Formula for local outlier factor, to identify that given point is outlier is not is given in below image.

If LOF(xi) is large then it’s outlier, otherwise not.

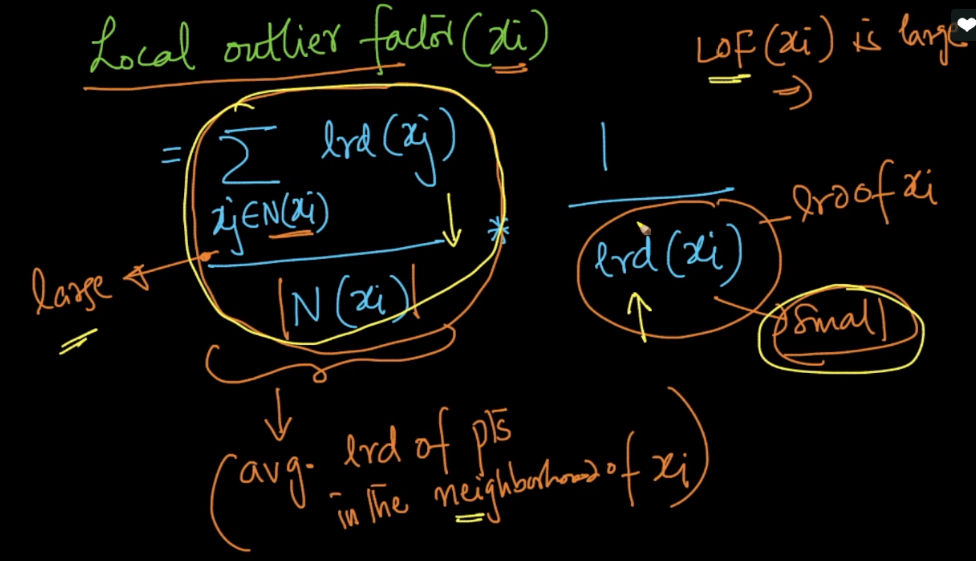
Let’s break this formula into two parts, one on the left side of multiplication and other on the right side of multiplication.

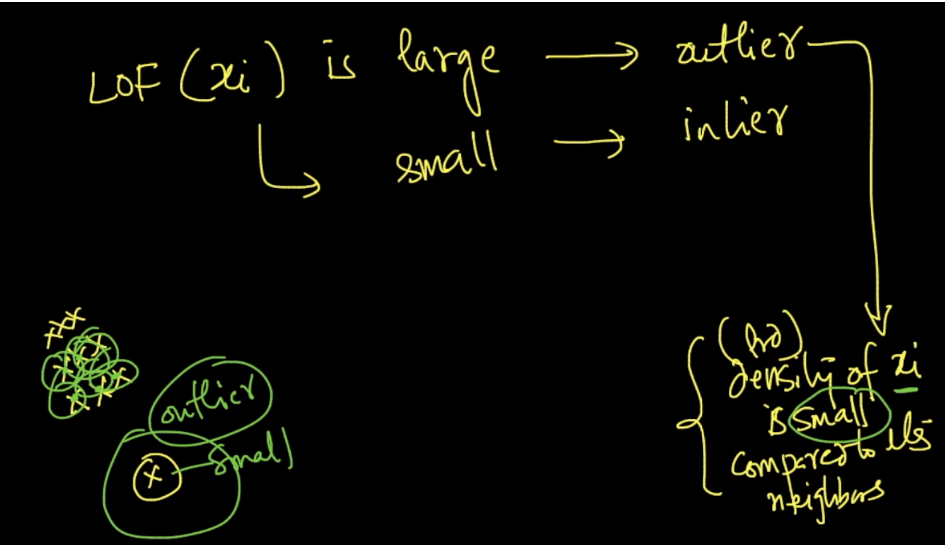
**Left Part**: It’s average of LRD of all points in the neighbourhood of xi, which shows the density of all the points present in neighbourhood of xi. If it’s large then LOF will also be large.

**Right Part**: It’s the inverse of LRD of xi (which shows density of xi), which tells the more the density of xi,  the less will be it’s value. If density of xi is small the LOF will be large.

Here is the intuition, if density of xi is significantly less as compare to density of all the points in it’s neighbourhood then LOF will be large, that means xi is an outlier.

On the other side if density is xi is similar to all other points in its neighbourhood then it’s inliner.





Let’s understand using the example given in below image.

There are two clusters c1, c2. Let’s say we take a point from c1 say x1 and from c2 say x3, and x2 and x4 are outliers.

Let’s say avg reach-dist of each point(x1, x2, x3, x4) is d1, d2, d3, d4 respectively.

We can clearly see that d1 < d2 < d3 < d4.

Now LOF detects them as follows:

Here LRD(density) of x1 is almost similar to LRD of it’s neighbourhood points, that’s why LOF will be lower.

LRD of x2 is significantly less than LRD of it’s neighbourhood points, that’s why LOF will be higher.

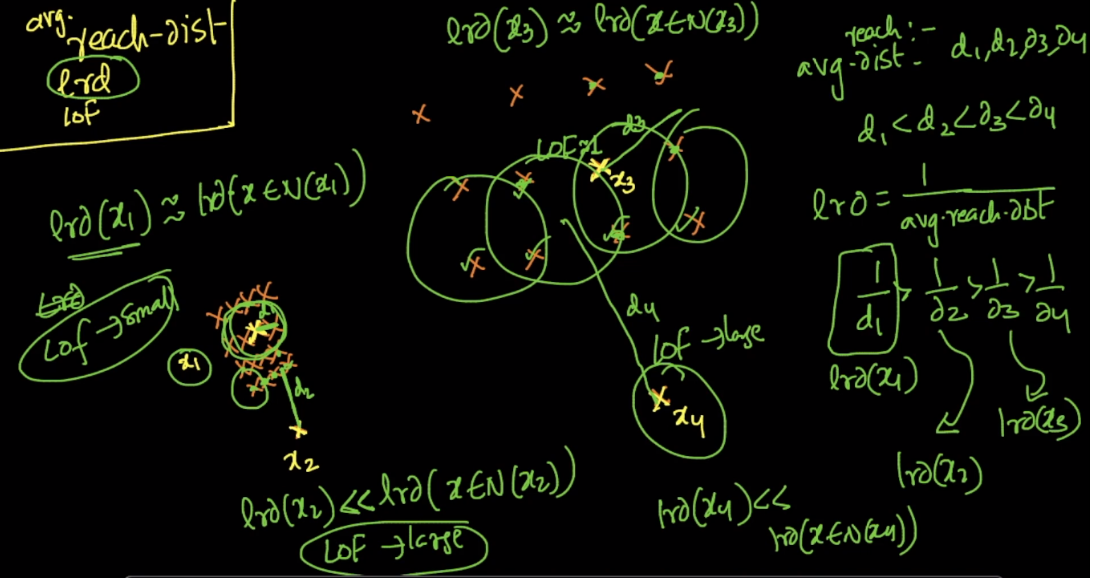
LRD of x3 is almost similar to LRD of it’s neighbourhood points, that’s why LOF will be lower.

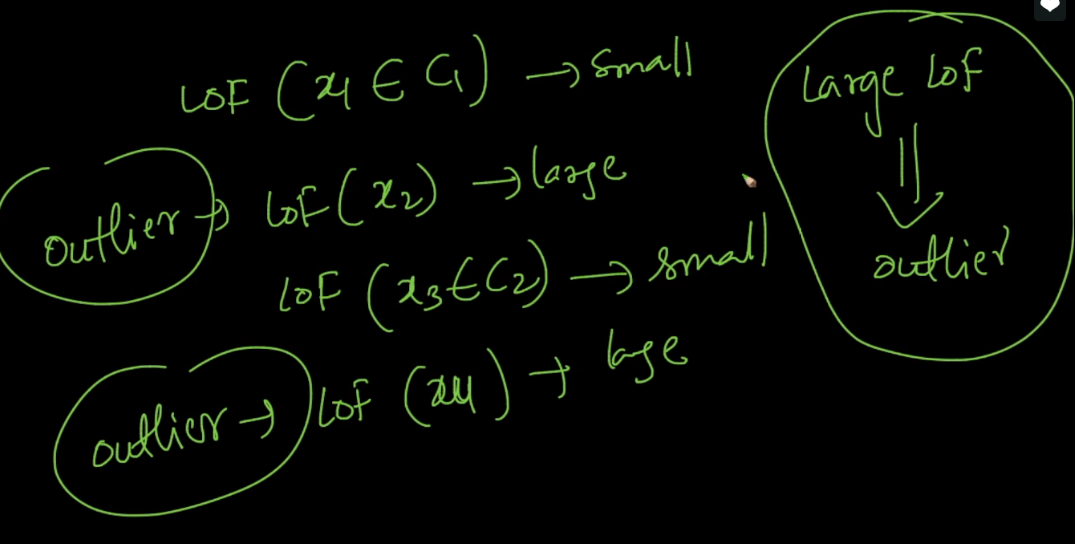
LRD of x4 is significantly less than LRD of it’s neighbourhood points, that’s why LOF will be higher.

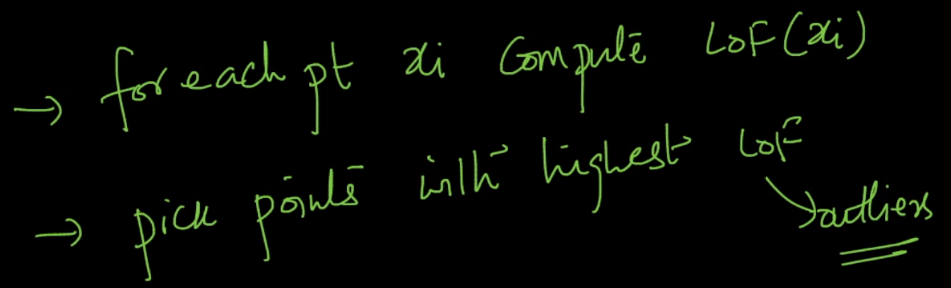
Now let’s say what if we do it only using right part of formula saying if density of point is smaller then it’s an outlier.

Since the avg reach-dist of each point are as d1 < d2 < d3 < d4.

Therefore 1/d1 > 1/d2 > 1/d3 > 1/d4, that means LRD(d1) > LRD(d2) > LRD(d3) > LRD(d4), therefore seeing this we are saying that x2 is not an outlier, as it’s density is greater than x3. And therefore we’ve to merge it with Left part also



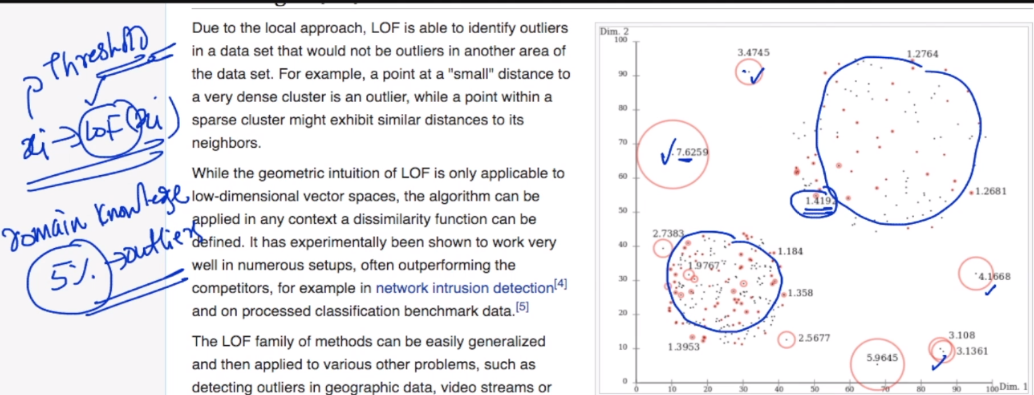


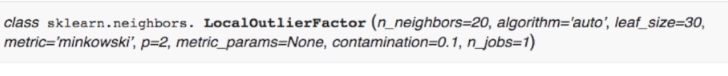


Now we got each point with LOF value, now how to know with how much LOF we’ve to remove.

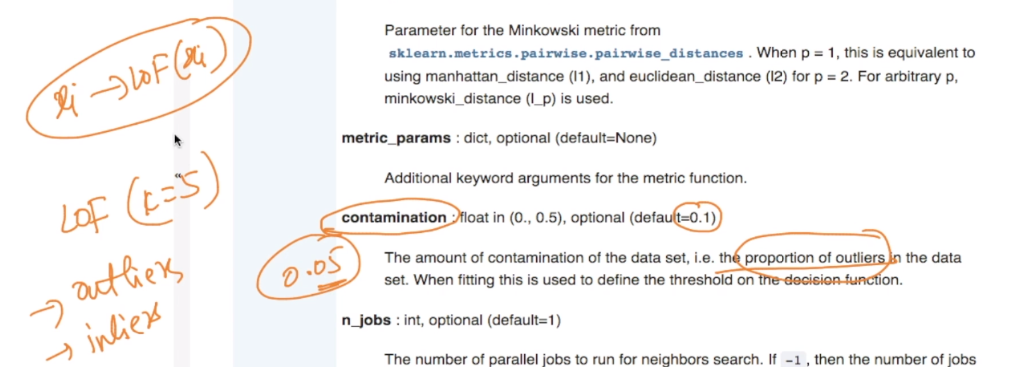
Here is a solution.

Sort points according to it’s LOF, Now with domain knowledge you get to know that in a given data some percentage of data is outlier let say 5% is outlier so we remove top 5% data have large LOF value.





Here contamination specify how many much proportion of data is outlier.



**Notes:**

* So we have to compute LOF for every point as which ever points have higher values, they are considered as outliers.  
  But again we need to check with the domain expert regarding how many points should be removed as outliers as the threshold value differs from problem to problem and domain to domain.
* how does the first term in the equation of LOF exactly contribute

The first term in LOF equation is basically calculating the average local reachability density of all the points near to that point. And it is then multiplying with the inverse of that point's own local reachability density. Hence if the densities are similar it will just give a value close to 1.

What happens if reachability of the neighbours as well as the point being considered is low. Even in this case do we get an LOF of around 1.What happens with these kind of datasets?

You will get low LOF values. But here again LOF takes the percentagee of the number of points to be treated as outliers. If we mention the value as 0.3(say), then it considers the top 30% of the points with high LOF as outliers. Even if the LOF values are low, but still aong them the top 30% will be be treated as outliers

**Outliers detection techniques learned till now:**

1. Boxplots
2. Interquartile Range (IQR) if point lies less than Q1-1.5IQR or more than Q2+1.5IQR then it is outlier
3. LOF