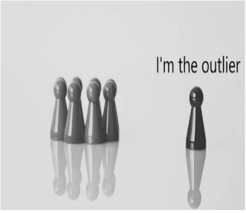
Dr.Siddique Ibrahim
VIT-AP University

Today's Topic: Outlier Analysis

1. INTRODUCTION
2. K-DISTANCE AND K-NEIGHBORS
3. REACHABILITY DISTANCE (RD)
4. LOCAL REACHABILITY DISTANCE (LRD)
5. LOCAL OUTLIER FACTOR (LOF)
6. EXAMPLE

What is Outlier?

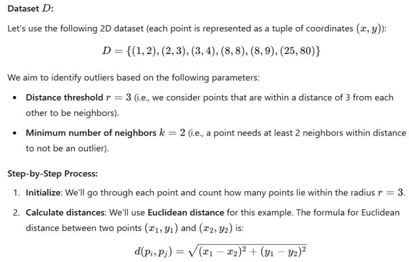
- An outlier is a data point that is different or far from the rest of the data points.
- The question that arises here is that can we identify the outliers present in the data?



Distance Based Outlier Detection

1. Index Based Algorithm
2. Nested Loop Algorithm
3. Cell Based Algorithm

Distance-based outlier detection with a nested loop algorithm



Dataset D:

Let's use the following 2D dataset (each point is represented as a tuple of coordinates (x, y)):

$$D = \{(1,2), (2,3), (3,4), (8,8), (8,9), (25,80)\}$$

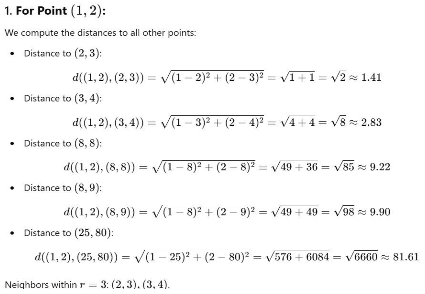
We aim to identify outliers based on the following parameters:

- **Distance threshold** $r = 3$ (i.e., we consider points that are within a distance of 3 from each other to be neighbors).
- **Minimum number of neighbors** $k = 2$ (i.e., a point needs at least 2 neighbors within distance r to not be an outlier).

Step-by-Step Process:

1. **Initialize**: We'll go through each point and count how many points lie within the radius $r = 3$.
2. **Calculate distances**: We'll use **Euclidean distance** for this example. The formula for Euclidean distance between two points (x_1, y_1) and (x_2, y_2) is:

$$d(p_i, p_j) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$



1. **For Point $(1, 2)$:**

We compute the distances to all other points:

- Distance to $(2, 3)$:

$$d((1,2),(2,3)) = \sqrt{(1-2)^2 + (2-3)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$$

- Distance to $(3, 4)$:

$$d((1,2),(3,4)) = \sqrt{(1-3)^2 + (2-4)^2} = \sqrt{4+4} = \sqrt{8} \approx 2.83$$

- Distance to $(8, 8)$:

$$d((1,2),(8,8)) = \sqrt{(1-8)^2 + (2-8)^2} = \sqrt{49+36} = \sqrt{85} \approx 9.22$$

- Distance to $(8, 9)$:

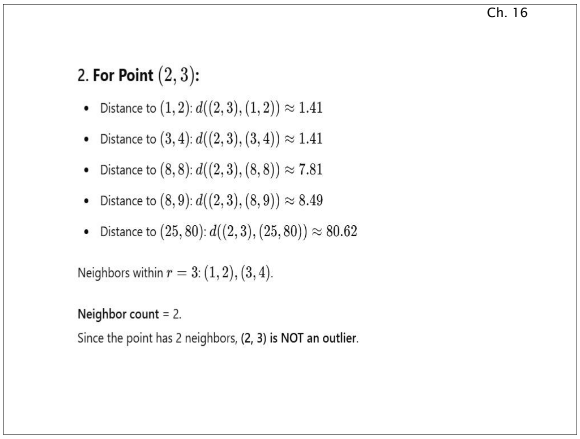
$$d((1,2),(8,9)) = \sqrt{(1-8)^2 + (2-9)^2} = \sqrt{49+49} = \sqrt{98} \approx 9.90$$

- Distance to $(25, 80)$:

$$d((1,2),(25,80)) = \sqrt{(1-25)^2 + (2-80)^2} = \sqrt{576+6084} = \sqrt{6660} \approx 81.61$$

Neighbors within $r = 3$: $(2, 3), (3, 4)$.

Neighbor count = 2.
Since the point has 2 neighbors (which meets $k = 2$), **(1, 2) is NOT an outlier.**

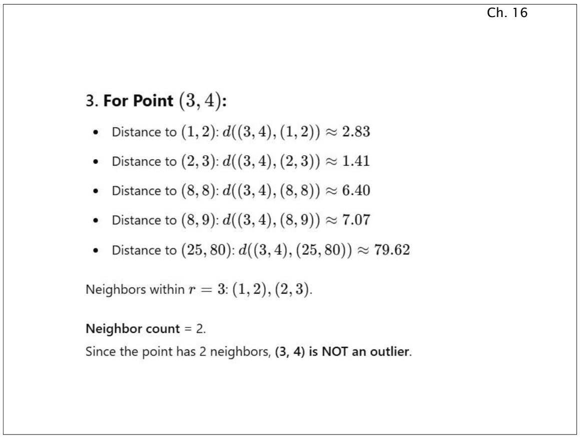


2. **For Point $(2, 3)$:**

- Distance to $(1, 2)$: $d((2, 3), (1, 2)) \approx 1.41$
- Distance to $(3, 4)$: $d((2, 3), (3, 4)) \approx 1.41$
- Distance to $(8, 8)$: $d((2, 3), (8, 8)) \approx 7.81$
- Distance to $(8, 9)$: $d((2, 3), (8, 9)) \approx 8.49$
- Distance to $(25, 80)$: $d((2, 3), (25, 80)) \approx 80.62$

Neighbors within $r = 3$: $(1, 2), (3, 4)$.

Neighbor count = 2.
Since the point has 2 neighbors, **(2, 3) is NOT an outlier**.

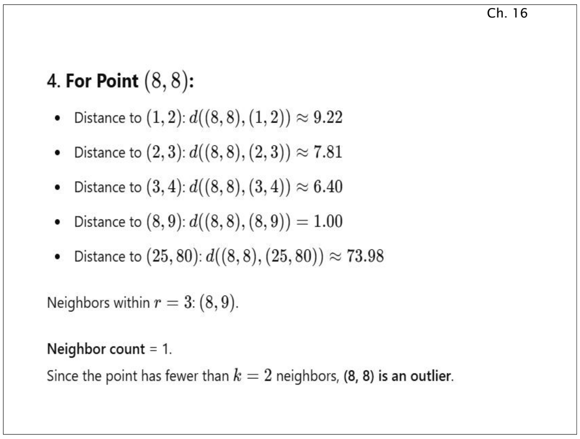


3. **For Point $(3, 4)$:**

- Distance to $(1, 2)$: $d((3, 4), (1, 2)) \approx 2.83$
- Distance to $(2, 3)$: $d((3, 4), (2, 3)) \approx 1.41$
- Distance to $(8, 8)$: $d((3, 4), (8, 8)) \approx 6.40$
- Distance to $(8, 9)$: $d((3, 4), (8, 9)) \approx 7.07$
- Distance to $(25, 80)$: $d((3, 4), (25, 80)) \approx 79.62$

Neighbors within $r = 3$: $(1, 2), (2, 3)$.

Neighbor count = 2.
Since the point has 2 neighbors, **(3, 4) is NOT an outlier**.

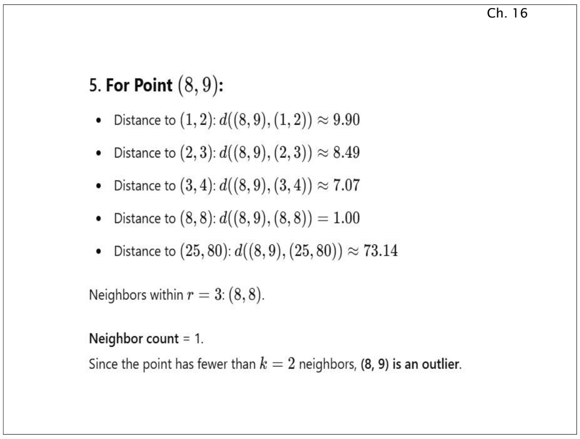


4. **For Point $(8, 8)$:**

- Distance to $(1, 2)$: $d((8, 8), (1, 2)) \approx 9.22$
- Distance to $(2, 3)$: $d((8, 8), (2, 3)) \approx 7.81$
- Distance to $(3, 4)$: $d((8, 8), (3, 4)) \approx 6.40$
- Distance to $(8, 9)$: $d((8, 8), (8, 9)) = 1.00$
- Distance to $(25, 80)$: $d((8, 8), (25, 80)) \approx 73.98$

Neighbors within $r = 3$: $(8, 9)$.

Neighbor count = 1.
Since the point has fewer than $k = 2$ neighbors, **(8, 8) is an outlier**.

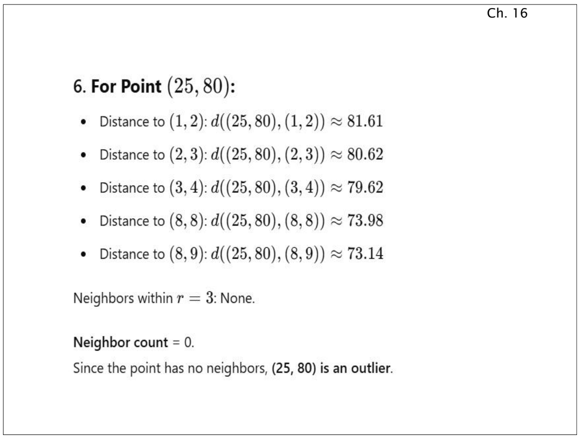


5. **For Point $(8, 9)$:**

- Distance to $(1, 2)$: $d((8, 9), (1, 2)) \approx 9.90$
- Distance to $(2, 3)$: $d((8, 9), (2, 3)) \approx 8.49$
- Distance to $(3, 4)$: $d((8, 9), (3, 4)) \approx 7.07$
- Distance to $(8, 8)$: $d((8, 9), (8, 8)) = 1.00$
- Distance to $(25, 80)$: $d((8, 9), (25, 80)) \approx 73.14$

Neighbors within $r = 3$: $(8, 8)$.

Neighbor count = 1.
Since the point has fewer than $k = 2$ neighbors, **(8, 9) is an outlier**.

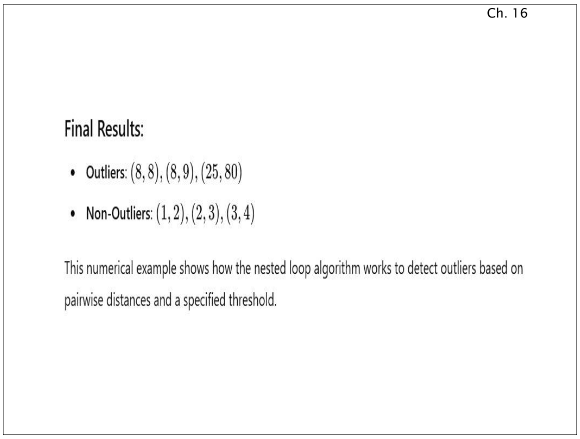


6. **For Point $(25, 80)$:**

- Distance to $(1, 2)$: $d((25, 80), (1, 2)) \approx 81.61$
- Distance to $(2, 3)$: $d((25, 80), (2, 3)) \approx 80.62$
- Distance to $(3, 4)$: $d((25, 80), (3, 4)) \approx 79.62$
- Distance to $(8, 8)$: $d((25, 80), (8, 8)) \approx 73.98$
- Distance to $(8, 9)$: $d((25, 80), (8, 9)) \approx 73.14$

Neighbors within $r = 3$: None.

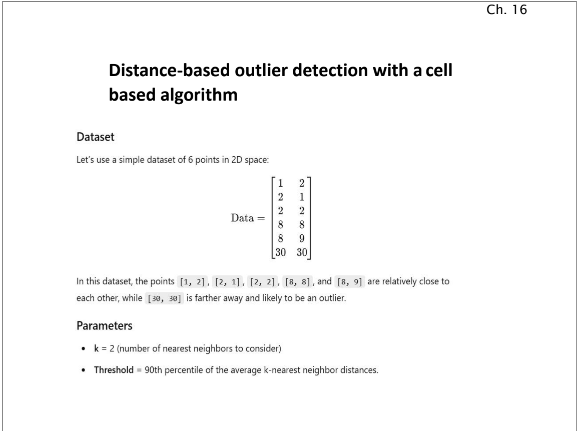
Neighbor count = 0.
Since the point has no neighbors, **(25, 80) is an outlier**.



Final Results:

- **Outliers**: $(8, 8), (8, 9), (25, 80)$
- **Non-Outliers**: $(1, 2), (2, 3), (3, 4)$

This numerical example shows how the nested loop algorithm works to detect outliers based on pairwise distances and a specified threshold.



Distance-based outlier detection with a cell based algorithm

Dataset

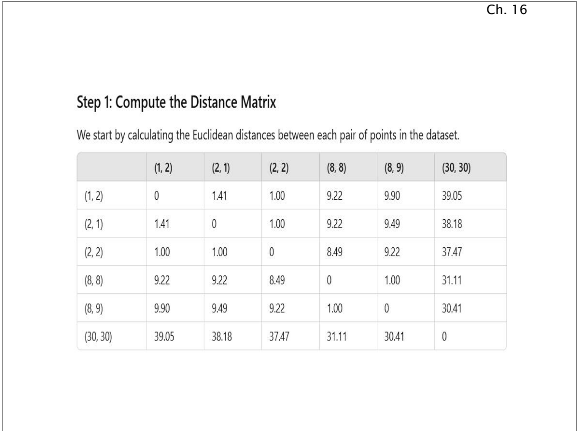
Let's use a simple dataset of 6 points in 2D space:

$$\text{Data} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 2 & 2 \\ 8 & 8 \\ 8 & 9 \\ 30 & 30 \end{bmatrix}$$

In this dataset, the points `[1, 2]`, `[2, 1]`, `[2, 2]`, `[8, 8]`, and `[8, 9]` are relatively close to each other, while `[30, 30]` is farther away and likely to be an outlier.

Parameters

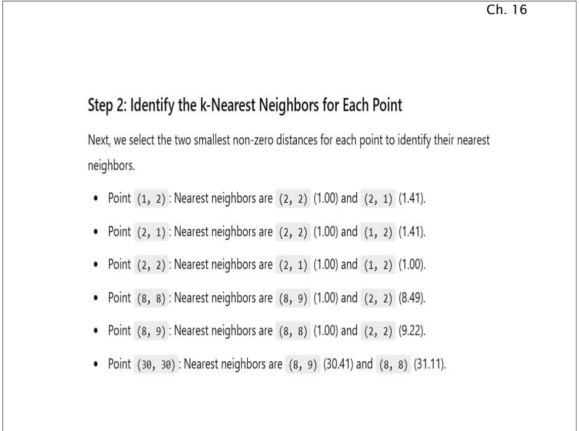
- **k** = 2 (number of nearest neighbors to consider)
- **Threshold** = 90th percentile of the average k-nearest neighbor distances.



Step 1: Compute the Distance Matrix

We start by calculating the Euclidean distances between each pair of points in the dataset.

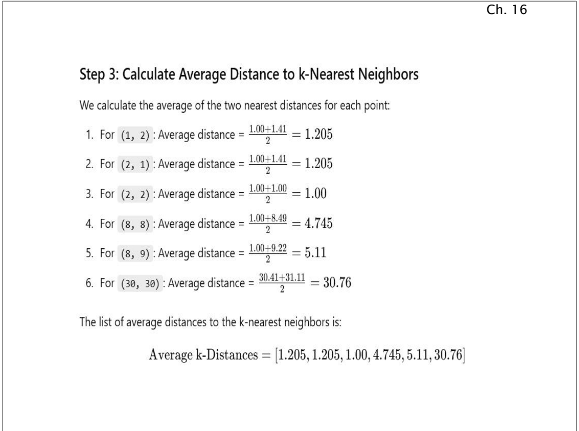
	(1, 2)	(2, 1)	(2, 2)	(8, 8)	(8, 9)	(30, 30)
(1, 2)	0	1.41	1.00	9.22	9.90	39.05
(2, 1)	1.41	0	1.00	9.22	9.49	38.18
(2, 2)	1.00	1.00	0	8.49	9.22	37.47
(8, 8)	9.22	9.22	8.49	0	1.00	31.11
(8, 9)	9.90	9.49	9.22	1.00	0	30.41
(30, 30)	39.05	38.18	37.47	31.11	30.41	0



Step 2: Identify the k-Nearest Neighbors for Each Point

Next, we select the two smallest non-zero distances for each point to identify their nearest neighbors.

- Point `(1, 2)`: Nearest neighbors are `(2, 2)` (1.00) and `(2, 1)` (1.41).
- Point `(2, 1)`: Nearest neighbors are `(2, 2)` (1.00) and `(1, 2)` (1.41).
- Point `(2, 2)`: Nearest neighbors are `(2, 1)` (1.00) and `(1, 2)` (1.00).
- Point `(8, 8)`: Nearest neighbors are `(8, 9)` (1.00) and `(2, 2)` (8.49).
- Point `(8, 9)`: Nearest neighbors are `(8, 8)` (1.00) and `(2, 2)` (9.22).
- Point `(30, 30)`: Nearest neighbors are `(8, 9)` (30.41) and `(8, 8)` (31.11).



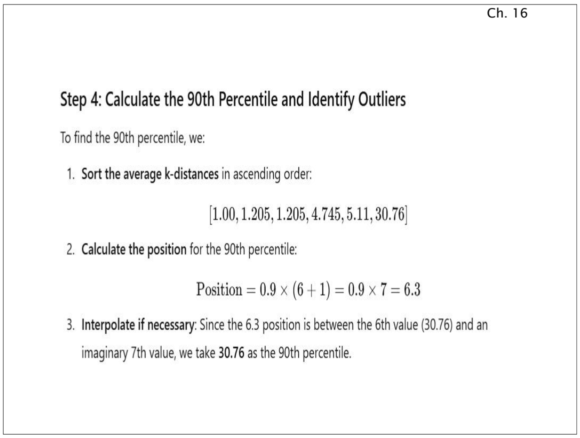
Step 3: Calculate Average Distance to k-Nearest Neighbors

We calculate the average of the two nearest distances for each point:

1. For `(1, 2)`: Average distance = $\frac{1.00+1.41}{2} = 1.205$
2. For `(2, 1)`: Average distance = $\frac{1.00+1.41}{2} = 1.205$
3. For `(2, 2)`: Average distance = $\frac{1.00+1.00}{2} = 1.00$
4. For `(8, 8)`: Average distance = $\frac{1.00+8.49}{2} = 4.745$
5. For `(8, 9)`: Average distance = $\frac{1.00+9.22}{2} = 5.11$
6. For `(30, 30)`: Average distance = $\frac{30.41+31.11}{2} = 30.76$

The list of average distances to the k-nearest neighbors is:

$$\text{Average k-Distances} = [1.205, 1.205, 1.00, 4.745, 5.11, 30.76]$$



Step 4: Calculate the 90th Percentile and Identify Outliers

To find the 90th percentile, we:

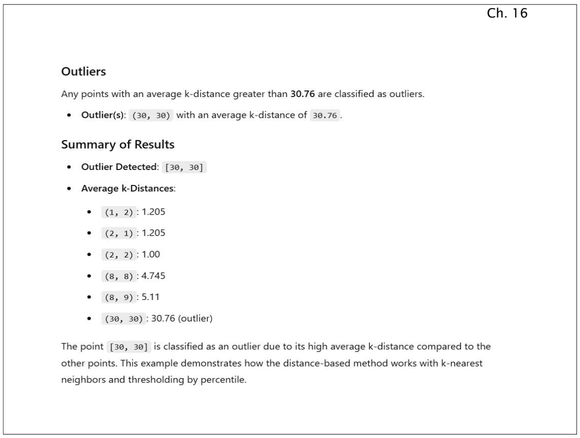
1. **Sort the average k-distances** in ascending order:

$$[1.00, 1.205, 1.205, 4.745, 5.11, 30.76]$$

2. **Calculate the position** for the 90th percentile:

$$\text{Position} = 0.9 \times (6+1) = 0.9 \times 7 = 6.3$$

3. **Interpolate if necessary**: Since the 6.3 position is between the 6th value (30.76) and an imaginary 7th value, we take **30.76** as the 90th percentile.



Outliers

Any points with an average k-distance greater than **30.76** are classified as outliers.

- **Outlier(s)**: `(30, 30)` with an average k-distance of `30.76`.

Summary of Results

- **Outlier Detected**: `[30, 30]`
- **Average k-Distances**:
  - `(1, 2)`: 1.205
  - `(2, 1)`: 1.205
  - `(2, 2)`: 1.00
  - `(8, 8)`: 4.745
  - `(8, 9)`: 5.11
  - `(30, 30)`: 30.76 (outlier)

The point `[30, 30]` is classified as an outlier due to its high average k-distance compared to the other points. This example demonstrates how the distance-based method works with k-nearest neighbors and thresholding by percentile.

To understand LOF, we have to learn a few concepts sequentially

1. K-distance and K-neighbors
2. Reachability distance (RD)
3. Local reachability density (LRD)
4. Local Outlier Factor (LOF)

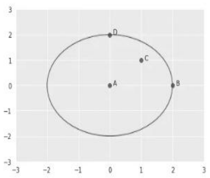
What is Outlier?

- Local Outlier Factor (LOF) is an algorithm that identifies the outliers present in the dataset.
- But what does the local outlier mean?
- When a point is considered as an outlier based on its local neighborhood, it is a local outlier.
- LOF will identify an outlier considering the density of the neighborhood.
- LOF performs well when the density of the data is not the same throughout the dataset.

K-DISTANCE AND K-NEIGHBORS

- K-distance is the distance between the point, and it's K^{th} nearest neighbor. K-neighbors denoted by $N_k(A)$ includes a set of points that lie in or on the circle of radius K-distance.
- K-neighbors can be more than or equal to the value of K. How's this possible?

K-DISTANCE AND K-NEIGHBORS



K-distance of A with K=2

If K=2, K-neighbors of A will be C, B, and D. Here, the value of K=2 but the $||N_2(A)|| = 3$. Therefore, $||N_k(\text{point})||$ will always be greater than or equal to K.

REACHABILITY DENSITY (RD)

$$\mathrm{RD}(X_i, X_j) = \max\left(K- \text{ distance } (X_j), \text{ distance } (X_i, X_j)\right)$$

It is defined as the maximum of K-distance of Xj and the distance between Xi and Xj. The distance measure is problem-specific (Euclidean, Manhattan, etc.)

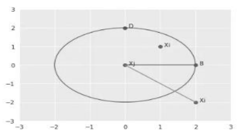
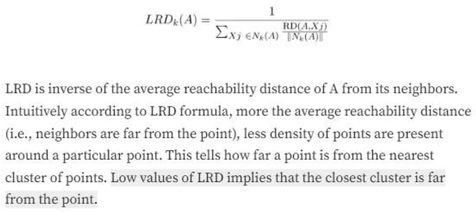


Illustration of reachability distance with K=2

In layman terms, if a point Xi lies within the K-neighbors of Xj, the reachability distance will be K-distance of Xj (blue line), else reachability distance will be the distance between Xi and Xj (orange line).

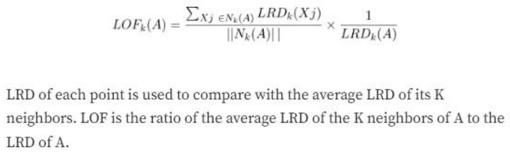
LOCAL REACHABILITY DENSITY (LRD)



$$LRD_k(A) = \frac{1}{\sum_{Xj \in N_k(A)} \frac{\text{RD}(A,Xj)}{\|N_k(A)\|}}$$

LRD is inverse of the average reachability distance of A from its neighbors. Intuitively according to LRD formula, more the average reachability distance (i.e., neighbors are far from the point), less density of points are present around a particular point. This tells how far a point is from the nearest cluster of points. Low values of LRD implies that the closest cluster is far from the point.

LOCAL OUTLIER FACTOR (LOF)

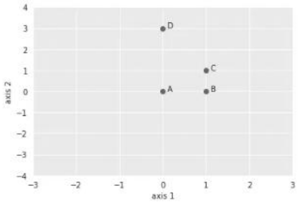


$$LOF_k(A) = \frac{\sum_{Xj \in N_k(A)} LRD_k(Xj)}{||N_k(A)||} \times \frac{1}{LRD_k(A)}$$

LRD of each point is used to compare with the average LRD of its K neighbors. LOF is the ratio of the average LRD of the K neighbors of A to the LRD of A.

Outlier Example?

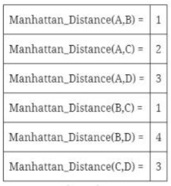
- 4 points: A(0,0), B(1,0), C(1,1) and D(0,3) and K=2. We will use **LOF** to detect one outlier among these 4 points.



A(0,0), B(1,0), C(1,1), and D(0,3)

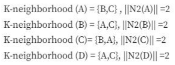
STEP:1(4 points: A(0,0), B(1,0), C(1,1) and D(0,3) and K=2)

- First, calculate the K-distance, distance between each pair of points, and K-neighborhood of all the points with **K=2.**
- We will be using Manhattan distance as a measure of distance.

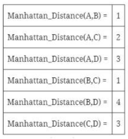


Manhattan_Distance(A,B) =	1
Manhattan_Distance(A,C) =	2
Manhattan_Distance(A,D) =	3
Manhattan_Distance(B,C) =	1
Manhattan_Distance(B,D) =	4
Manhattan_Distance(C,D) =	3

Outlier Example?

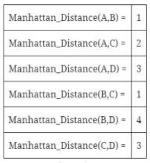


K-neighborhood (A) = {B,C} , ||N2(A)|| =2
K-neighborhood (B) = {A,C}, ||N2(B)|| =2
K-neighborhood (C)= {B,A}, ||N2(C)|| =2
K-neighborhood (D) = {A,C}, ||N2(D)|| =2

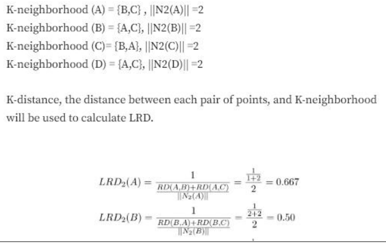


Manhattan_Distance(A,B) =	1
Manhattan_Distance(A,C) =	2
Manhattan_Distance(A,D) =	3
Manhattan_Distance(B,C) =	1
Manhattan_Distance(B,D) =	4
Manhattan_Distance(C,D) =	3

Ou



Manhattan_Distance(A,B) =	1
Manhattan_Distance(A,C) =	2
Manhattan_Distance(A,D) =	3
Manhattan_Distance(B,C) =	1
Manhattan_Distance(B,D) =	4
Manhattan_Distance(C,D) =	3



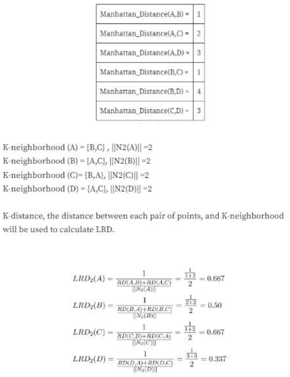
K-neighborhood (A) = {B,C} , ||N2(A)|| =2
K-neighborhood (B) = {A,C}, ||N2(B)|| =2
K-neighborhood (C)= {B,A}, ||N2(C)|| =2
K-neighborhood (D) = {A,C}, ||N2(D)|| =2

K-distance, the distance between each pair of points, and K-neighborhood will be used to calculate LRD.

$$LRD_2(A) = \frac{1}{\frac{RD(A,B)+RD(A,C)}{||N_2(A)||}} = \frac{\frac{1}{1+2}}{2} = 0.667$$

$$LRD_2(B) = \frac{1}{\frac{RD(B,A)+RD(B,C)}{||N_2(B)||}} = \frac{\frac{1}{2+2}}{2} = 0.50$$

Outlier Exam



Manhattan_Distance(A,B) =	1
Manhattan_Distance(A,C) =	2
Manhattan_Distance(A,D) =	3
Manhattan_Distance(B,C) =	1
Manhattan_Distance(B,D) =	4
Manhattan_Distance(C,D) =	3

K-neighborhood (A) = {B,C} , ||N2(A)|| =2
K-neighborhood (B) = {A,C}, ||N2(B)|| =2
K-neighborhood (C)= {B,A}, ||N2(C)|| =2
K-neighborhood (D) = {A,C}, ||N2(D)|| =2

K-distance, the distance between each pair of points, and K-neighborhood will be used to calculate LRD.

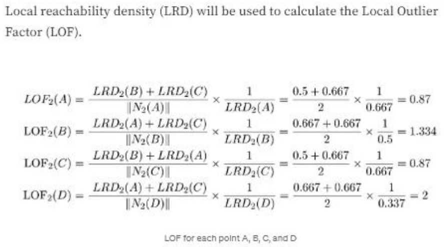
$$LRD_2(A) = \frac{1}{\frac{RD(A,B)+RD(A,C)}{||N_2(A)||}} = \frac{\frac{1}{1+2}}{2} = 0.667$$

$$LRD_2(B) = \frac{1}{\frac{RD(B,A)+RD(B,C)}{||N_2(B)||}} = \frac{\frac{1}{2+2}}{2} = 0.50$$

$$LRD_2(C) = \frac{1}{\frac{RD(C,B)+RD(C,A)}{||N_2(C)||}} = \frac{\frac{1}{1+2}}{2} = 0.667$$

$$LRD_2(D) = \frac{1}{\frac{RD(D,A)+RD(D,C)}{||N_2(D)||}} = \frac{\frac{1}{3+3}}{2} = 0.337$$

Outlier Example?



Local reachability density (LRD) will be used to calculate the Local Outlier Factor (LOF).

$$LOF_2(A) = \frac{LRD_2(B)+LRD_2(C)}{\|N_2(A)\|} \times \frac{1}{LRD_2(A)} = \frac{0.5+0.667}{2} \times \frac{1}{0.667} = 0.87$$

$$\mathrm{LOF}_2(B) = \frac{LRD_2(A)+LRD_2(C)}{\|N_2(B)\|} \times \frac{1}{LRD_2(B)} = \frac{0.667+0.667}{2} \times \frac{1}{0.5} = 1.334$$

$$\mathrm{LOF}_2(C) = \frac{LRD_2(B)+LRD_2(A)}{\|N_2(C)\|} \times \frac{1}{LRD_2(C)} = \frac{0.5+0.667}{2} \times \frac{1}{0.667} = 0.87$$

$$\mathrm{LOF}_2(D) = \frac{LRD_2(A)+LRD_2(C)}{\|N_2(D)\|} \times \frac{1}{LRD_2(D)} = \frac{0.667+0.667}{2} \times \frac{1}{0.337} = 2$$

LOF for each point A, B, C, and D

Highest LOF among the four points is LOF(D).
Therefore, D is an outlier.

Outlier Example?

Applications of Data Mining

Data mining has a wide range of applications across various industries, providing valuable insights by analyzing large datasets to uncover patterns, relationships, and trends. Here are some key applications of data mining:

1. **Customer Relationship Management (CRM)**

- **Objective:** Understand customer behavior, preferences, and purchasing patterns.
- **Example:** E-commerce companies use data mining to segment customers, predict customer churn, and design personalized marketing campaigns to enhance customer retention.

2. **Market Basket Analysis**

- **Objective:** Identify associations between products frequently bought together.
- **Example:** Retailers like supermarkets use data mining techniques such as association rule mining (e.g., the Apriori algorithm) to recommend complementary products, optimize store layout, and increase cross-selling opportunities.

3. **Fraud Detection**

- **Objective:** Detect and prevent fraudulent activities.
- **Example:** Financial institutions and credit card companies use data mining techniques such as anomaly detection and classification models to identify suspicious transactions and prevent fraud in real-time.

4. **Healthcare and Medical Diagnosis**

- **Objective:** Predict patient outcomes, detect diseases early, and enhance treatment plans.
- **Example:** Hospitals and healthcare providers use data mining to analyze patient records, medical history, and symptoms to predict the likelihood of diseases such as diabetes or cancer. Machine learning algorithms can also identify patterns that help in drug discovery and treatment optimization.

5. **Sentiment Analysis and Social Media Analytics**

- **Objective:** Analyze customer sentiment, public opinion, and social media trends.
- **Example:** Businesses use sentiment analysis on social media platforms to gauge customer reactions to new products, services, or marketing campaigns. It helps in brand monitoring, crisis management, and reputation management.

6. **Financial Forecasting and Risk Management**

- **Objective:** Predict financial trends, assess risks, and optimize investment strategies.
- **Example:** Investment firms and banks use data mining to analyze stock market trends, forecast economic conditions, and detect financial risks in lending, trading, or investments.

7. **Recommender Systems**

- **Objective:** Provide personalized recommendations to users based on their preferences and behavior.
- **Example:** Streaming services like Netflix and Spotify use data mining techniques such as collaborative filtering to recommend movies, music, or shows based on the user's past behavior and preferences.

8. **Supply Chain Optimization**

- **Objective:** Optimize logistics, inventory, and supplier relationships.
- **Example:** Data mining helps manufacturers and retailers analyze sales trends, forecast demand, and optimize inventory levels. It improves the efficiency of supply chain operations by predicting supply bottlenecks and optimizing delivery routes.

9. **Telecommunications**

- **Objective:** Reduce customer churn and optimize network performance.
- **Example:** Telecom companies use data mining to analyze usage patterns, detect potential churn, and predict network congestion. They also use predictive models to recommend personalized plans to their customers based on their data usage patterns.

10. **Education**

- **Objective:** Improve student performance and curriculum design.
- **Example:** Educational institutions use data mining to analyze student performance data, attendance, and engagement in online learning platforms. It helps identify at-risk students and develop intervention strategies for improved learning outcomes.

11. **Energy Management**

- **Objective:** Optimize energy consumption and reduce costs.
- **Example:** Smart grid systems use data mining to analyze energy consumption patterns, predict peak usage times, and recommend energy-saving measures to households and businesses. This also helps utility companies optimize their distribution and generation processes.

12. **Cybersecurity**

- **Objective:** Detect and prevent cybersecurity threats.
- **Example:** Data mining techniques like anomaly detection and clustering are used to analyze network traffic and detect unusual patterns indicative of cyber-attacks such as malware, phishing, or unauthorized access attempts.

13. **Retail Sales Forecasting**

- **Objective:** Predict future sales and trends.
- **Example:** Retailers use data mining to analyze historical sales data, seasonal trends, and promotional effects to predict future sales, optimize pricing strategies, and adjust inventory levels for peak demand periods.

14. **Recommendation Systems in E-learning**

- **Objective:** Provide personalized learning paths and materials.
- **Example:** E-learning platforms use data mining to recommend personalized learning content based on a student's learning progress, preferences, and interactions on the platform.

15. **Sports Analytics**

- **Objective:** Analyze performance data to improve strategies and player performance.
- **Example:** Sports teams use data mining to analyze player statistics, game footage, and performance metrics to develop strategies, assess player health, and improve team performance.

16. **Environmental and Ecological Studies**

- **Objective:** Predict environmental changes and assess ecological risks.
- **Example:** Data mining is used in climate science to analyze historical weather data, predict natural disasters like hurricanes and floods, and monitor ecological systems for changes in biodiversity or pollution levels.

17. **Real Estate and Property Valuation**

- **Objective:** Predict property prices and assess market trends.
- **Example:** Real estate firms use data mining to analyze property market trends, neighborhood development, and housing demand to predict future property values and investment opportunities.

Outlier Example?