# Q1. PCA

1.

Given the points (i,i), (i, i+1) for i = 1, …., 10

Let us represent the given space using a 20 x 2 matrix

|  |  |
| --- | --- |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 6 | 6 |
| 7 | 7 |
| 8 | 8 |
| 9 | 9 |
| 10 | 10 |
| 1 | 2 |
| 2 | 3 |
| 3 | 4 |
| 4 | 5 |
| 5 | 6 |
| 6 | 7 |
| 7 | 8 |
| 8 | 9 |
| 9 | 10 |
| 10 | 11 |

Mean of x-coordinates

*xmean* = (1 + 2 + … + 10 + 1 + 2 + …. + 10) / 20

= 5.5

Mean of y-coordinates

*ymean*

= (1 + 2 + … + 10 + 2 + 3 + …. + 11) / 20

= 6

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

*xjcentered* =

*yjcentered* =

*xj* −

*yj* −

*xmean xmean*

where, (xj, yj) are the j-th points in the given space.

Using above, the *Xcentered*

20 x 2 matrix can be represented as follow:



2. S = (1/*n* − 1) *XT X*

For our centered data, *Xcentered*

as computed in (1) can be used.

n = 20

(not displaying the matrix multiplication because of large dimensionality) Upon computation,

cov =

1. Using eigen computation, eigenvalues and eigenvectors of cov matrix are

ƛ1 = 17.500996753004870

v1 = (0.701730084756145, 0.712442901675730)

ƛ2 = 0.130582194363553

v2 = (-0.712442901675730, 0.701730084756145)

4.

Now, projection of a point X on a vector v is given by, *XcenteredvT*

Thus, projection of any given points on these two vectors would be (*Xcenteredv T* , *Xcenteredv*2*T* )

1

Using this for all the centered points, we can compute the projected values to be the following 20 x 2 matrix. (next page)

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

# Q2. Resolution

Knowledge Base: p => (q => r) …. (0)

Representing knowledge base in CNF, we get

p => (¬q v r) ……… using implication elimination

¬p v (¬q v r) ……… using implication elimination (¬p v ¬q v r) ……… using associativity of v

To Prove: (p ∧ q) => (q => r) …… (1) So, we can prove that (0) entails (1), i.e.,

Whenever (0) is true (which is given), then (1) is also true.

Representing (1) in CNF, we get

(p ∧ q) => (¬q v r) ……… using implication elimination (¬(p ∧ q)) v (¬q v r) ……… using implication elimination (¬p v ¬q) v (¬q v r) ……… using de Morgan

(¬p v ¬q v ¬q v r) ……… using associativity of v (¬p v ¬q v r)

Now, negation of (1) would be,

¬(¬p v ¬q v r)

(p ∧ q ∧ ¬r) ……… using de Morgan

Adding negation of (1) to Knowledge Base, we get a1: ¬p v ¬q v r

b1: p b2: q b3: ¬r

Goal: empty set

Step 1: resolve a1, b1: ¬q v r Step 2: resolve above, b2: r

Step 3: resolve above, b3: empty

Thus, (0) entails (1) Hence proved.

# Q3. Hierarchical Clustering

* 1. Calculating the euclidean distance between all the points,

*d* = ((*x*1 − *x*2)2 +

(*y*1

− *y*2)2)1/2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Madison (-89, 43) | Seattle  (-122, 48) | Boston  (-71, 42) | Vancouver (-123, 49) | Winnipeg (-97, 50) | Montreal (-74, 46) |
| Madison (-89, 43) | 0 | 33.376639 | 18.027756 | 34.525353 | 10.630146 | 15.297059 |
| Seattle  (-122, 48) | 33.376639 | 0 | 51.351728 | 1.414214 | 25.079872 | 48.041649 |
| Boston  (-71, 42) | 18.027756 | 51.351728 | 0 | 52.469038 | 27.202941 | 5 |
| Vancouver (-123, 49) | 34.525353 | 1.414214 | 52.469038 | 0 | 26.019224 | 49.091751 |
| Winnipeg (-97, 50) | 10.630146 | 25.079872 | 27.202941 | 26.019224 | 0 | 23.345235 |
| Montreal (-74, 46) | 15.297059 | 48.041649 | 5 | 49.091751 | 23.345235 | 0 |

# Iteration 1:

Closest pair of clusters => (Vancouver, Seattle)

Distance between them as defined by complete linkage => 1.414214 All clusters at the end of that iteration =>

(Vancouver, Seattle), (Madison), (Boston), (Winnipeg), (Montreal)

# Iteration 2:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Madison (-89, 43) | Boston  (-71, 42) | Winnipeg (-97, 50) | Montreal (-74, 46) | (Vancouver, Seattle) |
| Madison (-89, 43) | 0 | 18.027756 | 10.630146 | 15.297059 | 34.525353 |
| Boston  (-71, 42) | 18.027756 | 0 | 27.202941 | 5 | 52.469038 |
| Winnipeg (-97, 50) | 10.630146 | 27.202941 | 0 | 23.345235 | 26.019224 |
| Montreal (-74, 46) | 15.297059 | 5 | 23.345235 | 0 | 49.091751 |

Closest pair of clusters => (Montreal, Boston)

Distance between them as defined by complete linkage => 5

All clusters at the end of that iteration => (Vancouver, Seattle), (Montreal, Boston), (Madison), (Winnipeg)

# Iteration 3:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Madison (-89, 43) | Winnipeg (-97, 50) | (Vancouver, Seattle) | (Montreal, Boston) |
| Madison (-89, 43) | 0 | 10.630146 | 34.525353 | 18.027756 |
| Winnipeg (-97, 50) | 10.630146 | 0 | 26.019224 | 27.202941 |
| (Montreal, Boston) | 18.027756 | 27.202941 | 52.469038 | 0 |

Closest pair of clusters => (Madison, Winnipeg)

Distance between them as defined by complete linkage => 10.630146

All clusters at the end of that iteration => (Vancouver, Seattle), (Montreal, Boston), (Madison,Winnipeg)

# Iteration 4:

|  |  |  |  |
| --- | --- | --- | --- |
|  | (Vancouver, Seattle) | (Montreal, Boston) | (Madison, Winnipeg) |
| (Montreal, Boston) | 52.469038 | 0 | 27.202941 |
| (Vancouver, Seattle) | 0 | 52.469038 | 34.525353 |
| (Madison, Winnipeg) | 34.525353 | 27.202941 | 0 |

Closest pair of clusters => ((Madison, Winnipeg), (Montreal, Boston)) Distance between them as defined by complete linkage => 27.202941

All clusters at the end of that iteration => (Vancouver, Seattle), (Madison, Winnipeg, Montreal, Boston)

* 1. Calculating the euclidean distance between all the points, constraint being US to Canada city distance = ND (infinite),=

*d* = ((*x*1 − *x*2)2 +

(*y*1

− *y*2)2)1/2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Madison (-89, 43) | Seattle  (-122, 48) | Boston  (-71, 42) | Vancouver (-123, 49) | Winnipeg (-97, 50) | Montreal (-74, 46) |
| Madison (-89, 43) | 0 | 33.376639 | 18.027756 | ND | ND | ND |
| Seattle  (-122, 48) | 33.376639 | 0 | 51.351728 | ND | ND | ND |
| Boston  (-71, 42) | 18.027756 | 51.351728 | 0 | ND | ND | ND |
| Vancouver (-123, 49) | ND | ND | ND | 0 | 26.019224 | 49.091751 |
| Winnipeg (-97, 50) | ND | ND | ND | 26.019224 | 0 | 23.345235 |
| Montreal (-74, 46) | ND | ND | ND | 49.091751 | 23.345235 | 0 |

# Iteration 1:

Closest pair of clusters => (Madison, Boston)

Distance between them as defined by complete linkage => 18.027756

All clusters at the end of that iteration => (Madison, Boston), Seattle, Vancouver, Winnipeg, Montreal

# Iteration 2:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Seattle  (-122, 48) | Vancouver (-123, 49) | Winnipeg (-97, 50) | Montreal (-74, 46) | (Madison, Boston) |
| Seattle  (-122, 48) | 0 | ND | ND | ND | 51.351728 |
| Vancouver (-123, 49) | ND | 0 | 26.019224 | 49.091751 | ND |
| Winnipeg (-97, 50) | ND | 26.019224 | 0 | 23.345235 | ND |
| Montreal (-74, 46) | ND | 49.091751 | 23.345235 | 0 | ND |

Closest pair of clusters => (Winnipeg, Montreal)

Distance between them as defined by complete linkage => 23.345235

All clusters at the end of that iteration => (Madison, Boston), (Winnipeg, Montreal), Seattle, Vancouver

# Iteration 3:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Seattle  (-122, 48) | Vancouver (-123, 49) | (Madison, Boston) | (Winnipeg, Montreal) |
| Seattle  (-122, 48) | 0 | ND | 51.351728 | ND |
| Vancouver (-123, 49) | ND | 0 | ND | 49.091751 |
| (Madison, Boston) | 51.351728 | ND | 0 | ND |

Closest pair of clusters => (Vancouver, Winnipeg, Montreal)

Distance between them as defined by complete linkage => 49.091751

All clusters at the end of that iteration => (Madison, Boston), (Vancouver, Winnipeg, Montreal), Seattle

# Iteration 4:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Seattle  (-122, 48) | (Madison, Boston) | (Vancouver, Winnipeg, Montreal) |
| Seattle  (-122, 48) | 0 | 51.351728 | ND |
| (Madison, Boston) | 51.351728 | 0 | ND |
| (Vancouver, Winnipeg, Montreal) | ND | ND | 0 |

Closest pair of clusters => (Madison, Boston, Seattle)

Distance between them as defined by complete linkage => 51.351728

All clusters at the end of that iteration => (Madison, Boston, Seattle), (Vancouver, Winnipeg, Montreal)

# Q4. K-means Clustering

1.

# Iteration 1:

|  |  |  |  |
| --- | --- | --- | --- |
|  | dist\_c1 (0) | dist\_c2 (9) | y |
| x1 (10) | 10 | 1 | y1 = 2 |
| x2 (8) | 8 | 1 | y2 = 2 |
| x3 (6) | 6 | 3 | y3 = 2 |
| x4 (4) | 4 | 5 | y4 = 1 |
| x5 (3) | 3 | 6 | y5 = 1 |
| x6 (2) | 2 | 7 | y6 = 1 |

Updated centers

c1 = (x4 + x5 + x6) / 3

= (4 + 3 + 2) / 3

= (9) / 3

= 3

c2 = (x1 + x2 + x3) / 3

= (10 + 8 + 6) / 3

= (24) / 3

= 8

Energy = *d*(*x*1, *c*2)2 + *d*(*x*2, *c*2)2 +

*d*(*x*3, *c*2)2 +

*d*(*x*4, *c*1)2 +

*d*(*x*5, *c*1)2 +

*d*(*x*6, *c*1)2

= 12 + 12 +

32 +

42 +

32 + 22

= 1 + 1 + 9 + 16 + 9 + 4

= 40

# Iteration 2:

|  |  |  |  |
| --- | --- | --- | --- |
|  | dist\_c1 (3) | dist\_c2 (8) | y |
| x1 (10) | 7 | 2 | y1 = 2 |
| x2 (8) | 5 | 0 | y2 = 2 |
| x3 (6) | 3 | 2 | y3 = 2 |
| x4 (4) | 1 | 4 | y4 = 1 |
| x5 (3) | 0 | 5 | y5 = 1 |
| x6 (2) | 1 | 6 | y6 = 1 |

As clustering remains the same as Iteration 1, implying the centroids have converged, hence we stop here.

Energy = *d*(*x*1, *c*2)2 + *d*(*x*2, *c*2)2 +

*d*(*x*3, *c*2)2 +

*d*(*x*4, *c*1)2 +

*d*(*x*5, *c*1)2 +

*d*(*x*6, *c*1)2

= 22 + 02 +

22 +

12 +

02 + 12

= 4 + 0 + 4 + 1 + 0 + 1

= 10

2.

# Iteration 1:

|  |  |  |  |
| --- | --- | --- | --- |
|  | dist\_c1 (8) | dist\_c2 (9) | y |
| x1 (10) | 2 | 1 | y1 = 2 |
| x2 (8) | 0 | 1 | y2 = 1 |
| x3 (6) | 2 | 3 | y3 = 1 |
| x4 (4) | 4 | 5 | y4 = 1 |
| x5 (3) | 5 | 6 | y5 = 1 |
| x6 (2) | 6 | 7 | y6 = 1 |

Updated centers

c1 = (x2 + x3 + x4 + x5 + x6) / 5

= (8 + 6 + 4 + 3 + 2) / 5

= (23) / 5

= 4.6

c2 = (x1) / 1

= (10) / 1

= 10

Energy = *d*(*x*1, *c*2)2 + *d*(*x*2, *c*2)2 +

*d*(*x*3, *c*2)2 +

*d*(*x*4, *c*1)2 +

*d*(*x*5, *c*1)2 +

*d*(*x*6, *c*1)2

= 12 + 02 +

22 +

42 +

52 + 62

= 1 + 0 + 4 + 16 + 25 + 36

= 82

# Iteration 2:

|  |  |  |  |
| --- | --- | --- | --- |
|  | dist\_c1 (4.6) | dist\_c2 (10) | y |
| x1 (10) | 5.4 | 0 | y1 = 2 |
| x2 (8) | 3.4 | 2 | y2 = 2 |
| x3 (6) | 1.4 | 4 | y3 = 1 |
| x4 (4) | 0.6 | 6 | y4 = 1 |
| x5 (3) | 1.6 | 7 | y5 = 1 |
| x6 (2) | 2.6 | 8 | y6 = 1 |

Updated centers

c1 = (x3 + x4 + x5 + x6) / 4

= (6 + 4 + 3 + 2) / 4

= (15) / 4

= 3.75

c2 = (x1 + x2) / 2

= (10 + 8) / 2

= (18) / 2

= 9

Energy = *d*(*x*1, *c*2)2 + *d*(*x*2, *c*2)2 +

*d*(*x*3, *c*2)2 +

*d*(*x*4, *c*1)2 +

*d*(*x*5, *c*1)2 +

*d*(*x*6, *c*1)2

= 02 + 22 +

1.42 +

0.62 +

1.62 +

2.62

= 0 + 4 + 1.96 + 0.36 + 2.56 + 6.76

= 15.64

# Iteration 3:

|  |  |  |  |
| --- | --- | --- | --- |
|  | dist\_c1 (3.75) | dist\_c2 (9) | y |
| x1 (10) | 6.25 | 1 | y1 = 2 |
| x2 (8) | 4.25 | 1 | y2 = 2 |
| x3 (6) | 2.25 | 3 | y3 = 1 |
| x4 (4) | 0.25 | 5 | y4 = 1 |
| x5 (3) | 0.75 | 6 | y5 = 1 |
| x6 (2) | 1.75 | 7 | y6 = 1 |

As clustering remains the same as Iteration 2, implying the centroids have converged, hence we stop here.

Energy = *d*(*x*1, *c*2)2 + *d*(*x*2, *c*2)2 +

*d*(*x*3, *c*2)2 +

*d*(*x*4, *c*1)2 +

*d*(*x*5, *c*1)2 +

*d*(*x*6, *c*1)2

= 12 + 12 +

2.252 +

0.252 +

0.752 +

1.752

= 1 + 1 + 5.0625 + 0.0625 + 0.5625 + 3.0625

= 10.75

3.

As the final energy in part 1 is less, hence starting with centers c1 (0) and c2 (9) is providing a better minima of distortion. Thus the k-means solution of part 1 is better.