Local Outlier Factor

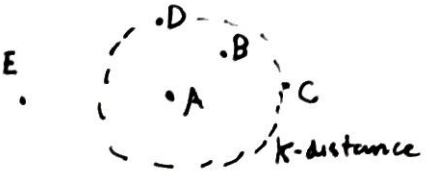
- Created in 2000 for identifying anomalous points based on local neighbors
- Think of probability density as $\frac{\text{mass}}{\text{volume}}$; (Elementary school science)
- Goal: Assign a LOF score to each point; LOF score of 1 is normal, higher LOF scores indicate possible anomaly.

Heuristic: Data points that are more isolated will have higher LOF score.

K-Distance: For a point A, the K-distance is the distance to the K^{th}-nearest neighbor. Also, denote K-neighborhood, $N_K(A)$, as all points within this distance of A.

Reachability Distance: $\text{Reach}_K(A,B) = \max\{K\text{-distance}(B), d(A,B)\}$

- Points in B's neighborhood are equally distant. Not a real distance function
- If I see you as a neighbor, do you see me as a neighbor?
- Heuristic: Neighbors of an anomaly do not view the anomaly as a nearest neighbor.



Local Reachability Density: Inverse of average reachability distance from A to neighbors

$$LRD(A) := \frac{|N_K(A)|}{\sum_{B \in N_K} \text{Reach-Dist}_K(A,B)} \quad \frac{(\text{Mass})}{(\text{Volume})}$$

[Average distance that A can be reached by its neighbor]

Local Outlier Factor: Average reachability density of neighbors compared to your reachability density.

$$LOF_K(A) := \frac{\sum_{B \in N_K(A)} LRD(B)}{|N_K(A)|} \cdot \frac{1}{LRD(A)}$$

[Anomalous points have low LRD, so $\frac{LRD(B)}{LRD(A)}$ will be high]

- If a point is normal, the average LRD of its neighbors should equal its LRD, so LOF for normal points is approximately one.

Interpretation: Ratio of how much denser a region your neighbors are in compared to you.

Disadvantages: Quotient is hard to intrepret; What threshold for LOF to use? Also bad for high-dimensional data (curse of dimensionality).