Anomaly Detection

Anomaly Detection

- Anomaly Detection in Data Mining, also known as outlier detection, detects patterns in data that do not match the expected behavior.

- These anomalies might indicate unexpected network activity, reveal a malfunctioning sensor, or highlight data that has to be cleaned before analysis.

- Generally, anomalies are either removed before analysis or are thoroughly investigated to gain an in-depth understanding of data points that are out of standard patterns.

Types of Anomalies

Anomalies are classified as follows:

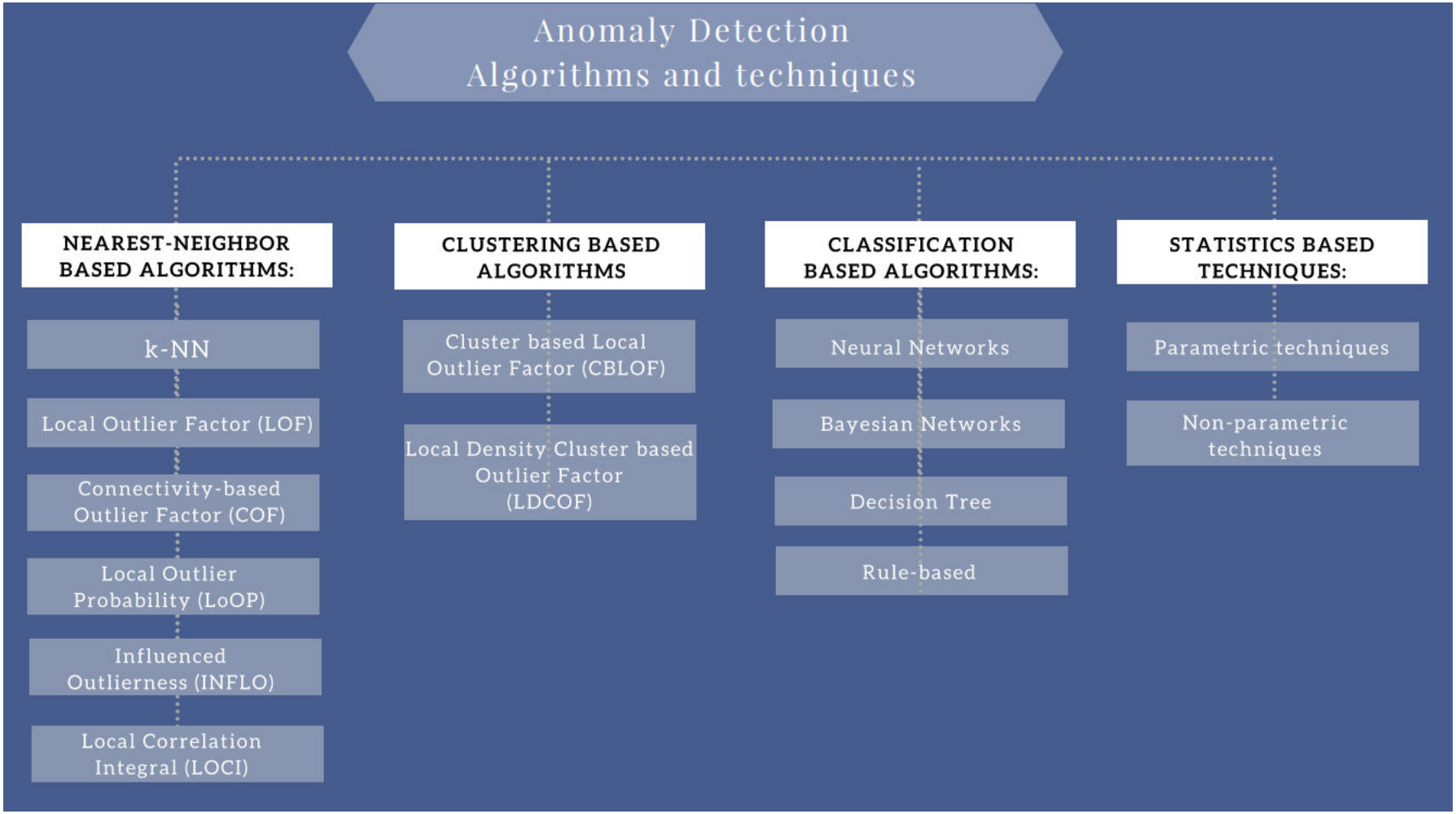
- Contextual Anomalies:
  - Anomalies that are situation-specific; the abnormality is context-based.
  - Usually, in time-series data, this form of aberration is prevalent.
  - For example, spending more on food every day during the holidays is normal, but it’s unusual otherwise.

Types of Anomalies

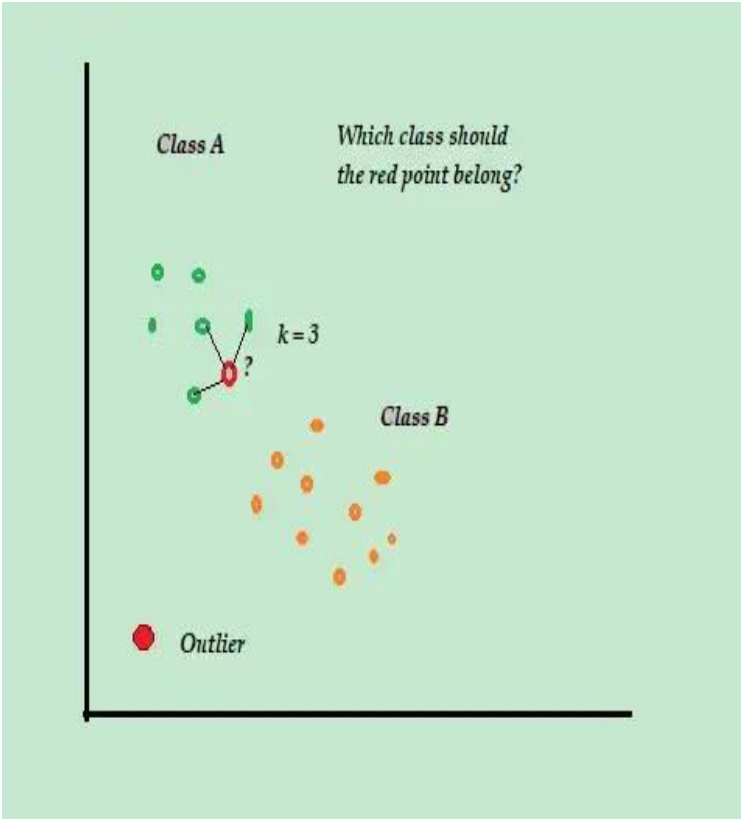
Anomalies are classified as follows:

- Collective Anomalies:
  - A group of data instances that occur together and do not show usual patterns are called collective anomalies.
  - In other words, the data points with the same behavior individually might not be an anomaly.
  - However, when they occur collectively, it is considered an anomaly.

Anomaly Detection Algorithms



K-NN Algorithm for Anomaly Detection

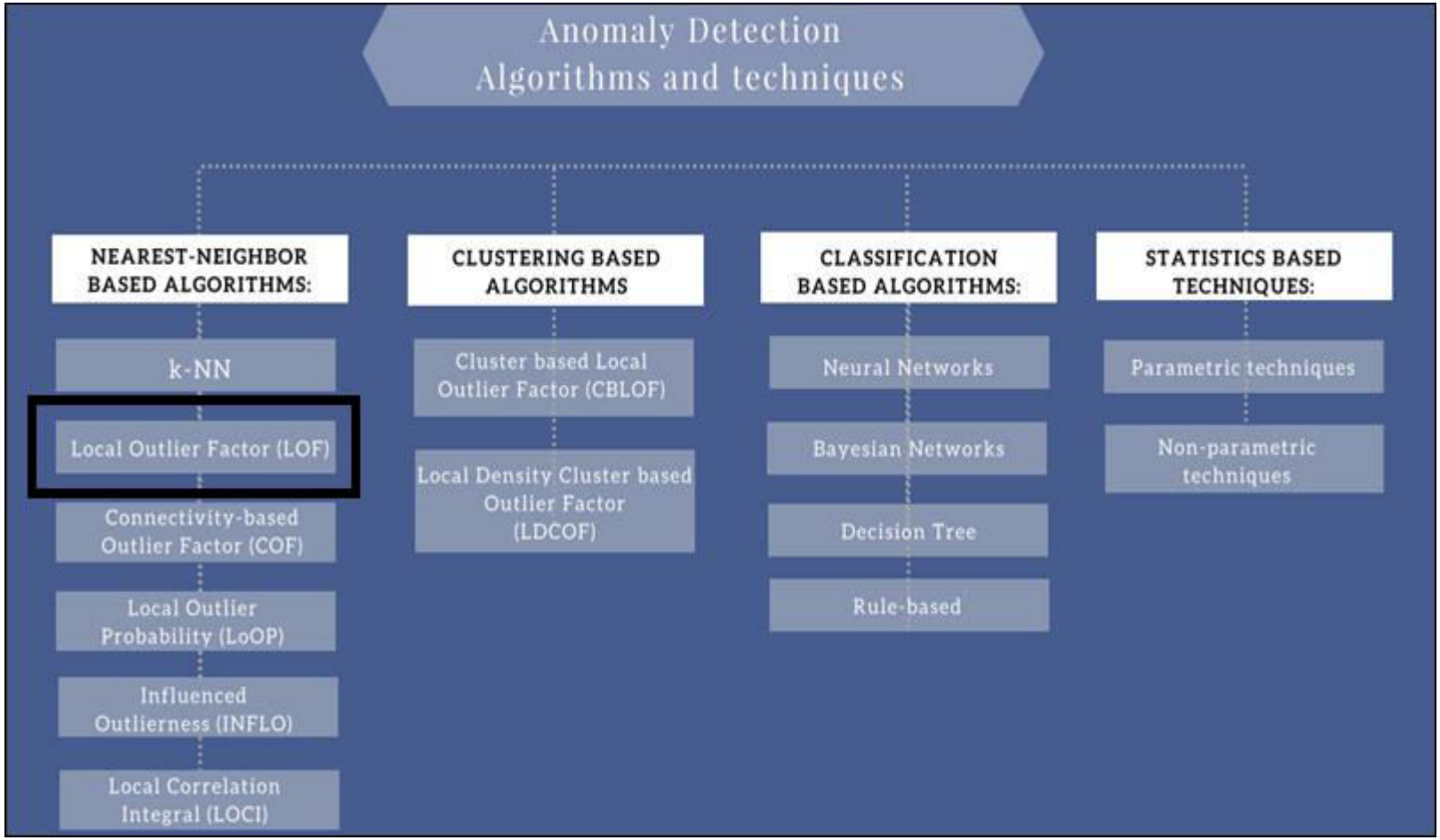


- Although kNN is a supervised ML algorithm, when it comes to anomaly detection it takes an unsupervised approach.

- This is because there is no actual "learning" involved in the process and there is no pre-determined labeling of "outlier" or "not-outlier" in the dataset, instead, it is entirely based upon threshold values.

- It has many applications in business and finance field. For example, k-NN helps for detecting and preventing credit card fraudulent transactions.

Anomaly Detection Algorithms



Local Outlier Factor (LOF)

- The LOF is a key anomaly detection algorithm based on a concept of a local density. It uses the distance between the k nearest neighbors to estimate the density.

- LOF compares the local density of an item to the local densities of its neighbors. Thus one can determine areas of similar density and items that have a significantly lower density than their neighbors. These are the outliers.

Local Outlier Factor (LOF) Algorithm

1. Calculate the reachability distance from all the neighbours (in the k-distance neighbourhood of the given point) to the given point.
2. Calculate the local reachability distance of the point.
3. Calculate the LOF score of the point using the local reachability of its neighbours and the point itself.
4. If LOF score>threshold, object is an outlier.
5. Repeat steps 1 through 4 for all points in the dataset.

Local Outlier Factor (LOF)

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

- K-distance and K-neighbors
- Reachability distance (RD)
- Local reachability density (LRD)
- Local Outlier Factor (LOF)

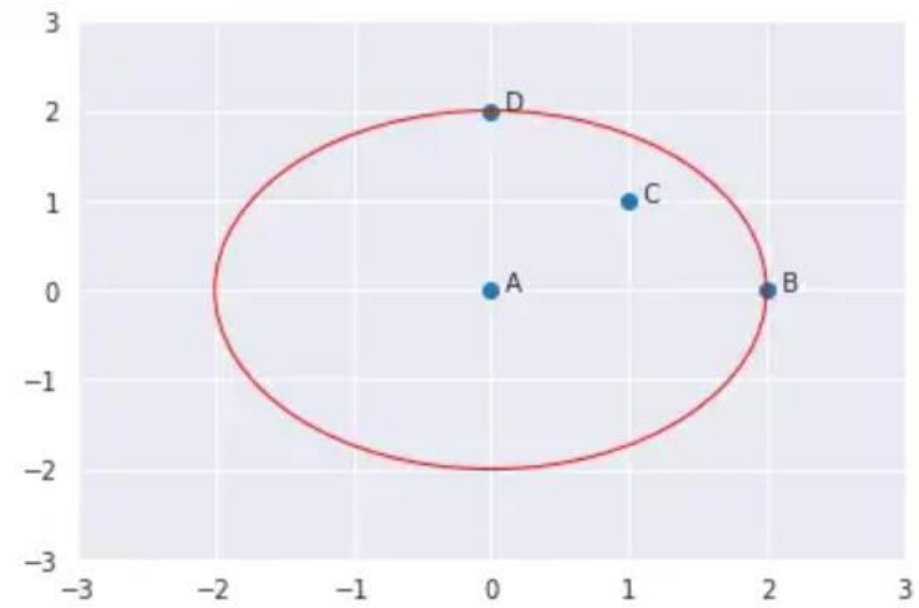
Local Outlier Factor (LOF)

- 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.

Local Outlier Factor (LOF)

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.



K-distance of A with K=2

Local Outlier Factor (LOF)

- **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.)

Local Outlier Factor (LOF)

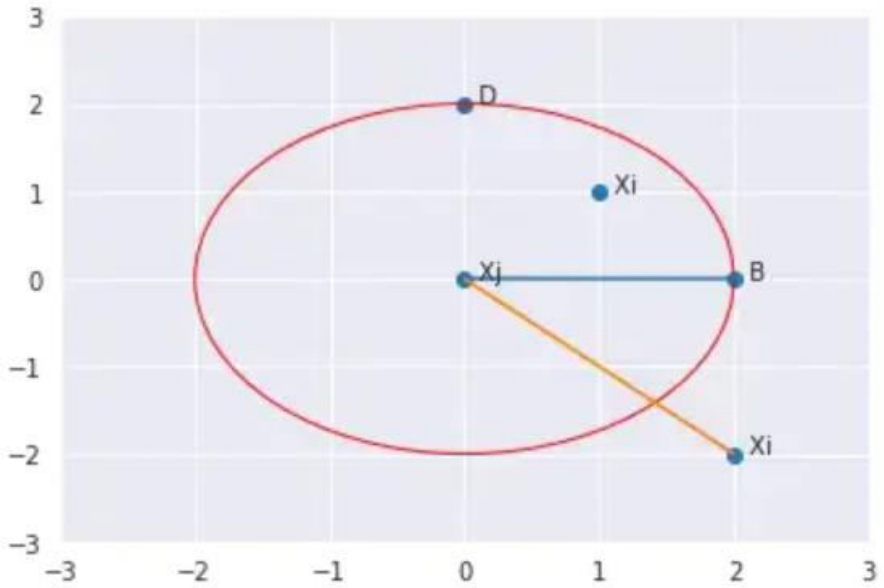


Illustration of reachability distance with K=2

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 Outlier Factor (LOF)

- LOCAL REACHABILITY DENSITY (LRD)

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

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.

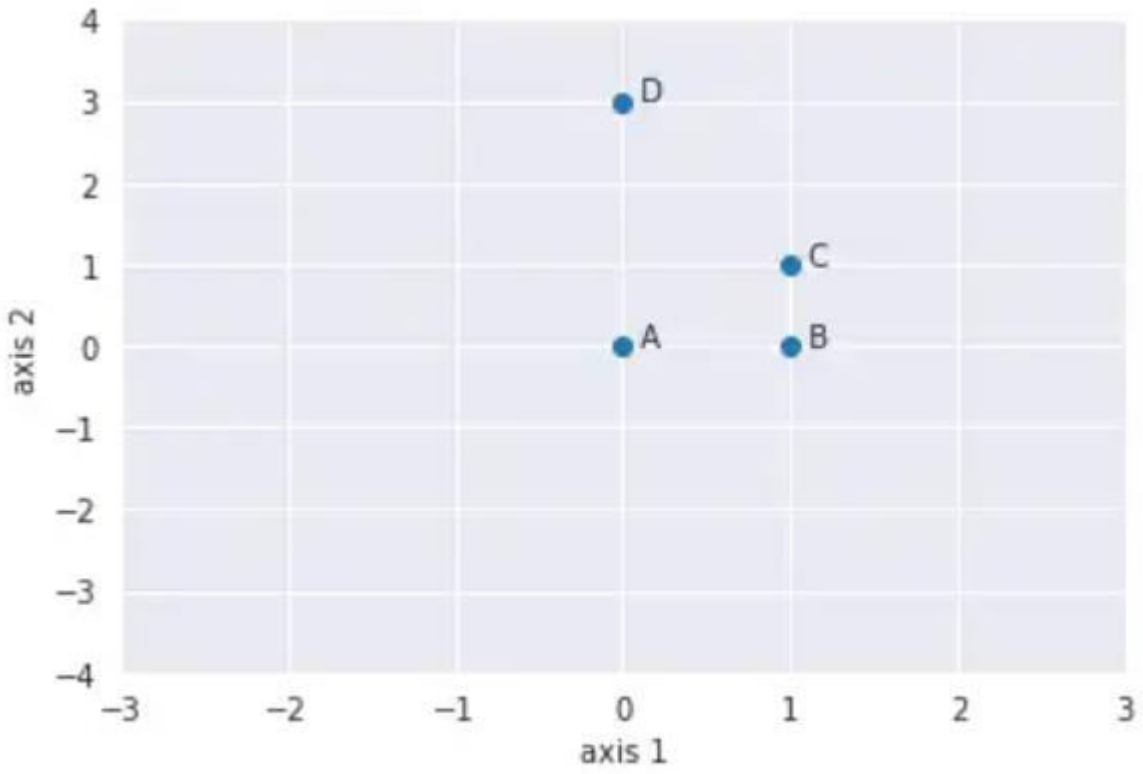
Local Outlier Factor (LOF)

- Intuitively, if the point is not an outlier (inlier), the ratio of average LRD of neighbors is approximately equal to the LRD of a point (because the density of a point and its neighbors are roughly equal).

- In that case, LOF is nearly equal to 1. On the other hand, if the point is an outlier, the LRD of a point is less than the average LRD of neighbors. Then LOF value will be high.

Local Outlier Factor (LOF)

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



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

Local Outlier Factor (LOF)

- Calculate the **K-distance**, **distance between each pair of points**, and **K-neighborhood** of all the points with K=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

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

Local Outlier Factor (LOF)

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

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

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

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

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

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

Local Outlier Factor (LOF)

- 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.

Local Outlier Factor (LOF)

- Advantage

  A point will be considered as an outlier if it is at a small distance to the extremely dense cluster.

  The global approach may not consider that point as an outlier. But the LOF can effectively identify the local outliers.

- Disadvantage

  Since LOF is a ratio, it is tough to interpret.

  There is no specific threshold value above which a point is defined as an outlier.

  The identification of an outlier is dependent on the problem and the user.