



It is a question from outlier analysis. We need to compare point B and E and predict that which point is more likely to be an outlier. As per the question, we will do this with two different approaches. The first is distance-based outlier scores and the second Local Outlier Factor (LOF).

So the funda behind distance based outlier is that we look at the k-th neighbor (3rd neighbor) in this case to call someone an outlier. We could have easily looked at the first nearest neighbor. But think of this like, two people live next to each other in a desert. Now if you ask anyone of them ‘do you have a neighbor’? they would say yes. And if you just decide based on it that they are social people, then it will be a wrong interpretation. Its same as, ‘references from the best friend are not valid’). So, instead of looking at immediate neighbor, we look at the third nearest neighbor.

For B, there are three points, (C, A and E) that are at unit distance. So. The third nearest neighbor of B is 1 unit far from them.

For E, there are four points (D, F, H, and B) that are at the unit distance. So, the third nearest neighbor of E is 1 unit foar from them.

Hence, both E and B know at three guys within 1 unit distance. So as per the distance-based outlier score, both points are similar in terms of being outliers.

Now, the second method we have to use is, Local Outlier Factor.

Now, LOF of a point X is defined as the ratio of its average reachability with mean of average reachability of its neighbours.

So, for point B, LOF is,

AR3(B)/mean (AR3(A) + AR3(C) + AR3(E)) as A, C, and E are neighbors of B.

So, we need to calculate average reachability of 4 points.

Now, AR3(X) = Mean R3(X, Y) for Y in V3 of X.

So, AR3(B) = mean(R3(B, A), R3(B, C), R3(B, E))

Now, R3(X, Y) = max(Distt(X, Y), V3(X))

So, R3(B, A) = max(Distt(B, A), V3(A)) = sqrt(2)

So, R3(B, C) = max(Distt(B, C), V3(C)) = sqrt(2)

So, R3(B, E) = max(Distt(B, E), V3(E)) = 1

-----------------------------------------------------------

So, AR3(B) = (sqrt(2) + sqrt(2) + 1)/3

Now, similarly, we need to calculate AR3(C), AR3(E), and AR3(A)

AR3(C) = mean(R3(C, B), R3(C, F), R3(C, E))

So, R3(C, B) = max(Distt(C, B), V3(B)) = 1

So, R3(C, F) = max(Distt(C, F), V3(F)) = 1

So, R3(C, E) = max(Distt(C, E), V3(E)) = sqrt(2)

So, AR3(C) = (sqrt(2) + 1+ 1)/3

Now, A is symmetrical to C, so AR3(A) = AR3(C) = (sqrt(2) + 1+ 1)/3

Now, we have 3 out of 4 Average leachabilities with us. We still need to calculate, AR3(E)

AR3(E) = mean(R3(E, B), R3(E, D), R3(E, F), R3(E, H))

Now,

So, R3(E, B) = max(Distt(E, B), V3(B)) = 1

So, R3(E, D) = max(Distt(E, D), V3(D)) = 1

So, R3(E, F) = max(Distt(E, F), V3(F)) = 1

So, R3(E, D) = max(Distt(E, D), V3(D)) = 1

AR3(E) = (1 + 1 + 1 + 1)/4 = 1

Now, we have all the required average reachability

Now, we will calculate LOF of B using the formula we wrote earlier

AR3(B)/mean (AR3(A) + AR3(C) + AR3(E))

AR3(B) = (sqrt(2) + sqrt(2) + 1)/3 = 1.276

AR3(A) = (sqrt(2) + 1+ 1)/3 = 1.138

AR3(C) = (sqrt(2) + 1+ 1)/3 = 1.138

AR3(E)) = 1

mean (AR3(A) + AR3(C) + AR3(E) = (1.138 + 1.138 + 1)/3 = 1.092

so, LOF = 1.276/1.092 = 1.168

Now, we will calculate LOF of E.

LOF(E) = AR3(E)/mean (AR3(B) + AR3(D) + AR3(H) + AR3(F)) as B,D,H, and F are neighbors of E

AR3(E)) = 1 (we already calculated that)

AR3(B) = 1.276(we already calculated that)

Now, D, H, and F are symmetrical to B. so that will have same AR3

So, AR3(B) = AR3(D) = AR3(H) = AR3(F) = 1.276

Also, mean (AR3(B) + AR3(D) + AR3(H) + AR3(F)) = 1.276 (as they are all same)

So,

LOF = 1/1.276 = 0.78

So, LOF of B is more than LOF of E. So, B is more likely to be an outlier.