CSE 40647/60647: Data Science	**Fall 2017**
Homework 5	
Handed Out: November 09, 2017	*Due: November 28, 2017 11:59 pm*

1 General Instructions

- This assignment is due at 11:59 PM on the due date.
- We will be using Sakai (https://sakailogin.nd.edu/portal/site/FA17-CSE-40647-CX-01) for collecting this assignment. Contact TA if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The homework MUST be submitted in pdf format. You can handwrite trees/figures and scan them into PDF. Name your pdf file as YourNetid-HW4.pdf.
- Please use Piazza if you have questions about the homework. Also feel free to send TA emails and come to office hours.

2 Question 1 (20 points)

Initialize with two centroids, (6, 4) and (6, 5). Use Manhattan distance as the distance metric. Please use K-Means to find two clusters.

Solutions: Given two centroids (6, 4) and (6, 5), use Manhattan distance, we can have the following table. Manhattan distance for X_1 to centroid (6, 4) is $|5-6|+|7-4|=4$.

Team	X-2016	Y-2017	(6, 4)	(6, 5)
X_1	5	7	4	3
X_2	6	7	3	2
X_3	2	8	8	7
X_4	7	8	5	4
X_5	8	4	2	3
X_6	6	4	0	1
X_7	7	3	2	3
X_8	6	3	1	2
X_9	5	2	3	4
X_{10}	4	3	3	4

Then we can tell that X_1, X_2, X_3 and X_4 are closer to centroid (6, 5) and X_5, X_6, X_7, X_8, X_9 and X_{10} are closer to centroid (6, 4). Then we can move centroid (6, 5) to (5, 7.5) and move centroid (6, 4) to (6, 3.17).

Run Manhattan distance another time.

Team	X-2016	Y-2017	(5, 7.5)	(6, 3.17)
X_1	5	7	4.83	0.5
X_2	6	7	3.83	1.5
X_3	2	8	8.83	3.5
X_4	7	8	5.83	2.5
X_5	8	4	2.83	6.5
X_6	6	4	0.83	4.5
X_7	7	3	1.17	6.5
X_8	6	3	0.17	5.5
X_9	5	2	2.17	5.5
X_{10}	4	3	2.17	5.5

There is no relocation occur after we move the centroids, which means current partitions are our final clusters. Cluster 1: X_5, X_6, X_7, X_8, X_9, X_{10} and Cluster 2: X_1, X_2, X_3, X_4.

3 Question 2 (20 points)

Initialize with two centroids, (6, 4) and (6, 5). Use Euclidean distance as the distance metric. Please use K-Means to find two clusters.

Solutions: Given two centroids (6, 4) and (6, 5), use Manhattan distance, we can have the following table. Euclidean distance for X_1 to centroid (6, 4) is $sqrt((5-6)^2 + (7-4)^2) = 3.16$.

Team	X-2016	Y-2017	(6, 4)	(6, 5)
X_1	5	7	3.16	2.24
X_2	6	7	3	2
X_3	2	8	5.66	5
X_4	7	8	4.12	3.16
X_5	8	4	2	2.24
X_6	6	4	0	1
X_7	7	3	1.41	2.24
X_8	6	3	1	2
X_9	5	2	2.24	3.16
X_{10}	4	3	2.24	2.83

Then we can tell that X_1, X_2, X_3 and X_4 are closer to centroid (6, 5) and X_5, X_6, X_7, X_8, X_9 and X_{10} are closer to centroid (6, 4). Then we can move centroid (6, 5) to (5, 7.5) and move centroid (6, 4) to (6, 3.17).

Run Euclidean distance another time.

Team	X-2016	Y-2017	(6, 3.17)	(5, 7.5)
X_1	5	7	3.96	0.5
X_2	6	7	3.83	1.12
X_3	2	8	6.27	3.04
X_4	7	8	4.93	2.06
X_5	8	4	2.17	4.61
X_6	6	4	0.83	3.64
X_7	7	3	1.01	4.92
X_8	6	3	0.17	4.61
X_9	5	2	1.54	5.5
X_{10}	4	3	2.01	4.61

There is no relocation occur after we move the centroids, which means current partitions are our final clusters. Cluster 1: X_5, X_6, X_7, X_8, X_9, X_{10} and Cluster 2: X_1, X_2, X_3, X_4.

4 Question 3 (20 points)

Initialize with two centroids, (8, 7) and (2, 6). Use Manhattan distance as the distance metric. Please use K-Means to find two clusters.

Solutions: Given two centroids (8, 7) and (2, 6), use Manhattan distance, we can have the following table. Manhattan distance for X_1 to centroid (8, 7) is $|5-8|+|7-7|=3$.

Team	X-2016	Y-2017	(8, 7)	(2, 6)
X_1	5	7	3	4
X_2	6	7	2	5
X_3	2	8	7	2
X_4	7	8	2	7
X_5	8	4	3	8
X_6	6	4	5	6
X_7	7	3	5	8
X_8	6	3	6	7
X_9	5	2	8	7
X_{10}	4	3	8	5

Then we can tell that X_1, X_2, X_4, X_5, X_6, X_7 and X_8 are closer to centroid (8, 7) and X_3, X_9 and X_{10} are closer to centroid (2, 6). Then we can move centroid (8, 7) to (6.43, 5.14) and move centroid (2, 6) to (3.67, 4.33).

Run Manhattan distance another time.

Team	X-2016	Y-2017	(6.43, 5.14)	(3.67, 4.33)
X_1	5	7	3.29	4
X_2	6	7	2.29	5
X_3	2	8	7.29	5.34
X_4	7	8	3.43	7
X_5	8	4	2.71	4.66
X_6	6	4	1.57	2.66
X_7	7	3	2.71	4.66
X_8	6	3	2.57	3.66
X_9	5	2	4.57	3.66
X_{10}	4	3	4.57	1.66

There is no relocation occur after we move the centroids, which means current partitions are our final clusters. Cluster 1: X_1, X_2, X_4, X_5, X_6, X_7, X_8 and Cluster 2: X_3, X_9, X_{10}.

5 Question 4 (20 points)

Suppose we initialize with two medoids, (2, 8) and (8, 4). Use Euclidean distance as the distance metric. In K-Medoids clustering, given a non-medoid (5,7), do we swap the medoid (2,8) with (5,7)?

Solutions: Given two medoids (2, 8) and (8, 4), use Euclidean distance, we can have the following table.

Team	X-2016	Y-2017	(2, 8)	(8, 4)
X_1	5	7	3.16	4.24
X_2	6	7	4.12	3.60
X_3	2	8	0	7.21
X_4	7	8	5	4.12
X_5	8	4	7.21	0
X_6	6	4	5.66	2
X_7	7	3	7.07	1.41
X_8	6	3	6.40	2.24
X_9	5	2	6.71	3.61
X_{10}	4	3	5.39	4.12

Then $SSE = 3.16^2 + 3.60^2 + 4.12^2 + 2^2 + 1.41^2 + 2.24^2 + 3.61^2 + 4.12^2 = 80.9322$. If we swap the medoid (2, 8) to (5, 7), then we can have the following table.

Team	X-2016	Y-2017	(5, 7)	(8, 4)
X_1	5	7	0	4.24
X_2	6	7	1	3.60
X_3	2	8	3.16	7.21
X_4	7	8	2.24	4.12
X_5	8	4	4.24	0
X_6	6	4	3.16	2
X_7	7	3	4.47	1.41
X_8	6	3	4.12	2.24
X_9	5	2	5	3.61
X_{10}	4	3	4.12	4.12

Then $SSE = 1^2 + 3.16^2 + 2.24^2 + 2^2 + 1.41^2 + 2.24^2 + 3.61^2 + 4.12^2 = 57.0154$. $S = 57.0154 - 80.9322 = -23.9168 < 0$.

Therefore, **Yes, we swap them**.

6 Question 5 (20 points)

Suppose the original two features are x and y. We use a kernel function to generate three new features: x^2, xy and y^2. Now we initialize with two centroids, (6, 4) and (6, 5), that are now (36, 24, 16) and (36, 30, 25). Use Manhattan distance as the distance metric in the new feature space. Please use Kernel K-Means to find two clusters.

Solutions: Given two centroids (6, 4) and (6, 5), use Manhattan distance, we can have the following table..

Team	X-2016	Y-2017	X^2	XY	Y^2	(36, 24, 16)	(36, 30, 25)
X_1	5	7	25	35	49	55	40
X_2	6	7	36	42	49	51	36
X_3	2	8	4	16	64	88	85
X_4	7	8	49	56	64	93	78
X_5	8	4	64	32	16	36	39
X_6	6	4	36	24	16	0	15
X_7	7	3	49	21	9	23	38
X_8	6	3	36	18	9	13	28
X_9	5	2	25	10	4	37	52
X_{10}	4	3	16	12	9	39	54

Then we can tell that X_1, X_2, X_3 and X_4 are closer to centroid (36, 30, 25) and X_5, X_6, X_7, X_8, X_9 and X_{10} are closer to centroid (36, 24, 16). As for kernel based k-means clusters, we mapped the data first and then use the new label to do clustering, Therefore, we can move centroid (36, 30, 25) to (28.5, 37.25, 56.5) and move centroid (36, 24, 16) to (37.67, 19.5, 10.5).

Run Manhattan distance another time.

Team	X-2016	Y-2017	X^2	XY	Y^2	(37.67, 19.5, 10.5)	(28.5, 37.25, 56.5)
X_1	5	7	25	35	49	66.67	13.25
X_2	6	7	36	42	49	62.67	19.75
X_3	2	8	4	16	64	90.67	53.25
X_4	7	8	49	56	64	101.33	46.75
X_5	8	4	64	32	16	44.33	81.25
X_6	6	4	36	24	16	11.67	61.25
X_7	7	3	49	21	9	14.33	84.25
X_8	6	3	36	18	9	4.67	74.25
X_9	5	2	25	10	4	28.67	83.25
X_{10}	4	3	16	12	9	30.67	85.25

There is no relocation occur after we move the centroids, which means current partitions are our final clusters. Cluster 1: X_5, X_6, X_7, X_8, X_9, X_{10} and Cluster 2: X_1, X_2, X_3, X_4.