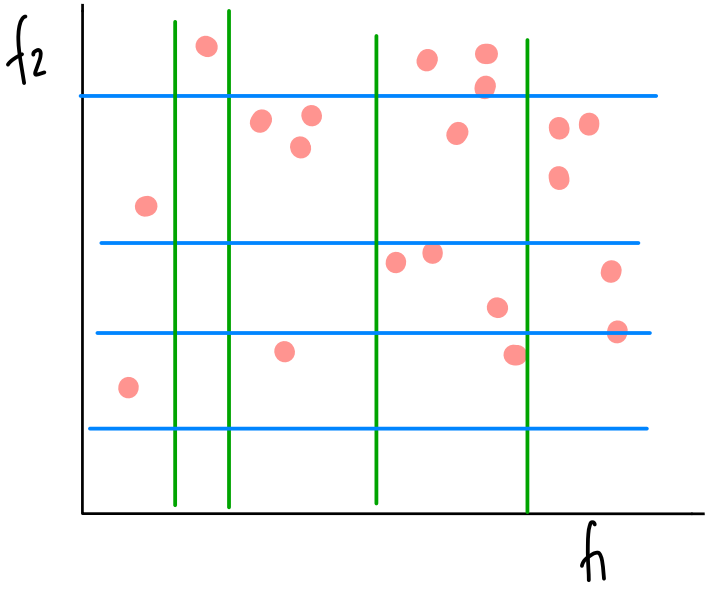
1. Isolation Forest => Based on DT
2. LOF : Local outlier Factor

Isolation Forest

↳ Collection of trees

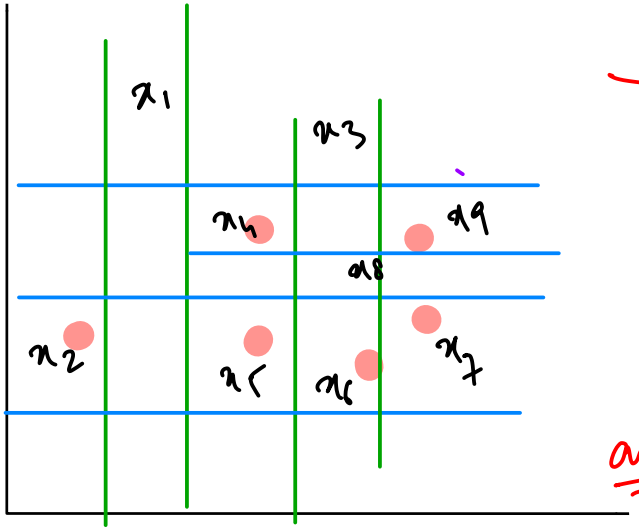
↳ Isolate points

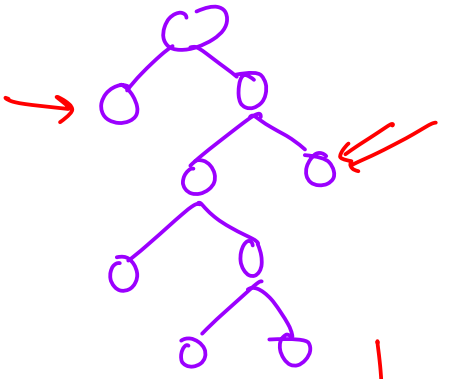
Create multiple decision trees on f's with thresholds and isolate point



For 2:

1. Randomly pick a feature
2. Randomly threshold the feature
3. Build the decision tree until each data point is a node itself.





100's

DT

1.6

avg(d) x_1 [2, 2, 1, 3, 2, 2, 3 . . . .]

1.8 x_2 [2, 1, 1, 1, 2, 3, 4, 2, 1]

2.9 x_3 [3 4, 3, 4, 5, 6, 3]

5 x_4

2.9 x_5

3.2 x_6

x_7

x_8

x_9

We use avg depth as a metric.

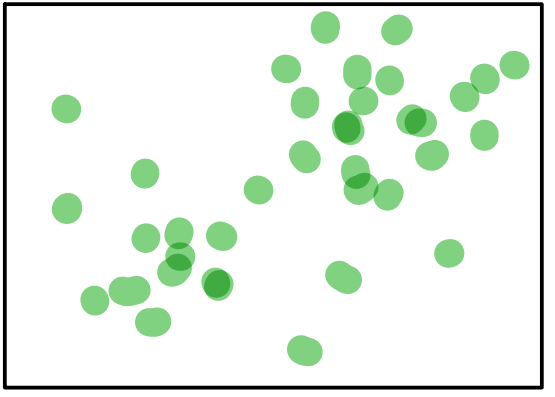
Soft point based on avg (d)

2%

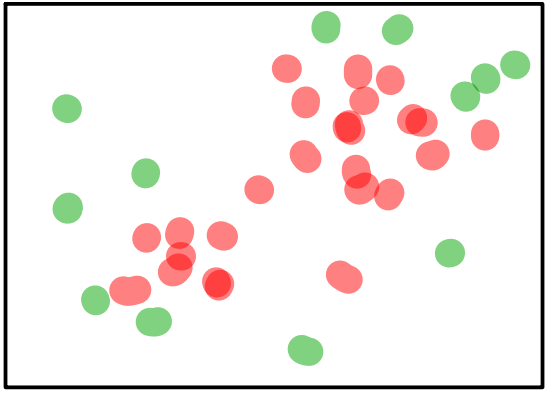
$x_1 < x_2 < x_3 \ldots$

100

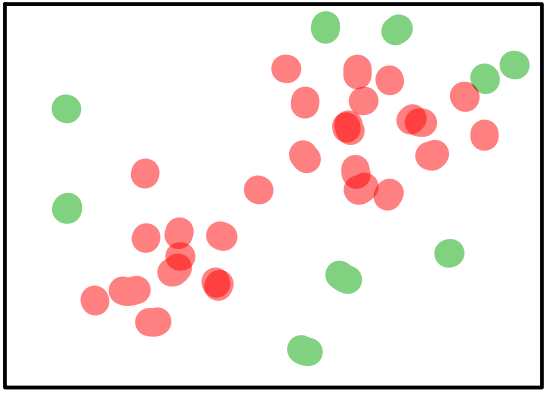
Contamination: 2%



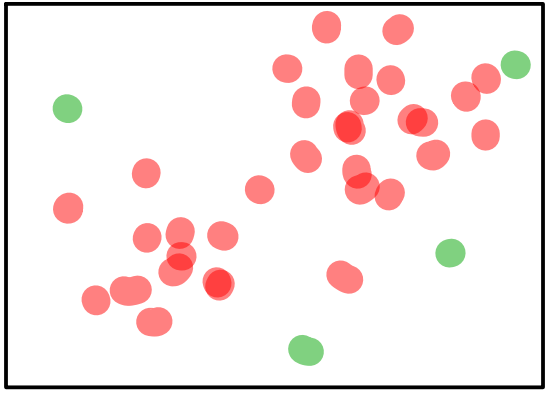
Contamination: 20



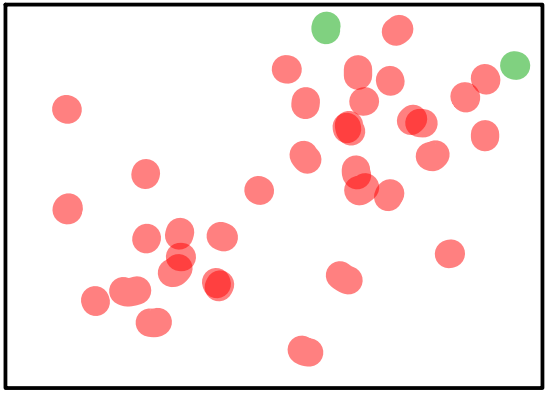
Contamination = 10



Contamination: 5

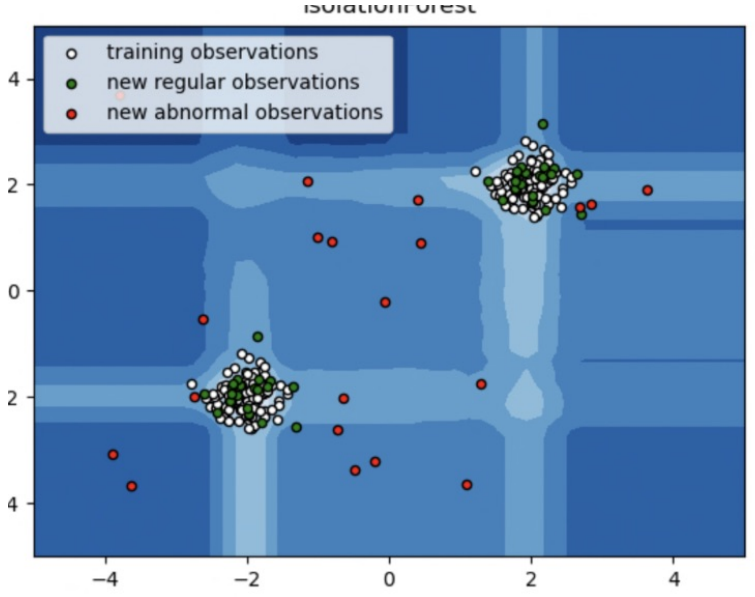


Contamination = 1



Disadvantages

* Slow if the size of dataset ↑↑↑ huge.
* Isolation Forests are biased towards axis parallel split.
* In the diagram below, shades of Color represent likelihood of a data point being inlier/outlier.



Advantage.

* Effective.
* Find out outliers b/w 2 clusters.
* Use probability / shade of color in determining effectively how sure are we in saying that a datapoint is outlier / inlier.

LOF . Local Outlier Factor

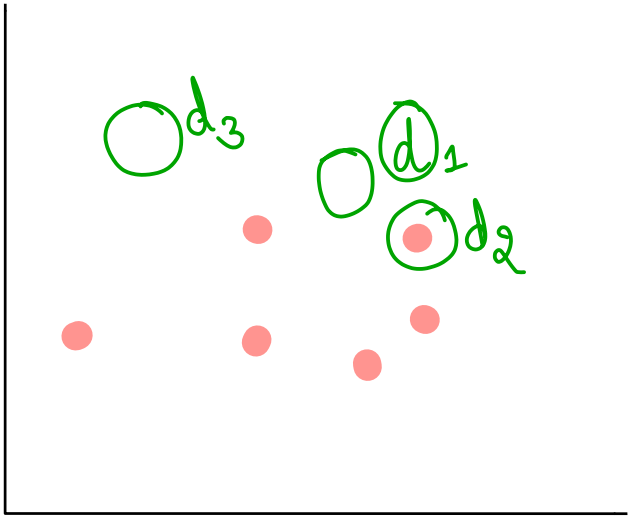
↳ KNN + Density

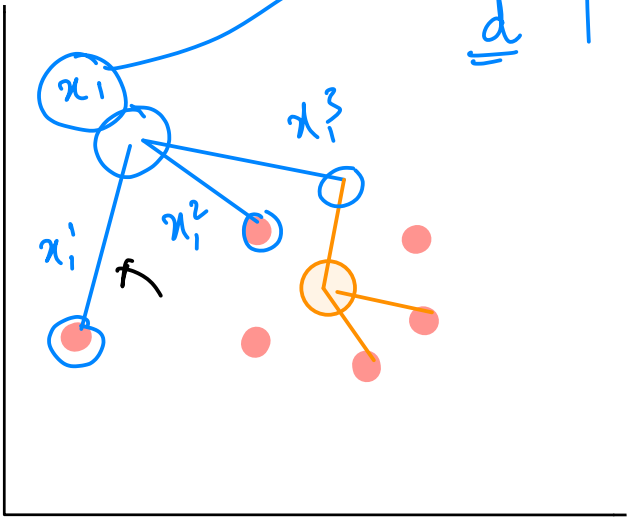
K_d

K distance

$x_1^2 < x_1^3 < x_1^1$

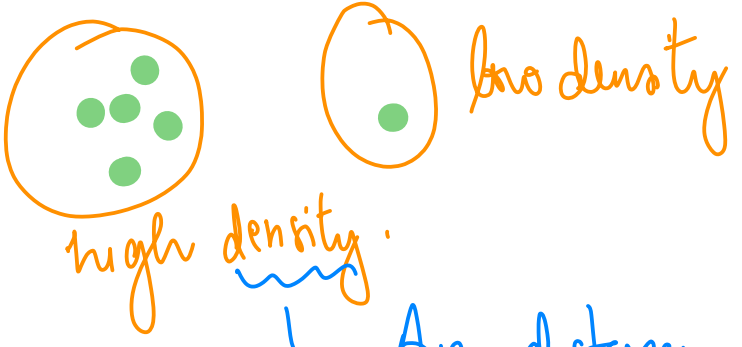
d





K_d ⇒ We will find the distance of farthest point among K Nearest Neighbour.

Idea: Outliers exist in low density region



↳ Avg distance b/w point.

$$\text{Density} \propto \frac{1}{\text{avg}(d)} \Rightarrow \text{Density} = \frac{1}{\text{avg}(d)}$$

(1) Find k nearest neighbours of point (A) $k = 3$.

(2) Calculate avg distance. μ.

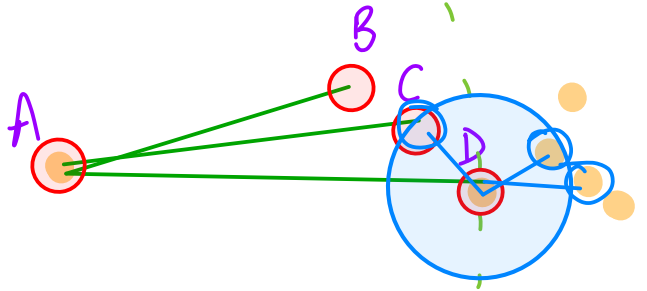
$$\text{Density}(A) = \frac{1}{d_1 + d_2 + d_3}$$

Variation → Instead of averages, we can also take maximum out of d_1, d_2 & d_3 → K dist [max value]

$$\text{density} = \frac{1}{K_D(A)}$$

max (k nearest neighbours) = K_d: distance from sorted K^{th} neighbour.

$k = 3$



low density region { A has a large radius

{ D has lower valued radius

high density region.

LOF

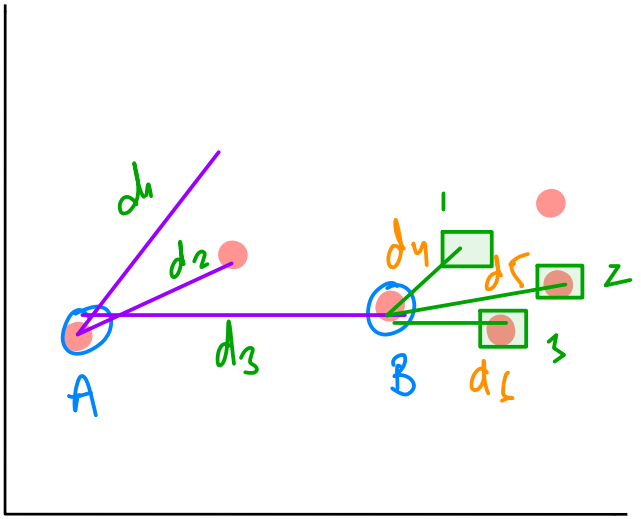
Local Outlier Factor: = $\frac{\text{Avg distance of k neighbours}}{\text{Density of A}}$

Local Reachability density

Reachability distance (A, B) = $rd(A, B)$

$$rd(A,B) = \max\{ \text{dist}(A,B), K\text{dist}(B)\}$$

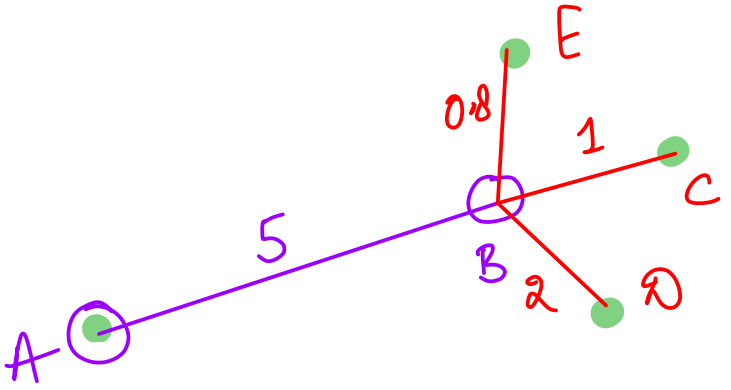
$$\text{Density}(A) = \frac{1}{\frac{1}{K}\sum RD(A, \text{neighbor})}$$



$$rd(A,B) = \max\{ d_3, d_5\}$$

Rd (me, you)

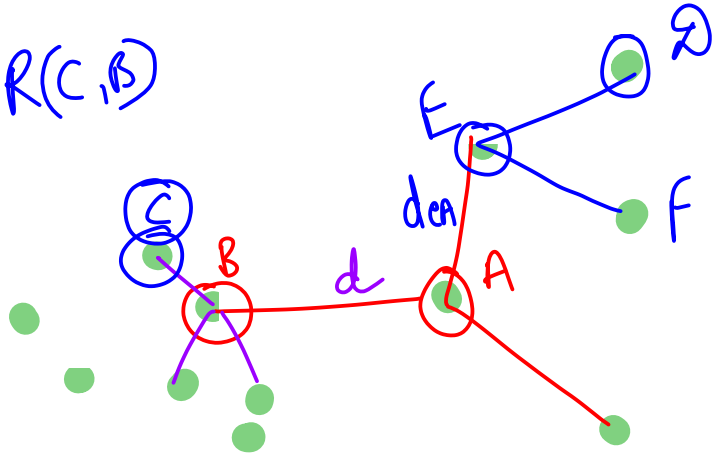
if you are in my 'top k' but I am not in you 'top k' then lets use actual distance b/w us, but if we both are in 'top k' of each other then lets use your k dist.



$R_d(A,B) = 5$ $\quad$ $Rd(E,B) = 2$

$R_d(C,B) = 2$

This is less intuitive bcos some points



$R(D,E) = d_{cA}$

$R(F,E) = d_{cA}$

① Code for all

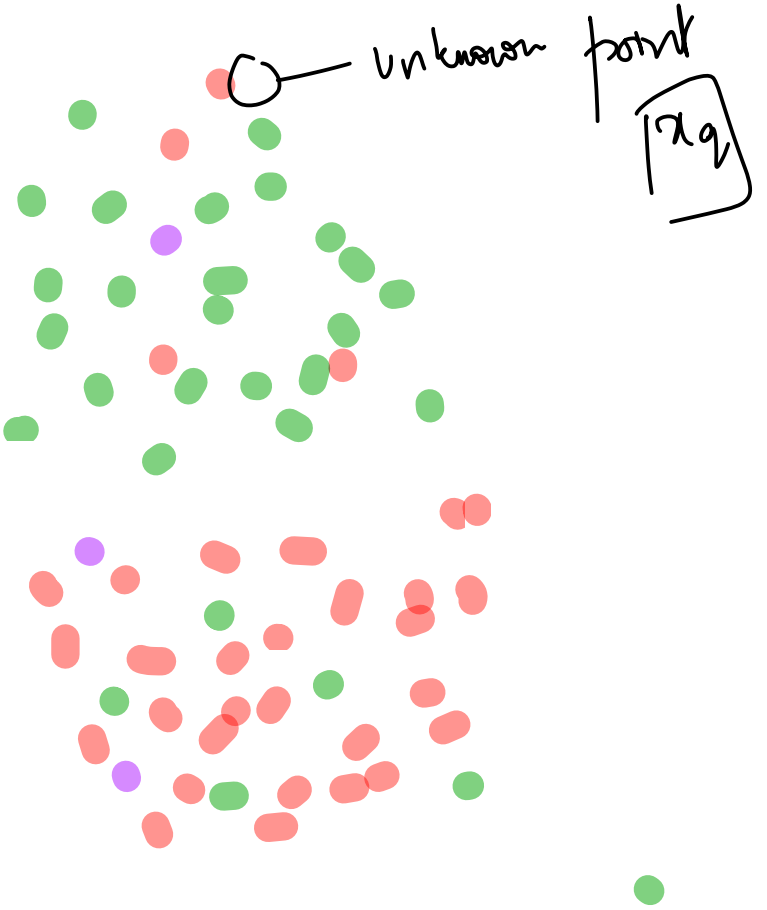
② LOF Interpretation + Adv + disadv. } next class.

Doubt

$n = 100$

$y_{+ve} = 60$

$y_{-ve} = 40$



$K = 100$ Impact of outlier ↓↓

$K = 1$ Impact of outlier ↑↑↑