Sindhuja Yerramalla
U00839259

Data Mining
Assignment 3

Q1: Given four 2-dimensional data points:

(2,2) (4,4) (1,5) and (5,1)

Given $L = 2$ and $K = 1$

Step1: Mean of datapoints & difference between mean & datapoints

Mean $\mu = \left[\frac{2+4+1+5}{4}, \frac{2+4+5+1}{4}\right]$

$= (3, 3)$

Difference between datapoints & mean.

$d_1 = (2-3, 2-3) = (-1, -1)$

$d_2 = (4-3, 4-3) = (1, 1)$

$d_3 = (1-3, 5-3) = (-2, 2)$

$d_4 = (5-3, 1-3) = (2, -2)$

Step2: Covariance Matrix.

$$X = \begin{bmatrix} -1 & 1 & -2 & 2 \\ -1 & 1 & 2 & -2 \end{bmatrix}$$ → from difference between datapoints & mean

Covariance Matrix $(\Sigma) = \frac{1}{4} \times X^T$

$$\Sigma = \frac{1}{4}\begin{bmatrix} -1 & 1 & -2 & 2 \\ -1 & 1 & 2 & -2 \end{bmatrix}\begin{bmatrix} -1 & -1 \\ 1 & 1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix}$$

$$= \frac{1}{4}\begin{bmatrix} 1+1+4+4 & 1+1-4-4 \\ 1+1-4-4 & 1+1+4+4 \end{bmatrix}$$

$$= \frac{1}{4}\begin{pmatrix} 10 & -6 \\ -6 & 10 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 5/2 & -3/2 \\ -3/2 & 5/2 \end{pmatrix}$$

Step 3: Obtaining Eigen values & Eigen vectors of Σ

Eigen values $= |\Sigma - \lambda I| = 0$

$$\Rightarrow \begin{vmatrix} 5/2-\lambda & -3/2 \\ -3/2 & 5/2-\lambda \end{vmatrix} = 0$$

$$(5/2-\lambda)^2 - (-3/2)^2 = 0$$

$$\frac{25}{4} + \lambda^2 - 5\lambda - \frac{9}{4} = 0$$

$$\frac{16}{4} + \lambda^2 - 5\lambda = 0$$

$$4\lambda^2 - 20\lambda + 16 = 0$$

$$\lambda^2 - 5\lambda + 4 = 0$$

$\Rightarrow \lambda^2 - 5\lambda + 4 = 0$

~~$\lambda(\lambda-4)$~~

$\Rightarrow \lambda^2 - \lambda - 4\lambda + 4 = 0$

$= \lambda(\lambda-1) - 4(\lambda-1) = 0$

$= (\lambda-4)\cdot(\lambda-1) = 0$

$\boxed{\lambda = 4}$ (or) $\boxed{\lambda = 1}$

Eigen vectors $\Rightarrow [\Sigma - \lambda I][x] = 0$

case(i) $\rightarrow \lambda = 4$

$$\begin{bmatrix} 5/2 - 4 & -3/2 \\ -3/2 & 5/2 - 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} -3/2 & -3/2 \\ -3/2 & -3/2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$x_1 + x_2 = 0$

$x_1 = -x_2.$

case(ii) $\rightarrow \lambda = 1$

$$\begin{bmatrix} 5/2 - 1 & -3/2 \\ -3/2 & 5/2 - 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} 3/2 & -3/2 \\ -3/2 & 3/2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$x_1 - x_2 = 0$

$x_1 = x_2.$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a \\ -a \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a \\ a \end{bmatrix}$$

$$\boxed{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \sqrt{1/2} \\ -\sqrt{1/2} \end{bmatrix}} \qquad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \sqrt{1/2} \\ \sqrt{1/2} \end{bmatrix}$$

Step 4 : Arrange eigen vectors in descending order of eigen values

$$\phi = \begin{pmatrix} \sqrt{1/2} & \sqrt{1/2} \\ -\sqrt{1/2} & \sqrt{1/2} \end{pmatrix}$$

Step 5 : Transformation from L to K dimension

let $Y = \phi X$.

for d_1 (2, 2)

$$Y = \begin{pmatrix} \sqrt{1/2} & \sqrt{1/2} \\ -\sqrt{1/2} & \sqrt{1/2} \end{pmatrix} \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

$$= \begin{pmatrix} 4\sqrt{1/2} \\ 0 \end{pmatrix} = \begin{pmatrix} 2\sqrt{2} \\ 0 \end{pmatrix}$$

$$\boxed{Y = \begin{pmatrix} 2.83 \\ 0 \end{pmatrix}}$$

for d_2 (4, 4)

$$Y = \begin{pmatrix} \sqrt{1/2} & \sqrt{1/2} \\ -\sqrt{1/2} & \sqrt{1/2} \end{pmatrix} \begin{pmatrix} 4 \\ 4 \end{pmatrix}$$

$$= \begin{bmatrix} 4\sqrt{1/2} + 4\sqrt{1/2} \\ 4\sqrt{1/2} - 4\sqrt{1/2} \end{bmatrix}$$

$$= \begin{bmatrix} 4\sqrt{2} \\ 0 \end{bmatrix}$$

$$\boxed{Y = \begin{matrix} 5.66 \\ 0 \end{matrix}}$$

For d_3: (1, 5)

$$Y = \begin{bmatrix} \sqrt{1/2} & \sqrt{1/2} \\ -\sqrt{1/2} & \sqrt{1/2} \end{bmatrix} \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

$$= \begin{bmatrix} 6\sqrt{1/2} \\ 4\sqrt{1/2} \end{bmatrix} \begin{bmatrix} 3\sqrt{2} \\ 2\sqrt{2} \end{bmatrix}$$

$$Y = \begin{bmatrix} 4.24 \\ 2.83 \end{bmatrix}$$

for d_4: (5, 1)

$$Y = \begin{bmatrix} \sqrt{1/2} & \sqrt{1/2} \\ -\sqrt{1/2} & \sqrt{1/2} \end{bmatrix} \begin{bmatrix} 5 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 6\sqrt{1/2} \\ -4\sqrt{1/2} \end{bmatrix} = \begin{bmatrix} 4.24 \\ -2.83 \end{bmatrix}$$

$$Y = \begin{bmatrix} 4.24 \\ -2.83 \end{bmatrix}$$

Step-6 : dimensions, whose eigen vectors is low or minimum

for $d_1(2,2) = (2.83, 0) \Rightarrow 0$

for $d_2(4,4) = (5.66 \quad 0) \Rightarrow 0$

for $d_3(1,5) = (4.24 \quad 2.83) \Rightarrow 2.83$

for $d_4(5,1) = (4.24 \quad -2.83) \Rightarrow -2.83$

∴ 2nd eigen vectors has minimum value, we select 2nd dimension points to projected subspace $\Rightarrow [0 \quad 0; 2.83 \quad -2.83]$

Q2 Given 2-dimensional points

a(5,5) b(5,7) c(7,8) d(8,4) e(3,6) f(4,8)

• Manhattan distance = $|x_1 - x_2| + |y_2 - y_1|$

(i) Distance Based model :

Given $\varepsilon = 4$ & size threshold = 3

Manhattan distances from point a(5,5)

a-b = |5-5| + |5-7| = 2

a-c = |5-7| + |5-8| = 5 → > 4

a-d = |5-8| + |5-4| = 4

a-e = |5-3| + |5-6| = 3

a-f = |5-4| + |5-8| = 4

ε-neighbourhood of point a N(a) = {a, b, d, e, f}

N(a) > size threshold (3), point a is not an outlier

Manhattan distances from point b(5,7)

b-a = |5-5| + |7-5| = 2

b-c = |5-7| + |7-8| = 3

b-d = |5-8| + |7-4| = 6 → > 4

b-e = |5-3| + |7-6| = 3

b-f = |5-4| + |7-8| = 2

Σ-neighbourhood of b $N(b) = \{a, b, c, e, f\}$

Since $N(5) >$ size threshold(3), b is not an outlier.

Manhattan distance from point c

$c-a = |7-5| + |8-5| = 5 > 4$

$c-b = |7-5| + |8-7| = 3$

$c-d = |7-8| + |8-4| = 5 > 4$

$c-e = |7-3| + |8-6| = 6 > 4$

$c-f = |7-4| + |8-8| = 3$

Σ-neighbourhood of c $N(c) = \{b, c, f\}$

$\because N(c) =$ size threshold(3), c is an outlier

Manhattan distance from point d

$d-a = |8-5| + |4-5| = 4$

$d-b = |8-5| + |4-7| = 6 > 4$

$d-c = |8-7| + |4-8| = 5 > 4$

$d-e = |8-3| + |4-6| = 7 > 4$

$d-f = |8-4| + |4-8| = 8 > 4$

Σ-neighbourhood of d $N(d) = \{a, d\}$

$\because N(d) <$ size threshold, d is an outlier

Manhattan distances from point e

$e-a = |3-5| + |6-5| = 3$

$e-b = |3-5| + |6-7| = 3$

$e-c = |3-7| + |6-8| = 6$

$e-d = |3-8| + |6-4| = 7$

$e-f = |3-4| + |6-8| = 3$

Σ - neighbourhood of point e $N(e) = \{a, b, e, f\}$

Since Σ - neighbourhood $N(e) >$ size threshold,

e is not an outlier

Manhattan distances from point f

$f-a = |4-5| + |8-5| = 4$

$f-b = |4-5| + |8-7| = 2$

$f-c = |4-7| + |8-8| = 3$

$f-d = |4-8| + |8-4| = 8$

$f-e = |4-3| + |8-6| = 3$

Σ - neighbourhood of point f $N(f) = \{a, b, c, f, e\}$

$\because N(f) >$ size threshold (3), f is not an outlier.

$\therefore$ The outliers are only $\{c, d\}$ //.

ii) Density Based Model:

Local outlier factors of d:

Manhattan distances from point d:

$d-a = |8-5| + |4-5| = 4$

$d-b = |8-5| + |4-7| = 6$

$d-c = |8-7| + |4-8| = 5$

$d-e = |8-3| + |4-6| = 7$

$d-f = |8-4| + |4-8| = 8$

Given $K=3$

$\therefore$ 3-neighbourhood of d $N_3(d) = \{a, c, b\}$

Distance between d & k^{th} nearest neighbour = 6

$\therefore$ local reachability distance of d $lrd_3(d) = 1/6$

$\therefore$ local outlier factor of d $= \dfrac{\sum_{o \in N_3(d)} \frac{lrd_3(o)}{lrd_3(d)}}{k}$

$$= \frac{\frac{lrd_3(a) + lrd_3(c) + lrd_3(b)}{lrd_3(d)}}{k} \quad \text{---(1)}$$

~~[we know from distance based model, $N(a) = \{a, b, d, e, f\}$~~

~~$\therefore lrd_3(a) = 1/4$.~~

~~$N(b) = \{a, b, c, e, f\}$ $\therefore lrd_3(b) = 1/4$~~

local reachability distances

Manhattten distances for points

$a-b = |5-5| + |5-7| = 2$

$a-c = |5-7| + |5-8| = 5$

$a-d = |5-8| + |5-4| = 4$

$a-e = |5-3| + |5-6| = 3$

$a-f = |5-4| + |5-8| = 4$

$\therefore$ Σ-neighbour hood of a: $N(a) = \{b, e, d, f\}$

Distance between point a & k^{th} nearest neighbour $\Sigma = 4$

$\therefore$ $lrd_3(a) = 1/4$

Manhattan distance for point b

$b-a = |5-5| + |7-5| = 2.$

$b-c = |5-7| + |7-8| = 3$

$b-d = |5-8| + |7-4| = 6$

$b-e = |5-3| + |7-6| = 3$

$b-f = |5-4| + |7-8| = 2$

Σ-neighborhood of b $N_3(b) = \{a, f, c, e\}$

Distance between point b & k^{th} nearest distance $\Sigma = 3$

$\therefore$ $lrd_3(b) = 1/3$

Manhattan distances for point c

$c-a = |7-5| + |8-5| = 5$

$c-b = |7-5| + |8-7| = 3$

$c-d = |7-8| + |8-4| = 5$

$c-e = |7-3| + |8-6| = 6$

$c-f = |7-4| + |8-8| = 3$

$\therefore$ Σ-neighbourhood of c; $N_3(c) = \{b, f, a, d\}$

$\therefore$ Distance between point b & kth nearest neighbour $\Sigma = 5$

$\therefore$ $\boxed{lrd_3(c) = 1/5}$

$\therefore$ (1) $\rightarrow$ local outlier factor of d $= \dfrac{\frac{1}{4} + \frac{1}{5} + \frac{1}{3}}{1/6.}$

$\dfrac{}{3}$

$\boxed{lof_3(d) = 1.57}$

$\rightarrow$ Local Outlier Factor of C

Manhattan distances from point c.

$c-a = |7-5| + |8-5| = 5$

$c-b = |7-5| + |8-7| = 3.$

$c-d = |7-8| + |8-4| = 5$

$c-e = |7-3| + |8-6| = 6$

$c-f = |7-4| + |8-8| = 3$

$\therefore$ Σ-neighbourhood of c $N_3(c) = \{5, f, a, d\}$

→ Distance between point b & k^{th} nearest neighbour $\Sigma = 5$

$\therefore$ local reachability distance of point c $\boxed{lrd_3(c) = 1/5}$

$\therefore$ Local outlier factor of c = $\dfrac{\dfrac{lrd_3(b) + lrd_3(f) + lrd_3(a) + lrd_3(d)}{lrd_3(c)}}{k.}$ — (1)

Local reachability distances

Manhattan distances for point b

$b-a = |5-5| + |7-5| = 2$

$b-c = |5-7| + |7-8| = 3$

$b-d = |5-8| + |7-4| = 6$

$b-e = |5-3| + |7-6| = 3$

$b-f = |5-4| + |7-8| = 2$

$\therefore$ Σ-neighbourhood of b $N_3(5) = \{a, f, c\}$

Distance between point c & k^{th} nearest neighbour $\Sigma = 3$

$\therefore$ $\boxed{lrd_3(5) = 1/3}$

Manhattan distance for point f

$f-a = |4-5| + |8-5| = 4$

$f-b = |4-5| + |8-7| = 2.$

$f-c = |4-7| + |8-8| = 3$

$f-d = |4-8| + |8-4| = 4$

$f-c = |4-3| + |8-6| = 3$

$\therefore$ Σ-neighbourhood of $N_3(f) = \{b, c, e\}$

Distance between f & 3rd nearest neighbour $\Sigma = 3$

$\therefore$ local reachability distance of point f $\boxed{lrd_3(f) = 1/3}$

Manhattan distance of point a

$a-b = |5-5| + |5-7| = 2$

$a-c = |5-7| + |5-8| = 5$

$a-d = |5-8| + |5-4| = 4$

$a-c = |5-3| + |5-6| = 3$

$a-f = |5-4| + |5-8| = 4$

$\therefore$ Σ-neighbourhood of $N_3(a) = \{n, e, d\}$

Distance $\Sigma = 4.$

$\therefore$ $\boxed{lrd_3(a) = 1/4}$

Manhattan distances for d

$d-a = |8-5| + |4-5| = 4$

$d-b = |8-5| + |4-7| = 6$

$d-c = |8-7| + |4-8| = 5$

$d-e = |8-3| + |4-6| = 7$

$d-f = |8-4| + |4-8| = 8$

$\therefore N_3(d) = \{a, c, b\}$

$\Sigma = 6$

$$lrd_3(d) = 1/6$$

$\therefore$ (1) $\Rightarrow$ local outlier factor of point c $= \dfrac{\dfrac{lrd_3(b) + lrd_3(f) + lrd_3(a) + lrd_3(d)}{lrd_3(c)}}{K}$

$$= \frac{\frac{1/3 + 1/3 + 1/4 + 1/6}{1/5}}{3}$$

$$= \frac{\frac{0.333 + 0.333 + 0.25 + 0.166}{0.2}}{3}$$

$$lof_3(c) = 1.803$$ //