Weekly Homework 9

Ava Chong
CS 1675: Intro to Machine Learning

April 4, 2019

Problem 1. K-means clustering

(a) Calculate the euclidean distance between points
The points we are looking at are (0,0), (0,5), (6,7), and (7,0).

For means (0,0) and (7,0)

Point A	Point B	Distance
(0,0)	(0,0)	0
(0,0)	(0,5)	5
(0,0)	(6,7)	9.219544
(0,0)	(7,0)	7
(7,0)	(0,0)	7
(7,0)	(0,5)	8.602325
(7,0)	(6,7)	7.071068
(7,0)	(7,0)	0

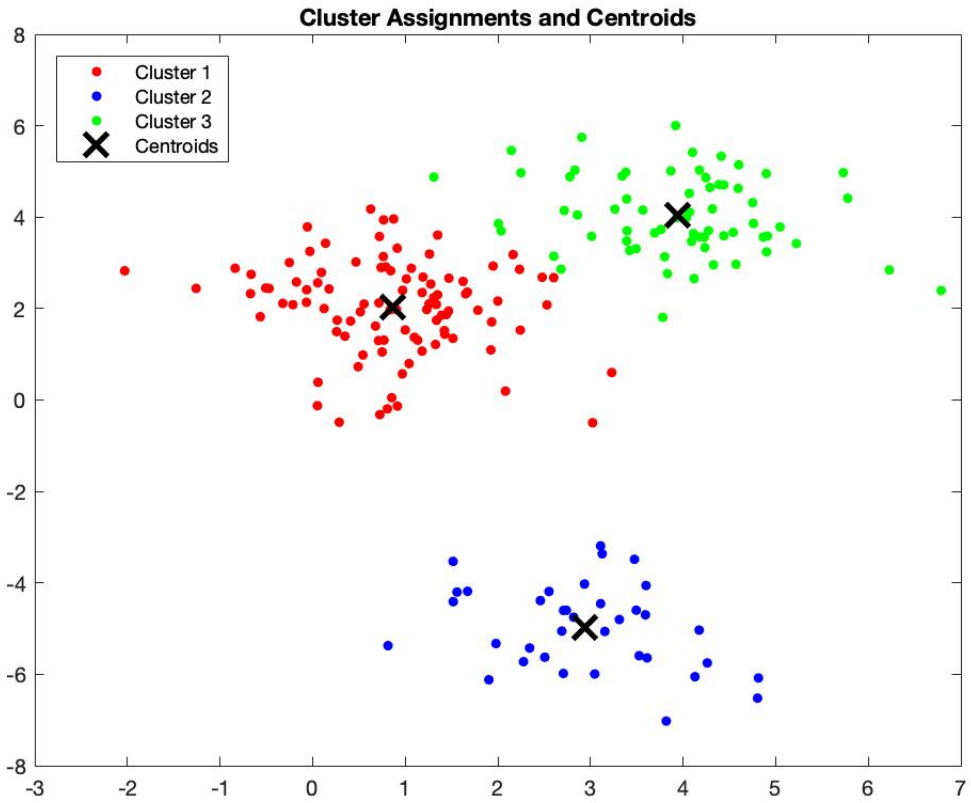
We cluster the data based on the the smallest distance between the mean. The mean that is closet to the data point is the group that data point gets put in.
Points (0,0) and (0,5) belong to mean point group (0,0), group 1.
Points (6,7) and (7,0) belong to mean point group (7,0), group 2.

(b) Using the means (3,3) and (7,0)

Point A	Point B	Distance
(3,3)	(0,0)	4.242
(3,3)	(0,5)	3.605551
(3,3)	(6,7)	5
(3,3)	(7,0)	5
(7,0)	(0,0)	7
(7,0)	(0,5)	8.602325
(7,0)	(6,7)	7.071068
(7,0)	(7,0)	0

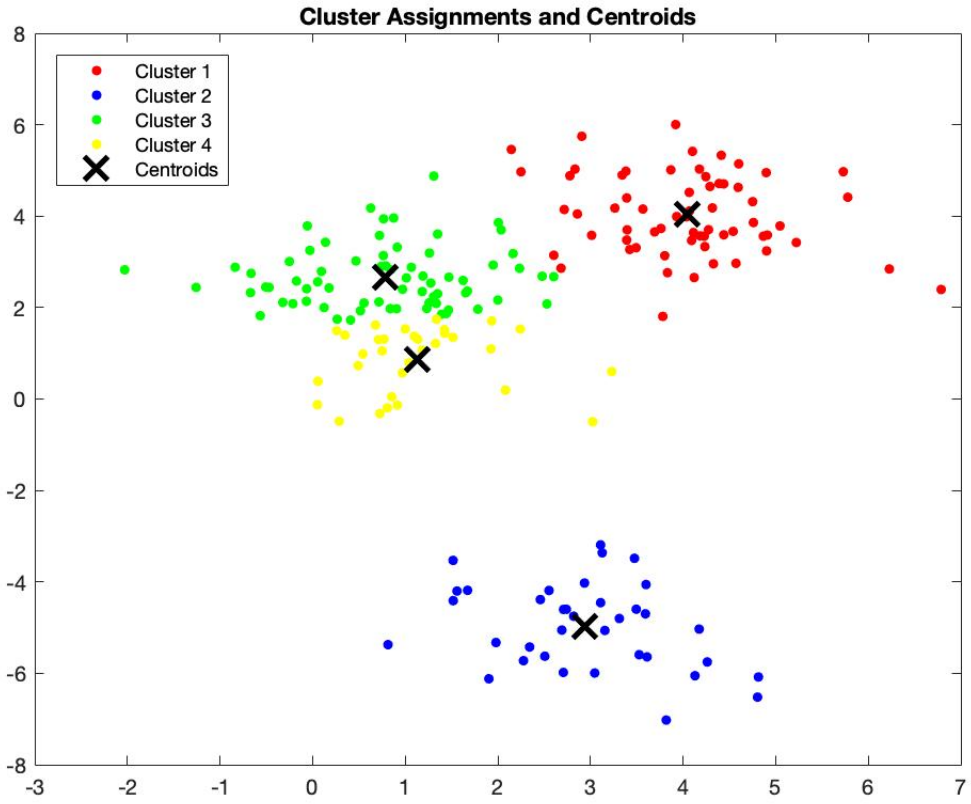
Points (0,0), (0,5) and (6,7) belong to mean point group (3,3), group 1.
Point (7,0) belong to mean point group (7,0), group 2.

Problem 2. K-means clustering experiments
(*a*)



Sizes of the three groups:
Cluster 1 = 98
Cluster 2 = 36
Cluster 3 = 66

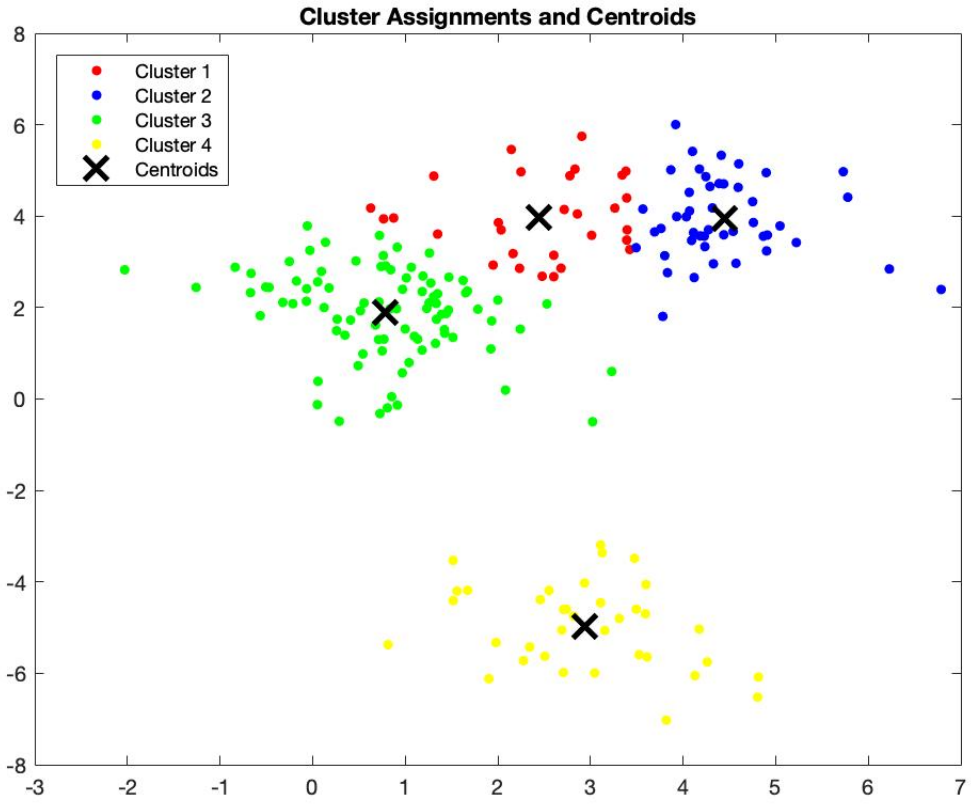
(*b*)



Sizes of the three groups:
Cluster 1 = 63
Cluster 2 = 36
Cluster 3 = 69
Cluster 4 = 32

Means found buy kmeans
C1 = 4.044, 4.033
C2 = 2.940, -4.969
C3 = 0.778, 2.659
C4 = 1.129, 0.863

(*c*)



When the program is re-run, the centers and clusters change.

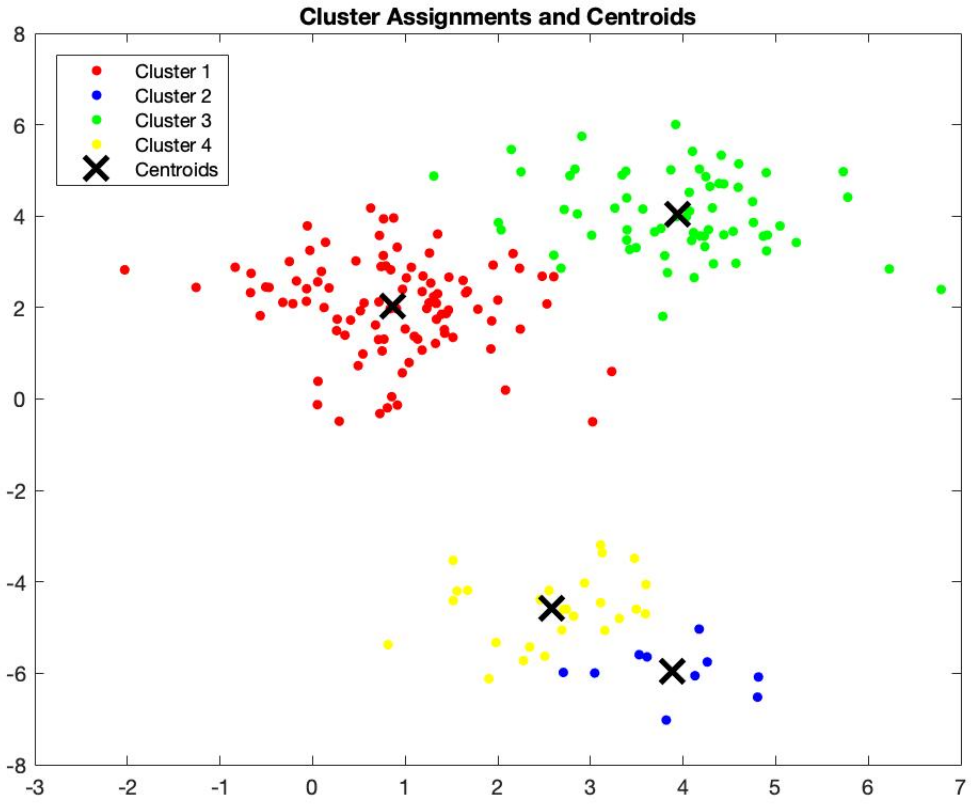
(*d*) To find the best clustering we should use the sum of all the distances from points to the mean of their cluster. This will punish poor groupings and reward better clustering of the data.
Sum of Square Error mathematical formula is defined as:

$$\text{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

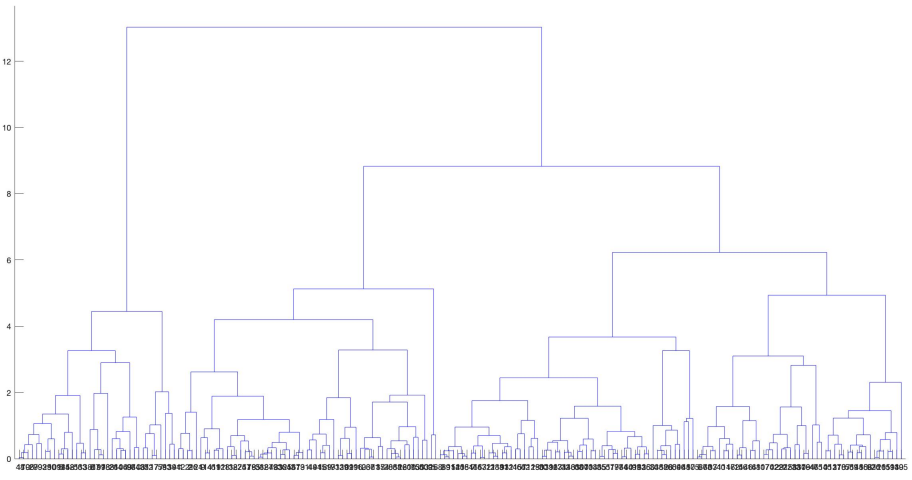
(*e*)

	1	2	3	4
1	63	61	36	40
2	46	36	29	89
3	98	26	10	66
4	98	66	10	26
5	36	37	95	32
6	47	36	61	56
7	39	36	65	60
8	98	66	10	26
9	61	61	36	42
10	99	25	64	12
11	60	36	50	54
12	36	26	53	85
13	53	85	26	36
14	98	10	26	66
15	28	97	36	39
16	26	36	85	53
17	26	98	10	66
18	36	63	63	38
19	40	36	61	63
20	26	36	85	53
21	98	66	10	26
22	47	36	61	56
23	49	52	36	63
24	69	63	36	32
25	63	61	36	40
26	36	85	53	26
27	66	26	98	10
28	51	36	61	52
29	97	36	28	39
30	63	36	51	50

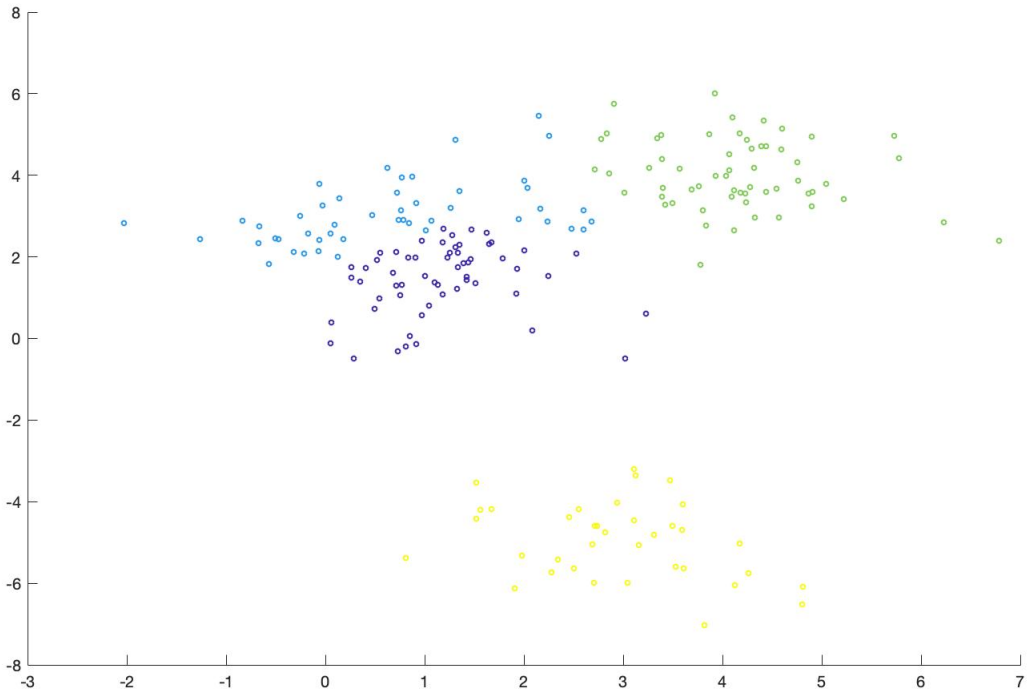
The best scatter-plot:



Problem 3. Hierarchical clustering experiments
(*a*) Dendrogram of the full cluster tree with 4 clusters



(*b*) Scatterplot



The scatterplot using matlab's cluster functions looks similar to the scatterplots we had before.