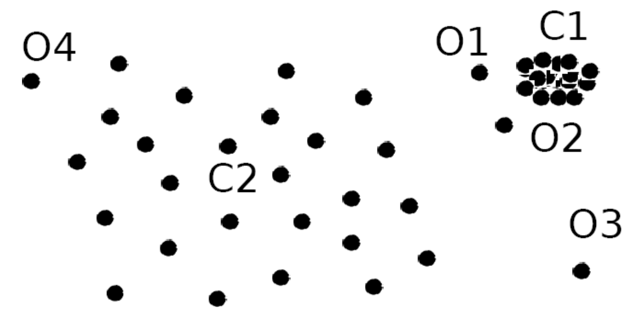
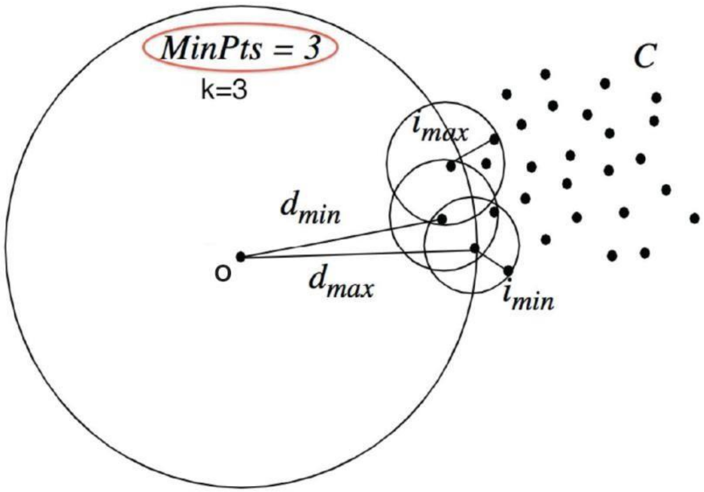
Density-Based Outlier Detection

- Local outliers: Outliers comparing to their local neighborhoods, instead of the global data distribution
- In Fig., o_1 and o2 are local outliers to C_1, o_3 is a global outlier, but o_4 is not an outlier. However, proximity-based clustering cannot find o_1 and o_2 are outlier (e.g., comparing with O_4).



- Intuition (density-based outlier detection): The density around an outlier object is significantly different from the density around its neighbors
- Method: Use the relative density of an object against its neighbors as the indicator of the degree of the object being outliers
- *k-distance* of an object o, $dist_k(o)$: distance between o and its k-th NN
- *k-distance neighborhood* of o, $N_k(o) = \{o' \mid o' \text{ in } D, dist(o, o') \leq dist_k(o)\}$
  - $N_k(o)$ could be bigger than k since multiple objects may have identical distance to o

Local Outlier Factor: LOF



- Reachability distance from *o'* to *o*:

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

  – where k is a user-specified parameter

- Local reachability density of *o*:

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$

- LOF (Local outlier factor) of an object o is the average of the ratio of local reachability of *o* and those of *o*'s k-nearest neighbors

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \frac{\sum_{o' \in N_k(o)} lrd_k(o') * \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}{||N_k(o)||^2}$$

- The lower the local reachability density of o, and the higher the local reachability density of the kNN of o, the higher LOF
- This captures a local outlier whose local density is relatively low comparing to the local densities of its kNN

LOF(Local Outlier Factor) Example

- Consider the following 4 data points: a(0, 0), b(0, 1), c(1, 1), d(3, 0)

Calculate the LOF for each point and show the top 1 outlier, set k = 2 and use Manhattan Distance.

Step 1: calculate all the distances between each two data points

- There are 4 data points:

  a(0, 0), b(0, 1), c(1, 1), d(3, 0)

(Manhattan Distance here)

dist(a, b) = 1
dist(a, c) = 2
dist(a, d) = 3
dist(b, c) = 1
dist(b, d) = 3+1=4
dist(c, d) = 2+1=3

Step 2: calculate all the $dist_2(o)$

- **$dist_k(o)$: distance between o and its k-th NN(k-th nearest neighbor)**

$dist_2(a)$ **= dist(a, c) = 2** ***(c is the 2nd nearest neighbor)***

$dist_2(b)$ **= dist(b, a) = 1** ***(a/c is the 2nd nearest neighbor)***

$dist_2(c)$ **= dist(c, a) = 2** ***(a is the 2nd nearest neighbor)***

$dist_2(d)$ **= dist(d, a) = 3** ***(a/c is the 2nd nearest neighbor)***

Step 3: calculate all the $N_k(o)$

- *k-distance neighborhood* of o, $\mathbf{N_k(o) = \{o' | o' \text{ in } D, dist(o, o') \leq dist_k(o)\}}$

$\mathbf{N_2(a)} =$ {b, c}

$\mathbf{N_2(b)} =$ {a, c}

$\mathbf{N_2(c)} =$ {b, a}

$\mathbf{N_2(d)} =$ {a, c}

Step 4: calculate all the $lrd_k(o)$

- $lrd_k(o)$: Local Reachability Density of o

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

$\| N_k(o) \|$ means the number of objects in $N_k(o)$,
For example: $\| N_2(a) \| = \| \{b, c\} \| = 2$

$$lrd_k(a) = \frac{\| N_2(a) \|}{reachdist_2(b \leftarrow a) + reachdist_2(c \leftarrow a)}$$

Step 4: calculate all the $lrd_k(o)$

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

$reachdist_2(b \leftarrow a) = max\{\mathrm{dist}_2(b), dist(b, a)\}$

$= \max\{1, 1\} = 1$

$reachdist_2(c \leftarrow a) = max\{\mathrm{dist}_2(c), dist(c, a)\}$

$= \max\{2, 2\} = 2$

Thus, $lrd_2(a)$

$$= \frac{\| N_2(a) \|}{reachdist_2(b \leftarrow a) + reachdist_2(c \leftarrow a)} = 2/(1+2) = 0.667$$

Step 4: calculate all the $lrd_k(o)$

Similarly,

$$lrd_2(b) = \frac{\| N_2(b) \|}{reachdist_2(a \leftarrow b) + reachdist_2(c \leftarrow b)} = 2/(2+2) = 0.5$$

$$lrd_2(c) = \frac{\| N_2(c) \|}{reachdist_2(b \leftarrow c) + reachdist_2(a \leftarrow c)} = 2/(1+2) = 0.667$$

$$lrd_2(d) = \frac{\| N_2(b) \|}{reachdist_2(a \leftarrow d) + reachdist_2(c \leftarrow d)} = 2/(3+3) = 0.33$$

Step 5: calculate all the $LOF_k(o)$

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \frac{\sum_{o' \in N_k(o)} lrd_k(o') * \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}{||N_k(o)||^2}$$

$LOF_2(a) = (lrd_2(b) + lrd_2(c)) * (reachdist_2(b \leftarrow a) + reachdist_2(c \leftarrow a)) / ||N_k(a)||^2$

$= (0.5+0.667) * (1+2) / 4 = 3.501 / 4 = 0.875$

$LOF_2(b) = (lrd_2(a) + lrd_2(c)) * (reachdist_2(a \leftarrow b) + reachdist_2(c \leftarrow b)) / ||N_k(b)||^2$

$= (0.667+0.667) * (2+2) / 4 = 5.336 / 4 = 1.334$

$LOF_2(c) = (lrd_2(b) + lrd_2(a)) * (reachdist_2(b \leftarrow c) + reachdist_2(a \leftarrow c)) / ||N_k(c)||^2$

$= (0.5+0.667) * (1+2) / 4 = 3.501 / 4 = 0.875$

$LOF_2(d) = (lrd_2(a) + lrd_2(c)) * (reachdist_2(a \leftarrow d) + reachdist_2(c \leftarrow d)) / ||N_k(d)||^2$

$= (0.667+0.667) * (3+3) / 4 = 8.004 / 4 = 2.001$

Step 6: Sort all the $LOF_k(o)$

The sorted order is:

$LOF_2(\mathbf{d}) = 2.001$

$LOF_2(\mathbf{b}) = 1.334$

$LOF_2(\mathbf{a}) = 0.875$

$LOF_2(\mathbf{c}) = 0.875$

Obviously, top 1 outlier is point d.