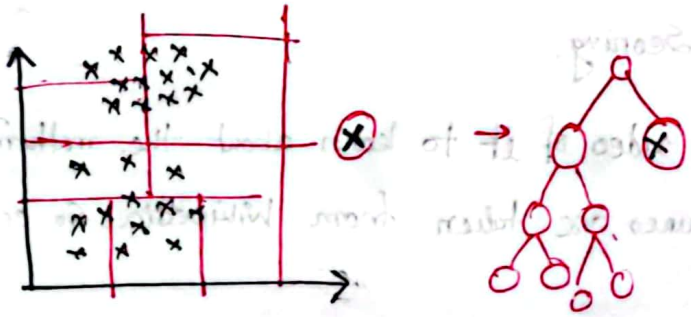
Anomaly detection Isolation Forces: (To detect outliers)

Suppose, we have a disease dataset, where the normal points shows the healthy people and the outliers shows the people who are in diseases.
So, not always we have to remove outliers.
In this type of case, to detect outliers, we use Anomaly Detection algorithm.

There are two types of Anomaly Detection:

(1) Isolation Forest (Decision Tree)

We will run decision Tree on data points and will find the leaf nodes



The datapoints who are isolated, their leafnode can be found easily.
Like in our case we can find it in depth 2.

So, the isolated node, how quickly it is found, that becomes the high probability to be a outlier. For that, we will calculate Anomaly Score.

Because with many features, many decision tree will be created and many isolated node will be found. To measure the probability of a leaf node to become an outlier, we will use Anomaly Score.

Mathematical score of anomaly score:

Anomaly score for a new score:

$$S(x,m) = 2^{\frac{-E(h(x))}{c(m)}}$$

m = no. of data points

x = data point

$E(h(x))$ = Average search depth of x from the isolated tree

$c(m)$ = Average depth of all the datapoints.

$E(h(x)) \ll c(m) = S(x,m) \approx 1 \Rightarrow$ Anomaly score → outlier

We will define the threshold, suppose 0.5

>0.5 will be outliers.

$E(h(x)) \gg c(m) = S(x,m) \approx 0.5 \Rightarrow$ Normal data

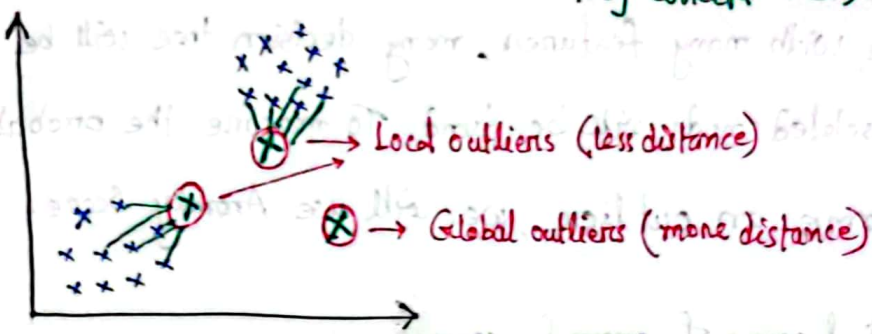
DBSCAN Clustering For Anomaly Detection:

(Already the basics of it is discussed)

It can also be used to detect outliers.

Local outlier factor anomaly detection: (LOF)

Key concept: Measuring local density



It follow the technique of K Nearest Neighbour. It calculates the distance of point x with its K Nearest Neighbour (Suppose K=5).

So it calculate the distance of x with it's 5 nearest neighbour and then calculate each neighbours distance with it's other 5 nearest neighbours Then it compares the density of it self and the density of the neighbours with their neighbours. If the density of x is less than it's neighbours then it is an outlier.