Local outlier Factor [LOF] method

The Local Outlier Factor (LOF) method detects outliers based on the local density of data points. A point is considered an outlier if it has a significantly lower density than its neighbours. LOF assigns a score to each data point, where a higher score indicates a higher likelihood of being an outlier.

Steps to calculate LOF.

1. Calculate the distance between points:
   First, compute the distances between each point and its neighbors.
2. Compute the local Reachability density (LRD):-
   For each point, determine its density by considering the distances to its neighbors.
3. Calculate the LOF score:-
   Compare the local density of each point with the density of its neighbors.

Points that have significantly lower density compared to their neighbors are flagged as outliers with higher LOF scores.

Solved example

Consider the 2D dataset representing points in a 2-dimensional space.

points : A(1,2), B(2,3), C(3,4), D(10,10)

Calculate LOF score for each point using K=2 nearest neighbors

Step1: Calculate the Euclidean Distances

The Euclidean distance b/w two points (x_1, y_1) and (x_2, y_2)

$$\text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Distances b/w points

* $A(1,2)$ to $B(2,3)$:

$$\text{distance}(A,B) = \sqrt{(2-1)^2 + (3-2)^2} = \sqrt{1^2 + 1^2} = \sqrt{2} \approx 1.414$$

* $A(1,2)$ to $C(3,4)$:

$$\text{distance}(A,C) = \sqrt{(3-1)^2 + (4-2)^2} = \sqrt{2^2 + 2^2} = \sqrt{8} \approx 2.828$$

* $A(1,2)$ to $D(10,10)$:

$$\text{distance}(A,D) = \sqrt{(10-1)^2 + (10-2)^2} = \sqrt{9^2 + 8^2} = \sqrt{81 + 64} = \sqrt{145} \approx 12.042$$

• B(2,3) to C(3,4):

distance (B, C) = $\sqrt{(3-2)^2+(4-3)^2} = \sqrt{1^2+1^2} = \sqrt{2}$
≈ 1.414

• B(2,3) to D(10,10):

distance (B, D) = $\sqrt{(10-2)^2+(10-3)^2}$

$= \sqrt{8^2+7^2} = \sqrt{64+49} = \sqrt{113}$
≈ 10.630

• C(3,4) to D(10,10):

distance (C, D) = $\sqrt{(10-3)^2+(10-4)^2} = \sqrt{7^2+6^2}$

$= \sqrt{49+36} = \sqrt{85} \approx 9.220$

Step 2

Find the K Nearest-Neighbors for each point for K=2, we need to find the two Nearest Neighbors for each point.

• A(1,2):

Nearest Neighbors: B(2,3) and C(3,4).

Distances: [1.414, 2.828].

- B(2,3)

  Nearest Neighbors : A(1,2) and C(3,4)

  Distances : [1.414 , 1.414]

- C(3,4) :

  Nearest neighbors : B(2,3) and A(1,2)

  Distances : [1.414 , 2.828]

- D(10,10) :

  Nearest Neighbors : A(1,2) and B(2,3)

  Distances : [12.042 , 10.630]

Step 3.

Calculate the Reachability Distances

The reachability distance b/w two points p and q is calculated as :-

$$\boxed{\text{reach-dist}(p,q) = \max(\text{dist}(p,q), \text{k-distance}(q))}$$

where k-distance(q) is the distance to the kth ~~Neighbor~~ Nearest neighbor of q

For each point, we will calculate the reachability distance to its 2 nearest neighbors.

For point A(1,2):

* Reachability distance to B(2,3):
  reach-dist (A,B) = max(1.414, 1.414) = 1.414.
* Reachability distance to C(3,4):
  reach-dist (A,C) = max(2.828, $\overset{1.414}{\cancel{2.828}}$) = 2.828.

For point B(2,3):

* Reachability distance to A(1,2):
  reach-dist (B,A) = max (1.414, 1.414) = 1.414.
* Reachability distance to C(3,4):

For point C(3,4):

* Reachability distance to B(2,3).
  reach-dist (C,B) = max (1.414, 1.414) = 1.414
* Reachability distance to A(1,2).
  reach-dist (C,A) = max.

For point D(10,10)

* Reachability distance to A(1,2):
  reach-dist (D,A) = max (12.042, 2.828) = 12.042,
* Reachability distance to B(2,3).
  reach-dist (D,B) = max(10.630, 1.414) = 10.630

Step 4: The Local Reachability Density (LRD) is calculated as the inverse of the average reachability distance of a point's k-nearest neighbours.

$$LRD(p) = \frac{1}{\frac{1}{K}\sum_{q \in N_k(p)} \text{reach-dist}(p,q)}$$

For point A (1,2):

- The reachability distances to the neighbors are: [1.414, 2.828].
- the average reachability distance: $\frac{1.414+2.828}{2}$ $= 2.121$.
- The LRD of A is:

$$LRD(A) = \frac{1}{2.121} \approx 0.471$$

For point B (2,3)

- The reachability distances to the neighbors are: [1.414, 1.414].
- The average reachability distance: $\frac{1.414+1.414}{2}$ $= 1.414$.

The LRD for B is

$$LRD(B) = \frac{1}{1.414} \approx \underline{\underline{0.707}}$$

For point C (3, 4):

- The reachability distances to the neighbors are [1.414, 2.828].
- The average reachability distance: $\frac{1.414 + 2.828}{2}$
  $= 2.121$.
- The LRD for C is:

$$LRD(C) = \frac{1}{2.121} \approx \underline{\underline{0.471}}.$$

For point D (10, 10).

- The reachability distances to the neighbors are:- [12.042, 10.630].
- The average reachability distance:-

$$\frac{12.042 + 10.630}{2} = \underline{\underline{11.336}}.$$

- The LRD for D is:-

$$LRD(D) = \frac{1}{11.336} \approx \underline{\underline{0.088}}.$$

Step5 :- Calculate LOF scores

The LOF score for a point p is the ratio of the average LRD of its neighbors to its own LRD.

$$LOF(p) = \frac{\frac{1}{K} \sum_{q \in N_K(p)} LRD(q)}{LRD(p)}$$

For point A (1,2):

- The neighbors are B and C with LRD values: [0.707, 0.471].
- The LOF for A is:-

$$LOF(A) = \frac{0.707 + 0.471}{0.471} \approx 2.49$$

For point B (2,3):

The neighbours are A and C with LRD values: [0.471, 0.471].

The LOF for B is:-

$$LOF(B) = \frac{0.471 + 0.471}{0.707} \approx 1.33$$

For point C (3,4):

- The neighbors are B and A with LRD values: [0.707, 0.471].
- The LOF for C is

$$LOF(C) = \frac{0.707 + 0.471}{0.471} \approx \underline{\underline{2.49}}.$$

For point D (10,10)

- The neighbors are A and B with LRD values : [0.471, 0.707].
- The LOF for D is :-

$$LOF(D) = \frac{0.471 + 0.707}{0.088} \approx \underline{\underline{13.43}}.$$

~~Conc~~

point	LRD	LOF Scores
A	0.471	2.49
B	0.707	1.33
C	0.471	2.49
D	0.088	13.43

- point D has the highest LOF score [13.43], indicating it is a clear outlier.

- points A and C have similar LOF scores (2.49), suggesting that they are somewhat outlier-like, but less so than D.
- point B has the lowest LOF score (1.33), indicating it is not an outlier.

Z-Score based Outlier or Anomaly detection

Suppose we have a dataset of daily sales revenue for a retail store over the past 30days.

- [100, 150, 120, 125, 140, 130, 110, 135, 130, 150, 140, 100, 95, 80, 120, 125, 130, 100, 140, 135, 130, 145, 110, 120, 130, 135, 140, 125, 130, 120].
- We want to identify any days where the sales revenue is significantly different from the other days, which may indicate an anomaly or outlier.

Step 1 :

Mean : $\mu = \sum_{i=1}^{30} \frac{x_i}{n} = \frac{100+150+120+\ldots+130+120}{30}$

$= \underline{\underline{123.5}}$

standard deviation :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{30} (x_i - \mu)^2}{n-1}}$$

$$= \sqrt{\frac{(100-123.5)^2 + (150-123.5)^2 + \ldots (126-123.5)^2}{30-1}}$$

$= \underline{\underline{20}}$.

Step 2

Calculate z-scores

We then calculate the z-score for each data point, which represents how many standard deviations away from the mean the data point is:-

- $z = \frac{x - \mu}{\sigma}$

- $z = \frac{100 - 123.5}{20} = \underline{\underline{-1.17}}$

- [-1.17, 0.14, -0.28, -0.16, 0.56, 0.08, -0.89, 0.32, 0.08, ~~-0.89,~~ 0.14, 0.56, -1.17, -1.45, -2.22, -0.28, -0.16, 0.08, -1.17, 0.56, 0.32, 0.08, 0.86, -0.89, -0.28, 0.08, 0.32, 0.56, -0.16, 0.08, -0.28].

Step 3: Set a threshhold.

- z score = 3 is considered as a cut-off value to set the limit, which captures 99.7% of the data in a normal distribution

∴ Any z-score greater than +3 or less than -3 is considered as outlier which is pretty much similar to standard deviation method.

Step 4: Identify anomalies

[−1.17, 0.14, −0.28, −0.16, 0.56, 0.08,
−0.89, 0.32, 0.08, 0.14, 0.56, −1.17,
−1.45, −2.22, −0.28, −0.16, 0.08,
−1.17, 0.56, 0.32, 0.08, 0.86, −0.89,
−0.28, 0.08, 0.32, 0.56, −0.16, 0.08, −0.28]

No outliers found in the given data.