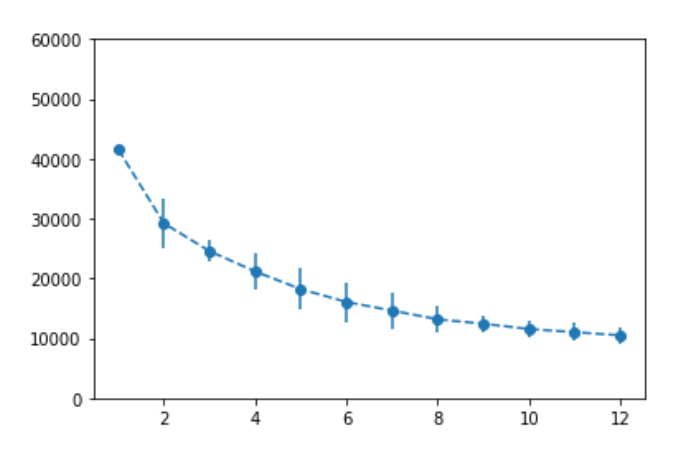
James Ho  
CISC Data Mining

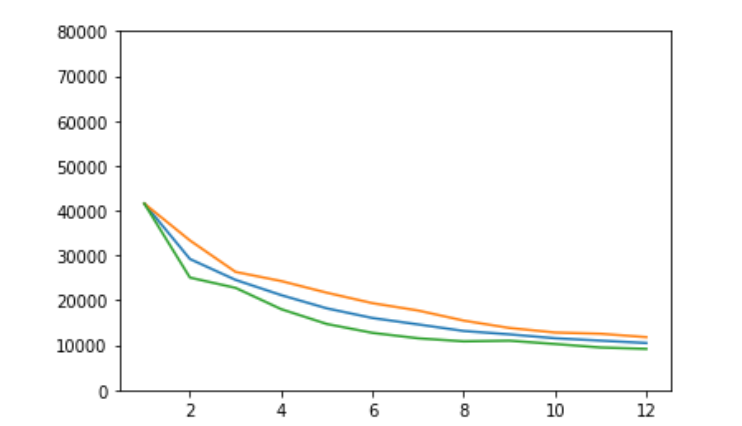
Question 1

Part A

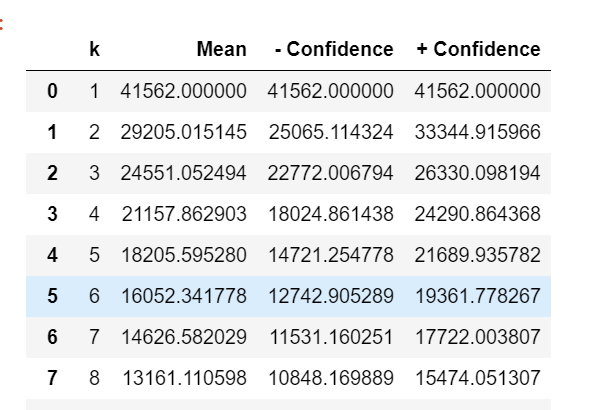
Graph with Error margins



Same graph with two lines instead of error margins.



Part B



Part C

As the value of K increases and there are more clusters that are set we reduce the value of SSE and it approaches zero one K equals the number of points N. The problem for this is when there is no expert opinion we do not know how many clusters to set and the elbow method may fail as we may not be able to infer the amount of clusters from SSE.

Part D

Another way one can evaluate a centroids clustering model is to use the scatter method that was done bellow for another data set. Alternatively we can also use other non centroid clustering methods in conjunction as a means to compare and see if a similar result was obtain to see the legitimacy of ones own result.

Question 2:

{ 0, 4, 5, 20, 25, 39, 43, 44 }

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 4 | 5 | 20 | 25 | 39 | 43 | 44 |
| 0 | 0 | 4 | 5 | 20 | 25 | 39 | 43 | 40 |
| 4 |  | 0 | 1 | 16 | 21 | 35 | 39 | 36 |
| 5 |  |  | 0 | 15 | 20 | 35 | 38 | 39 |
| 20 |  |  |  | 0 | 5 | 19 | 23 | 24 |
| 25 |  |  |  |  | 0 | 14 | 18 | 14 |
| 39 |  |  |  |  |  | 0 | 4 | 5 |
| 43 |  |  |  |  |  |  | 0 | 1 |
| 44 |  |  |  |  |  |  |  | 0 |

{0, (4,5), 20, 25, 39, 43, 44}

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 4,5 | 20 | 25 | 39 | 43 | 44 |
| 0 | 0 | 4 | 20 | 25 | 39 | 43 | 44 |
| 4,5 |  | 0 | 16 | 21 | 35 | 39 | 40 |
| 20 |  |  | 0 | 5 | 19 | 23 | 24 |
| 25 |  |  |  | 0 | 14 | 18 | 19 |
| 39 |  |  |  |  | 0 | 4 | 5 |
| 43 |  |  |  |  |  | 0 | 1 |
| 44 |  |  |  |  |  |  | 0 |

{0, (4,5), 20, 25, 39, (43 , 44)}

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 0 | (4,5) | 20 | 25 | 39 | (43, 44) |
| 0 | 0 | 4 | 20 | 25 | 39 | 43 |
| 4,5 |  | 0 | 15 | 20 | 34 | 38 |
| 20 |  |  | 0 | 5 | 19 | 33 |
| 25 |  |  |  | 0 | 14 | 18 |
| 39 |  |  |  |  | 0 | 4 |
| 43, 44 |  |  |  |  |  | 0 |

{(0, (4,5)), 20, 25, 39, (43 , 44)}

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0, (4,5) | 20 | 25 | 39 | 43,44 |
| 0, (4,5) | 0 | 15 | 20 | 34 | 38 |
| 20 |  | 0 | 5 | 19 | 23 |
| 25 |  |  | 0 | 14 | 18 |
| 39 |  |  |  | 0 | 4 |
| 43,44 |  |  |  |  | 0 |

{(0, (4,5)), 20, 25, (39, (43 , 44)) }

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (0, (4,5)) | 20 | 25 | (39, (43 , 44)) |
| (0, (4,5)) | 0 | 15 | 20 | 35 |
| 20 |  | 0 | 5 | 19 |
| 25 |  |  | 0 | 14 |
| (39, (43 , 44)) |  |  |  | 0 |

{(0, (4,5)), (20, 25) , (39, (43 , 44)) }

|  |  |  |  |
| --- | --- | --- | --- |
|  | (0, (4,5)) | (20, 25) | (39, (43 , 44)) |
| (0, (4,5)) | 0 | 15 | 34 |
| (20, 25) |  | 0 | 14 |
| (39, (43 , 44)) |  |  | 0 |

{(0, (4,5)), ((20, 25) , (39, (43 , 44))) }

{((0, (4,5)), ((20, 25) , (39, (43 , 44)))) }

Dendrogram Graph

{ 0, 4, 5, 20, 25, 39, 43, 44 }

{0, (4,5), 20, 25, 39, 43, 44}

| - |

4 5

{0, (4,5), 20, 25, 39, (43 , 44)}

| - | | - |

4 5 43 44

{(0, (4,5)), 20, 25, 39, (43 , 44)}

| --- |

| - | | | - |

4 5 0 43 44

{(0, (4,5)), 20, 25, (39, (43 , 44)) }

| --- | | --- |

| - | | | | - |

4 5 0 39 43 44

{(0, (4,5)), (20, 25) , (39, (43 , 44)) }

| --- | | --- |

| | - | | - | | | - |

0 4 5 20 25 39 43 44

{(0, (4,5)), ((20, 25) , (39, (43 , 44))) }

| ----- |

| --- | | | --- |

| | - | | - | | | - |

0 4 5 20 25 39 43 44

{((0, (4,5)), ((20, 25) , (39, (43 , 44)))) }

| ---------------- |

| | ----- |

| --- | | | --- |

| | - | | - | | | - |

0 4 5 20 25 39 43 44

Two clusters:

C1 = {(0, (4,5)), }

C2 = {((20, 25) , (39, (43 , 44))) }

Question 3:

C1 = {(1,1), (2,2), (3,3)}

C2 = {(5,2), (6,2), (7,2), (8,2), (9,2)}

Mean vectors

Mean of Vector 1:

Part A:

Sum of C1 times 1/3: 1 + 2 + 3, 1 + 2 + 3= (6, 6) \* 1/ 3 = (2, 2)

Sum of C2 times 1/5: 5 + 6 + 7 + 8 + 9, 2 + 2 + 2 + 2 +2 = (7, 2)

Part B

Total Sum times 1/8 = 1 + 2 + 3 + 5 + 6 + 7 + 8 + 9, 1 + 2 + 3 + 2 + 2 + 2 + 2 +2 = (5.125, 2)

Part C

S1

(X – mean) times (X – mean).transposed

(1,1) – (2,2) = (-1,-1) \*(-1,-1).transposed = 1 1

1 1

(2,2) – (2,2) = (0,0) \*(0,0).transposed = 0 0

0 0

(3,3) – (2,2) = (1,1) \*(1,1).transposed = 1 1

1 1

S1 = 1 1 + 0 0 + 1 1 = 2 2

1 1 0 0 + 1 1 2 2

S2

(X – mean) times (X – mean).transposed

(5,2) – (7,2) = (-2,0) \*(-2,-0).transposed = 4 0

0 0

(6,2) – (7,2) = (1,0) \*(1,0).transposed = 1 0

0 0

(7,2) – (7,2) = (0,0) \*(0,0).transposed = 0 0

0 0

(8,2) – (7,2) = (1,0) \*(1,0).transposed = 1 0

0 0

(9,2) – (7,2) = (2,0) \*(2,-0).transposed = 4 0

0 0

S2 = 4 0 + 1 0 + 0 0 + 1 0 + 4 0 = 10 0

0 0 0 0 0 0 0 0 0 0 0 0

Part D

Sw = 10 0 + 2 2 = 12 2

0 0 2 2 2 2

Part E

Sb = sb1 + sb2

Sb1 = (2,2) – (5.1,2) = (-3.125, 0) \* (3.125, 0 ).transposed \* 3 = 9.1 \* 3 = 27.3 0

0 0

Sb2 = (7, 2) – (5.1,2) = (1.9, 0) \* (1.9, 0).transposed \* 5 = 18.05 0

0 0

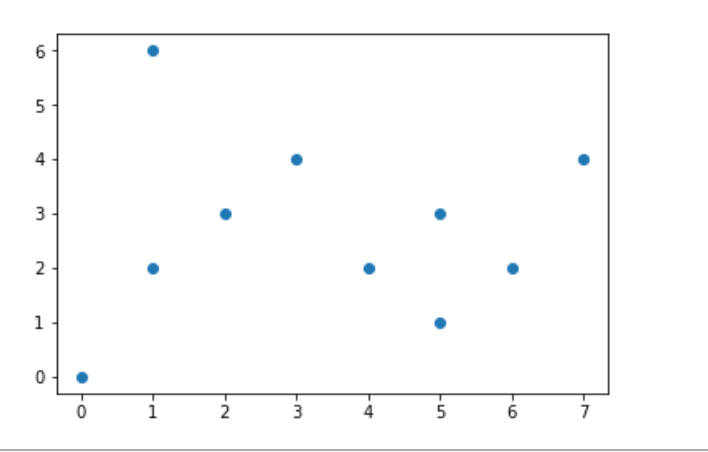
Sb = Sb1 + sb2 = 45 0

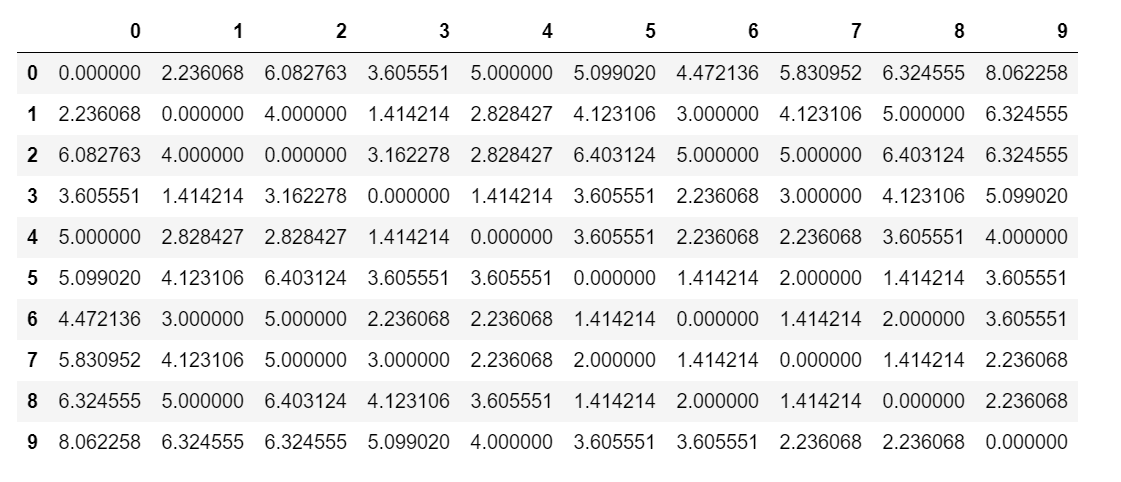
0 0

Part F

Tr(Sb)/Tr(Sw) = 45/14 = 3.21

Question 4:





Part A

Cluster A: {(1,2), (2,3), (3,4)}

Cluster B: {(3,4), (5,1), (4,2), (5,3), (6,2)}

Part B

For Cluster A

(1,2), connected to (2,3)

(2,3), connected to (3,4)

Therefore the points above are density connected but not all values are density reachable from each other.

For Cluster B

(3,4), connected to (5,1), (3,4) connected to(6,2)},

(4,2), connected to (5,3), and (5,3) connected to (6,2)

Therefore the points above are density connected to one another but not all of them are density reachable from each other.

Part C

The outliers are as follows:

(0,0), (1,6), (7,4)

Question 5:

|  |  |  |  |
| --- | --- | --- | --- |
| Instance | Predicted Score | Predicted Label | True (Ground Truth) |
| 1 | 0.95 | P | P |
| 2 | 0.85 | P | N |
| 3 | 0.78 | P | P |
| 4 | 0.66 | P | P |
| 5 | 0.60 | P | N |
| 6 | 0.55 | P | P |
| 7 | 0.43 | N | N |
| 8 | 0.42 | N | N |
| 9 | 0.41 | N | N |
| 10 | 0.40 | N | P |

Confusion Matrix:

Predicted

TP = 4 FP = 2 Actual

FN = 1 TN = 3

Accuracy: 7/10 = 70%

Precision: 4/6 = 66.6%

Recall: 4/5 = 80%

F1 Score: 8/11 = 72.27%

Specificity 3/5 = 60%