Sushma Kudlur Nirvanappa   [Student ID: 9079813292]

Question 1: PCA [25 points]

Consider a data set with 20 two-dimensional points of the form (i, i),(i, i + 1) for i = 1, ..., 10.

1. Given the points (i,i), (i, i+1) for i = 1, ..., 10. Below represent the data points

Mean of x-coordinates

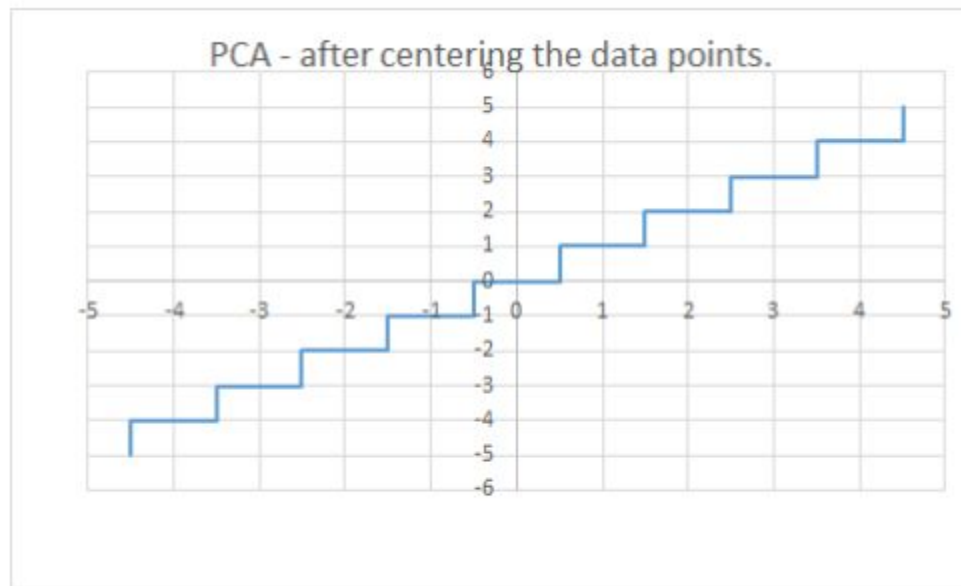$X(mean) = (1 + 2 + ... + 10 + 1 + 2 + ... + 10) / 20 = \mathbf{5.5}$

Mean of y-coordinates

$Y(mean) = (1 + 2 + ... + 10 + 2 + 3 + ... + 11) / 20 = \mathbf{6}$

Now, for computing the centered data points, we will shift all points using

X(i-centered) = X(i) − X(mean)
Y(i-centered) = Y(i) − Y(mean), where (Xi, Yi) are the i-th points in the given space.

Using above, the X(centered) data points can be represented as follow:

PCA - after centering the data points.

2. S = $(1/(n-1))X^{T}X$

For our centered data, X(centered) as computed above can be used.
n = 20, Upon computation we get below Covariance matrix.
= 1/(20-1) *
(165.000000 165.000000
165.000000 170.000000)

= (8.68421052 8.68421052
8.68421052   8.9473683)

3. Using eigen computation, eigenvalues and eigenvectors of covariance matrix are:

Eigenvalue λ_1 = 0.131
Eigenvector V_1 : [-0.712
                -0.702]

Eigenvalue λ_2 = 17.501
Eigenvector V_2 : [0.702
                - 0.712]

4. Compute the new two-dimensional coordinates of CENTERED points { corresponding to the old (i; i) and (i; i+1) { by projecting them onto v1; v2. (Hint: if you only keep the rst dimension, it would be a maximum spread projection of the original data set.)

Now, projection of a point X on a vector V is given by, X(centered) V^T

And, projection of any given points on these two vectors would be (X(centered) $V_1{}^T$, X(centered) $V_2{}^T$)

Using above for all the centered points, we can compute the projected values to be the following matrix.

original (x,y)	projected x	projected y
(1,1)	-6.71999989	-0.3026573662
(2,2)	-5.305826903	-0.3133701832
(3,3)	-3.891653917	-0.3240830001
(4,4)	-2.47748093	-0.334795817
(5,5)	-1.063307944	-0.3455086339
(6,6)	0.3508650424	-0.3562214508
(7,7)	1.765038029	-0.3669342678
(8,8)	3.179211015	-0.3776470847
(9,9)	4.593384002	-0.3883599016
(10,10)	6.007556988	-0.3990727185
(1,2)	-6.007556988	0.3990727185
(2,3)	-4.593384002	0.3883599016
(3,4)	-3.179211015	0.3776470847
(4,5)	-1.765038029	0.3669342678
(5,6)	-0.3508650424	0.3562214508
(6,7)	1.063307944	0.3455086339
(7,8)	2.47748093	0.334795817
(8,9)	3.891653917	0.3240830001
(9,10)	5.305826903	0.3133701832
(10,11)	6.71999989	0.3026573662

Question 2: Resolution [25 points]

Given the knowledge base

$$p \Rightarrow (q \Rightarrow r)$$

use resolution to prove the query

$$(p \wedge q) \Rightarrow (q \Rightarrow r)$$

Be sure to show what you convert to CNF and how (do not skip steps), and how you perform each resolution step.

Solution:

Knowledge Base, converting to CNF form: $p \Rightarrow (q \Rightarrow r)$

$p \Rightarrow (q \Rightarrow r)$

$\sim p \vee (q \Rightarrow r)$ // Removing implication

$\sim p \vee \sim q \vee r$ // Removing implication

Query (β), converting to CNF form: $(p \wedge q) \Rightarrow (q \Rightarrow r)$

$(p \wedge q) \Rightarrow (q \Rightarrow r)$

$\sim(p \wedge q) \vee (q \Rightarrow r)$ // Removing implication

$\sim(p \wedge q) \vee \sim q \vee r$ // Removing implication

$\sim p \vee \sim q \vee r$ //Moving negation inwards

Now, for resolution we add negation of Query (β),

So, $\sim\beta = \sim(\sim p \vee \sim q \vee r)$

$p \wedge q \wedge \sim r$ // Moving negation inwards

The components we have for to check resolution are:

A: $\sim p \vee \sim q \vee r$ // from KB

B1: p // from query

B2: q // from query

B3: $\sim r$ // from query

We can clearly see that conjunction of all above statements gives empty.

$\sim p \vee \sim q \vee r \wedge p \wedge q \wedge \sim r$

$(\sim p \wedge p) \vee (\sim q \wedge q) \vee (r \wedge \sim r)$

Everything will be nullified, thus proves the resolution. Thus we can say KB entails query.

Question 3: Hierarchical Clustering [25 points]

Consider the following six major cities. In the US: Madison, Seattle, Boston; and in Canada: Vancouver, Winnipeg, Montreal. For the purpose of this question ignore the curvature of the Earth, and compute the Euclidean distance. Suppose the cities are located at the following coordinates:
city coordinate
Madison (-89, 43)
Seattle (-122, 48)
Boston (-71, 42)
Vancouver (-123, 49)
Winnipeg (-97, 50)
Montreal (-74, 46)

1. Use hierarchical clustering with complete linkage to produce TWO clusters by hand. Specifically, show the following in each iteration: (1) the closest pair of clusters; (2) the distance between them as defined by complete linkage; (3) all clusters at the end of that iteration.

Iteration 1:

| Closest Pair Of Cluster	Euclidean Distance (Complete Linkage)	Cluster at the end of iteration
[Vancouver (-123, 49), Seattle (-122, 48)]	$\sqrt{\|-123-(-122)\|^2 + \|49-48\|^2}$ $= 1.414$	(Vancouver, Seattle), (Madison), (Boston), (Winnipeg), (Montreal)

Iteration 2:

Closest Pair Of Cluster	Euclidean Distance (Complete Linkage)	Cluster at the end of iteration
[Montreal (-74, 46), Boston (-71, 42)]	$\sqrt{\|-74-(-71)\|^2 + \|46-42\|^2}$ $= \sqrt{9+16} = 5$	(Vancouver, Seattle), (Montreal, Boston), (Madison), (Winnipeg)

Iteration 3:

Closest Pair Of Cluster	Euclidean Distance (Complete Linkage)	Cluster at the end of iteration
[Winnipeg (-97, 50), Madison (-89, 43)]	$\sqrt{\|-89-(-97)\|^2+\|50-43\|^2}$ $=\sqrt{113}=10.63$	(Vancouver, Seattle), (Montreal, Boston), (Madison,Winnipeg)

Iteration 4:

Closest Pair Of Cluster	Euclidean Distance (Complete Linkage)	Cluster at the end of iteration
[[Winnipeg (-97, 50), Madison (-89, 43)] , [Montreal (-74, 46), Boston (-71, 42)]]	$\sqrt{\|-97-(-71)\|^2+\|50-42\|^2}$ $=\sqrt{740}=27.2$	(Vancouver, Seattle), (Madison, Winnipeg, Montreal, Boston)
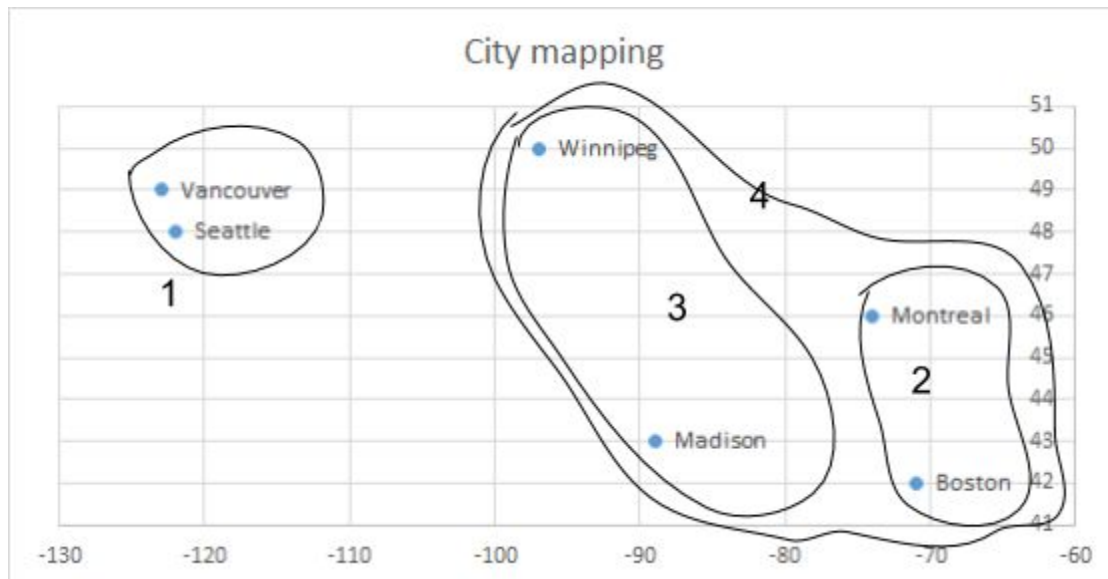
After fourth iteration it forms two different clusters.
1. One cluster is formed which is shown in Iteration-1 Output.
   Output: [Vancouver (-123, 49), Seattle (-122, 48)]
2. Another cluster is shown in Iteration-4 output.
   Output: [[Winnipeg (-97, 50), Madison (-89, 43)] , [Montreal (-74, 46), Boston (-71, 42)]]

2. Now repeat the above question, but with the following constraint: at no point should a US city and a Canadian city be put in the same cluster. Equivalently, whenever the complete linkage between two clusters is due to two cities in different countries, treat the two clusters as infinity apart, regardless of what other cities are in those two clusters. Show the same (1)(2)(3) as above in each iteration.

Iteration 1:

| Closest Pair Of Cluster	Euclidean Distance (Complete Linkage)	Cluster at the end of iteration
[Madison (-89, 43), Boston (-71, 42)]	$\sqrt{\lvert -89-(-71)\rvert^2 + \lvert 43-42\rvert^2} =$ $= \sqrt{325} = 18.03$	(Madison, Boston), Seattle, Vancouver, Winnipeg, Montreal

Iteration 2:

Closest Pair Of Cluster	Euclidean Distance (Complete Linkage)	Cluster at the end of iteration
[Winnipeg (-97, 50), Montreal (-74, 46)]	$\sqrt{\lvert -97-(-74)\rvert^2 + \lvert 50-46\rvert^2}$ $\sqrt{529+16} = \sqrt{545} = 23.345$	(Madison, Boston), (Winnipeg, Montreal), Seattle, Vancouver

Iteration 3:

Closest Pair Of Cluster	Euclidean Distance (Complete Linkage)	Cluster at the end of iteration
[Vancouver (-123, 49), [Winnipeg (-97, 50), Montreal (-74, 46)]]	$\sqrt{\lvert -123-(-74)\rvert^2 + \lvert 49-46\rvert^2}$ $\sqrt{2401+9} = \sqrt{2410} = 49.092$	(Madison, Boston), (Vancouver, Winnipeg, Montreal), Seattle

Iteration 4:

Closest Pair Of Cluster	Euclidean Distance (Complete Linkage)	Cluster at the end of iteration
[[Madison (-89, 43), Boston (-71, 42)] , Seattle (-122, 48)]	$\sqrt{\left\| -122 - (-71) \right\|^{2} + \left\| 48 - 42 \right\|^{2}}$ $\sqrt{2601 + 36} = \sqrt{2637} =$ 51.352	(Madison, Boston, Seattle), (Vancouver, Winnipeg, Montreal)
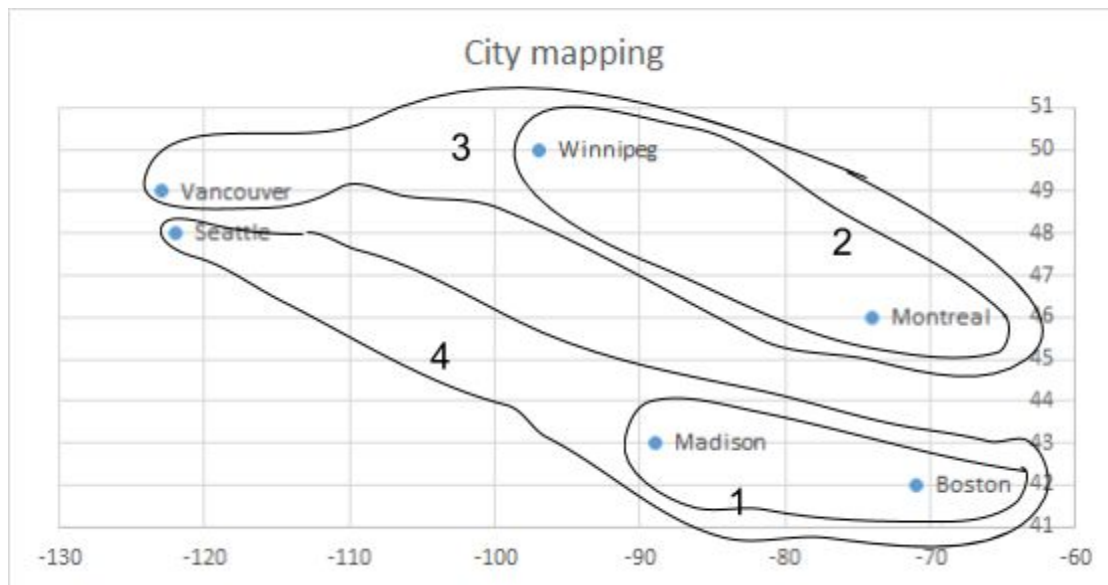
After fourth iteration it forms two different clusters.
1. One cluster is formed which is shown in Iteration-3 Output.
   Output: [Vancouver (-123, 49), [Winnipeg (-97, 50), Montreal (-74, 46)]]
2. Another cluster is shown in Iteration-4 Output.
   Output: [[Madison (-89, 43), Boston (-71, 42)] , Seattle (-122, 48)]

Question 4: K-means Clustering [25 points]

Given the following six items in 1D: $x1 = 10$; $x2 = 8$; $x3 = 6$; $x4 = 4$; $x5 = 3$; $x6 = 2$, perform k-means clustering to obtain $k = 2$ clusters by hand. Specifically,

1. Start from initial cluster centers $c1 = 0$; $c2 = 9$. Show your steps for all iterations: (1) the cluster assignments $y1; : : : ; y6$; (2) the updated cluster centers at the end of that iteration; (3) the energy at the end of that iteration.
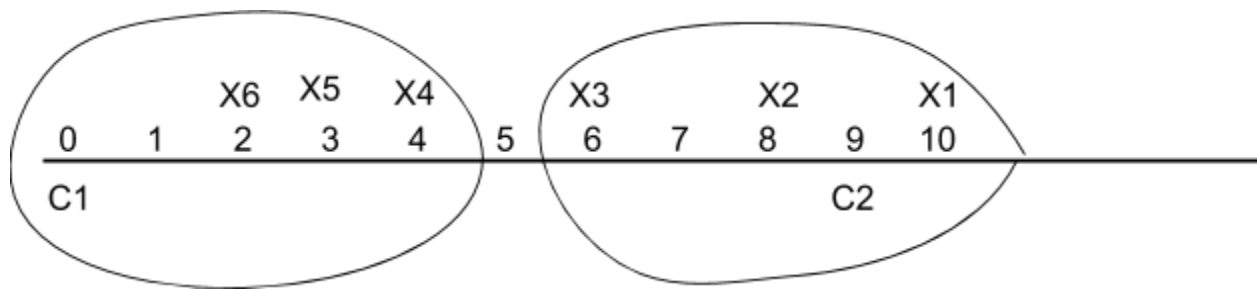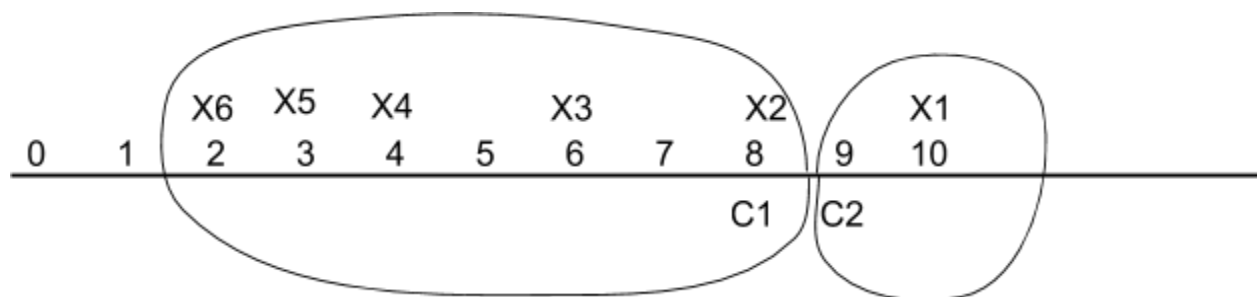
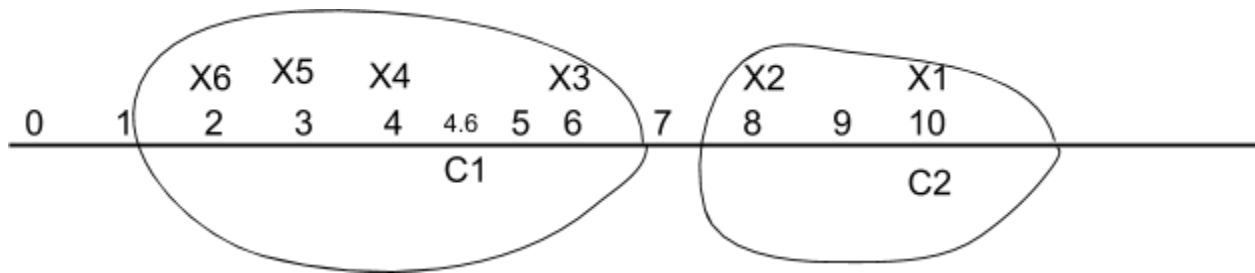$C1 = (2 + 3 + 4) \div (3) = 3$

$C2 = (6 + 8 + 10) \div (3) = 8$
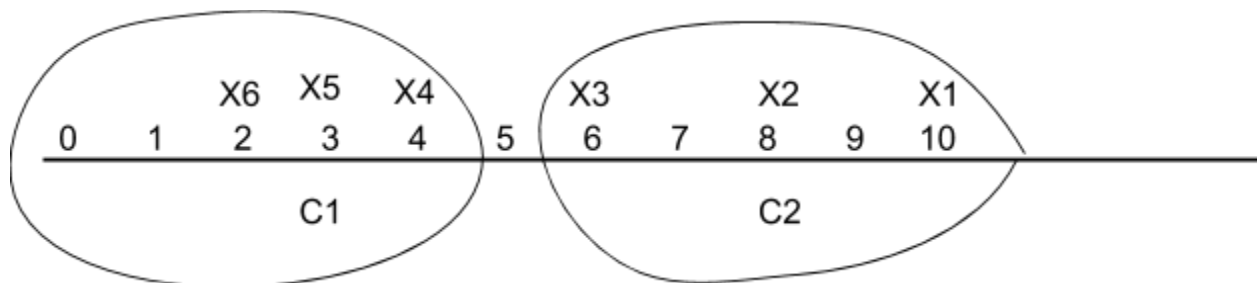
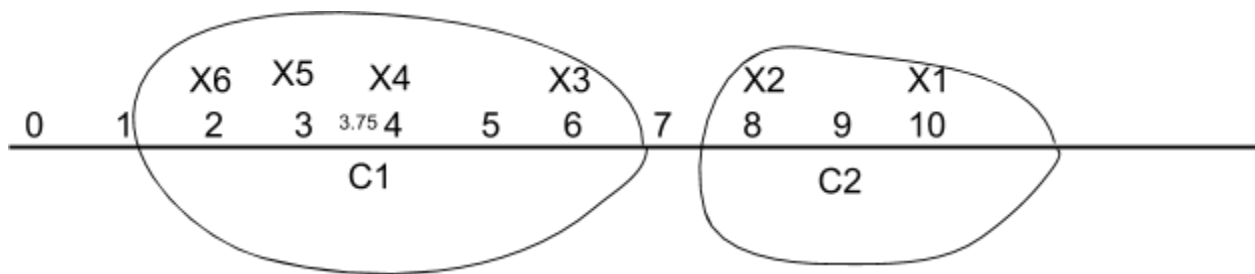2. Repeat the above but start from initial cluster centers $c1 = 8$; $c2 = 9$.

$C1 = (2 + 3 + 4 + 6 + 8) \div (5) = 4.6$

$C2 = (10) \div (1) = 10$

```
        X6  X5   X4            X3        X2            X1
  0   1   2   3    4   4.6  5  6   7    8    9    10
                       C1                        C2
```

$C1 = (2 + 3 + 4 + 6) \div (4) = 3.75$
$C2 = (8 + 10) \div (2) = 9$

```
        X6  X5   X4            X3        X2            X1
  0   1   2   3  3.75 4    5    6   7    8    9    10
                       C1                        C2
```

3. Which k-means solution is better? Why?

     First is better. All the points are closer to centroid (this shows less distortion or energy value) when compared to the cluster solution resulted in second one. Less distortion of points in the resulted clustering makes first better than second.

Reason for respective solution: Choosing cluster centroids apart will make clustering quickly compared to the having to choose the centroids close to each other. The reason is that with centroid at some extreme will includes points whose variance is quite more having to recenter the centroid multiple times.