d({2},{6}), d({5},{3}), d({5},{4}), d({5},{6})) = d({5},{6})

1. Using AVG as an inter-cluster measure

Iteration 1:

	p_1	p_2	p_3	p_4	p_5	p_6
p_1						
p_2	0.2421					
p_3	0.2159	0.1523				
p_4	0.3677	0.1965	0.1581			
p_5	0.3418	0.1334	0.2846	0.2842		
p_6	0.2354	0.2530	**0.1020**	0.2195	0.3860	

Cluster 1: {3, 6}

- d({1},{3,6}) = avg[d({1},{3}) + d({1},{6})] = 0.2256
- d({4},{3,6}) = avg[d({4},{3}) + d({4},{6})] = 0.1888
- d({2},{3,6}) = avg[d({2},{3}) + d({2},{6})] = 0.2026
- d({5},{3,6}) = avg[d({5},{3}) + d({5},{6})] = 0.3353

Iteration 2:

	p_1	p_2	p_3, p_6	p_4	p_5
p_1					
p_2	0.2421				
p_3, p_6	0.2256	0.2026			
p_4	0.3677	0.1965	0.1888		
p_5	0.3418	**0.1334**	0.3353	0.2842	

Cluster 1: {3, 6}
Cluster 1: {2, 5}

- d({1},{2,5}) = avg[d({1},{2}) + d({1},{5})] = 0.2919
- d({3,6},{2,5}) = avg[d({3,6},{2}) + d({3,6},{5})] = 0.2837
- d({4},{2,5}) = avg[d({4},{2}) + d({4},{5})] = 0.2404

Iteration 3:

	p_1	p_2, p_5	p_3, p_6	p_4
p_1				
p_2, p_5	0.2919			
p_3, p_6	0.2256	0.2185		
p_4	0.3677	0.2404	**0.1888**	

Cluster 1: {3, 6}
Cluster 2: {2, 5}
Cluster 3: {{3, 6}, {4}}

- d({1},{{3, 6}, {4}}) = avg[d({1},{3, 6}) + d({1},{4})] = 0.2967
- d({2, 5},{{3, 6}, {4}}) = avg[d({2, 5},{3, 6}) + d({2, 5},{4})] = 0.2295

Iteration 4:

	p_1	p_2, p_5	p_3, p_6
p_1			
p_2, p_5	0.2919		
p_3, p_4, p_6	0.2256	**0.2185**	

Cluster 1: {3, 6}
Cluster 2: {2, 5}
Cluster 3: {{3, 6}, {4}}
Cluster 4: {{3, 4, 6}, {2, 5}}

Iteration 5:

	p_1	p_2, p_3, p_4, p_5, p_6
p_1		
p_2, p_3, p_4, p_5, p_6	**0.2919**	

Cluster 1: {3, 6}
Cluster 2: {2, 5}
Cluster 3: {{3, 6},{4}}
Cluster 4: {{3, 4, 6}, {2, 5}}
Cluster 5: {1,{{3, 4, 6}, {2, 5}}}

3. DBSCAN

Solution: