Outline

- Basic concepts
- Statistical approaches
- Proximity-based approaches
  - Distance-based outlier detection
  - Density-based outlier detection
- Reconstruction-based approaches
- Clustering and classification based approaches

Proximity-Based Approaches

- This way of detecting outliers looks at how close data points are to each other.
- If a point is far away from most others, it's considered unusual or an outlier.
- Two types of proximity-based methods:
  - Distance-based: This method looks at each point and its nearest neighbors.
    - If the point is far from its neighbors beyond a certain limit (called a distance threshold), it's an outlier.
  - Density-based: This method checks how many points are around a particular point compared to its neighbors.
    - If there are a lot fewer points around it, it's an outlier.

Distance-based outlier detection

- A user sets a distance threshold, which is a way to say how far apart points should be before they're considered outliers.
- The method then checks each point to see how many other points are within this threshold distance.
- If a point has fewer neighbors than the set threshold percentage, it's an outlier.

Distance-Based Outlier Detection

- An object is an outlier if it does not have enough nearby points.
  - “Nearby” is determined by a distance threshold r .
  - “Enough” is determined by a fraction threshold π.
- Steps:
  1. For each object, count the points within distance r (its neighborhood).
  2. If this count is less than π × total points, the object is marked as an outlier
- Simplified formula: A point is an outlier if

$$\frac{Number\ of\ points\ within\ r}{total\ points} \leq \pi$$

Distance-Based Outlier Detection

- Let r ($r \geq 0$) be a distance threshold and
- π ($0 < \pi \leq 1$) be a fraction threshold
- where dist(·, ·) is a distance measure

$$\frac{\|\{o' | dist(o, o') \leq r\}\|}{\|D\|} \leq \pi,$$

Computational considerations:

- The basic way to find these outliers would be to compare every point with every other point, which can take a lot of time.
- However, in practice, this doesn't take as long as expected because the process can stop early for points that are clearly not outliers, which is most of them.

Distance-Based Outlier Detection - Algorithm

- It goes through each item in the list
- For each item, it starts a count at 0.
- Then, it compares this item to every other item to see how many are within the distance r.
- If an item has enough neighbors (at least the number decided by the percentage π), it's not an outlier.
- If an item doesn't have enough neighbors, then it's marked as an outlier.
- At the end, it tells you which items are the outliers.

Input:

- a set of objects $D = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_n\}$, threshold r $(r > 0)$ and π $(0 < \pi \leq 1)$;

Output: $DB(r, \pi)$ outliers in D.
Method:

for $i = 1$ to n **do**
  $count \leftarrow 0$
  for $j = 1$ to n **do**
    if $i \neq j$ **and** $dist(\boldsymbol{o}_i, \boldsymbol{o}_j) \leq r$ **then**
      $count \leftarrow count + 1$
      if $count \geq \pi \cdot n$ **then**
        exit {$\boldsymbol{o}_i$ cannot be a $DB(r, \pi)$ outlier}
      endif
    endif
  endfor
  print $\boldsymbol{o}_i$ {$\boldsymbol{o}_i$ is a $DB(r, \pi)$ outlier according to (Eq. 11.10)}
endfor;

Distance-based vs Distance-based detection

Distance-based outliers

- This method looks at the whole dataset to find outliers.
- An object is an outlier if it's far away from a certain percentage of all other objects.
- These are called "global outliers" because the method looks at the entire dataset.

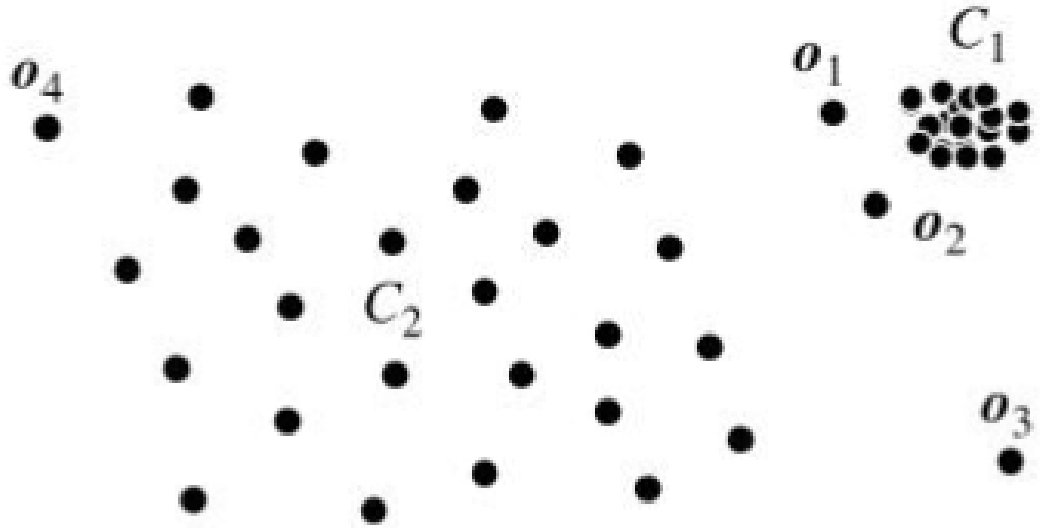
Density-based outlier detection

- Instead of looking at the entire dataset, this method focuses on smaller local areas or neighborhoods within the data.
- It compares how crowded (dense) or empty (sparse) these neighborhoods are.

Global vs. Local Outliers

- A global outlier is like someone standing alone far away from all the groups.
- A local outlier is like someone who is standing in a group but not talking to anyone, maybe because they are very different from everyone in that group

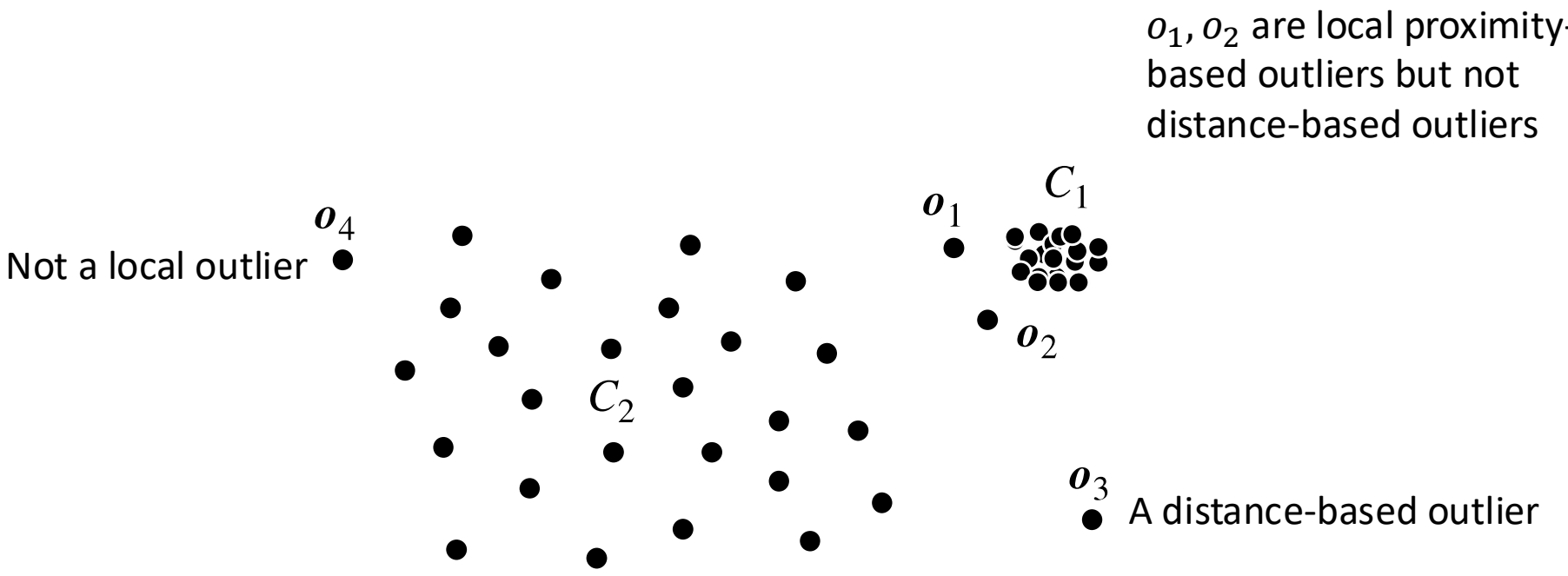
Density-Based Outlier Detection



- *Two clusters: C*1 and *C*2
  - *C*1 is crowded
  - *C*2 is sparse
- ***o*3** is an outlier because it's far from most points
- ***o*1** and ***o*2** are not outliers globally because they are close to the dense cluster ***C*1**
- However, if we just look at ***C*1** locally, ***o*1** and ***o*2** could be outliers because they're different from the other points in ***C*1**
- They're also not close to the points in ***C*2**, making them stand out when looking at each cluster separately.
- A distance-based outlier detection methods cannot capture local outliers like o1 and o2.
- Note that the distance between object o4 and its nearest neighbors is much greater than the distance between o1 and its nearest neighbors.
- However, because o4 is local to cluster C2 (which is sparse), o4 is not considered a local outlier.

Density-Based Outlier Detection

- Challenge: how to measure the relative density of an object?



k-Distance of a point

- The k-distance of o, denoted by $dist_k(o)$, is the distance, $dist(o,p)$, between o and another object, p ∈ D, such that
  - There are at least k objects $o' \in D\backslash\{o\}$ such that $dist(o,o') \leq dist(o,p)$
  - There are at most k–1 objects $o'' \in D\backslash\{o\}$ such that $dist(o,o'') < dist(o,p)$

- The k-distance $dist_k(o)$ is the distance between o and its k-nearest neighbor

k-Distance neighborhood

- The k-distance neighborhood of o contains all objects of which the distance to o is not greater than $dist_k(o)$, denoted by

$$N_k(o) = \{o' | o' \in D, dist(o, o') \leq dist_k(o)\}$$

- $N_k(o)$ may contain more than k objects because multiple objects may each be the same distance away from o

Reachability Distance

- **Reachability Distance**:
  - It's a way to measure how far away two objects are, but it has a lower limit.
  - This lower limit is the "k-distance" of the object being reached to.
  - If the actual distance is greater than the k-distance, then the actual distance is used.
- For two objects, o and o′, the reachability distance from o′ to o is $dist(o \leftarrow o')$ if $dist\ (o, o') > dist_k(o)$; $dist_k(o)$ otherwise
  - that is,

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

- Reachability distance is not symmetric, that is, in general, $reachdist_k(o \leftarrow o') \neq reachdist_k(o' \leftarrow o)$

Local Reachability Density

- **Local Reachability Density (LRD):** quantifies how close an object is to its nearest neighbors.
- This is calculated by taking the inverse of the average reachability distance of an object from its neighbors.
- It measures how close an object's nearest neighbors are. A lower LRD indicates that an object is further away from its neighbors, which could make it an outlier.
- The local reachability density of an object, o, is

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

- $|N_k(o)|$ is the set of k -nearest neighbors of o .

Local Outlier Factor

- The Local Outlier Factor (LOF) measures how isolated an object o is relative to its neighbors.
- It does this by comparing the local reachability density (LRD) of o to the LRDs of its neighbors.
- If an object o has a much lower density (lower LRD) compared to its neighbors, it is considered an outlier.
- The LOF value tells how "outlier-like" the object is:
  - $LOF_k(o)>1$: Object o is less dense than its neighbors → potential outlier.
  - $LOF_k(o)\approx1$: Object o has similar density as its neighbors → normal.
  - $LOF_k(o)<1$: Object o is denser than its neighbors → strongly non-outlier.
- The LOF is the average of the ratios of the LRD of an object's neighbors to the object's own LRD.

Local Outlier Factor

The local outlier factor of an object o is

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\left|N_{k(o)}\right|}$$