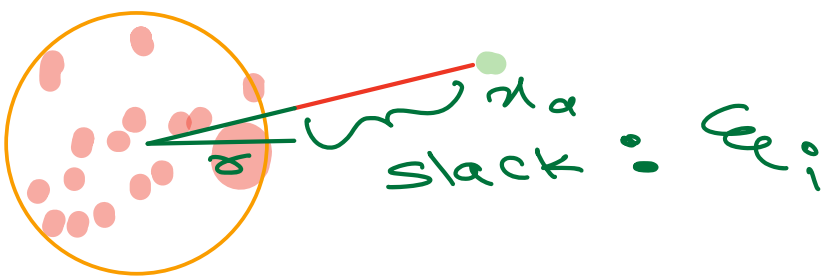
- One Class SVM
- LOF

One-Class Classifier

One-Class SVM

- Outlier Detection

- There is only one Category 'inlier'
- Everything that is not an inlier is considered Outlier



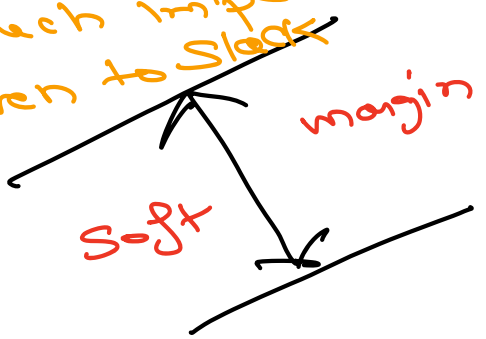
One-Class SVM

Find a radius r of hypersphere such that it captures all the inlier points

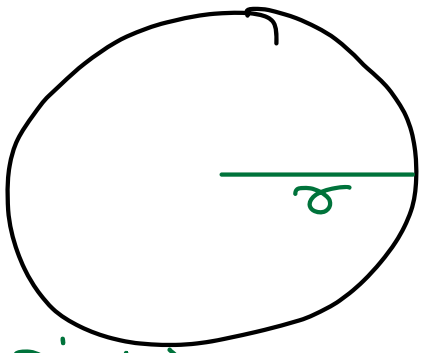
$$\min\left(\alpha^2\right) + \lambda \sum_{i=1}^{n} \epsilon_i$$

Gradient descent

How much importance is given to Slack



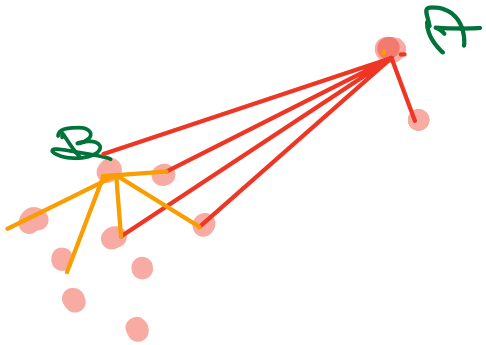
RBF work Better generally



- Disadvantages
  - Very Slow
  - As n (No of Data-points) increases, Time-Complexity increases as well
  - Many Hyperparamets to tune
    - Kernel
    - Lambda

Local Outlier Factor

inspired by KNN + Density



$k = 5$

Density for point A

$$D_A \propto \frac{1}{\sum_{i=1}^{k} (d_k)}$$

$d_1 + d_2 + d_3 + d_4 + d_5$

$$D_B \propto \frac{1}{\sum_{i=1}^{k} (d_k)}$$

$d_1 + d_2 + d_3 + d_4 + d_5$

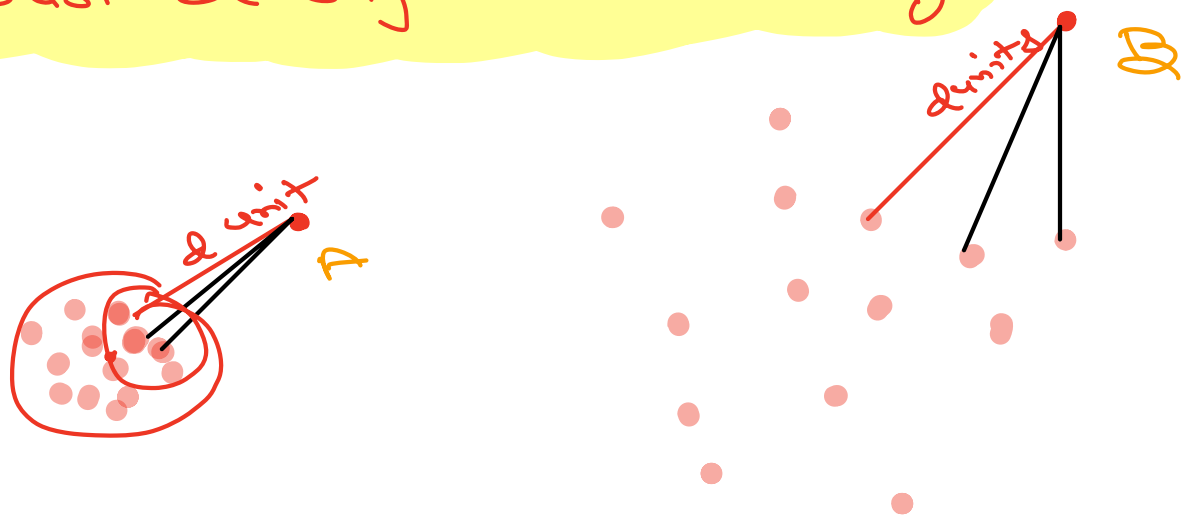
$$D \propto \frac{1}{\text{Sum of Distance of neighbors}}$$

$$\sum_{i=1}^{k} (d_k)_A \quad > \quad \sum_{i=1}^{k} (d_k)_B$$

$$D_A \quad \frac{1}{\text{Sum}_K} \quad < \quad \frac{1}{\text{Sum}(K)} \quad D_B$$

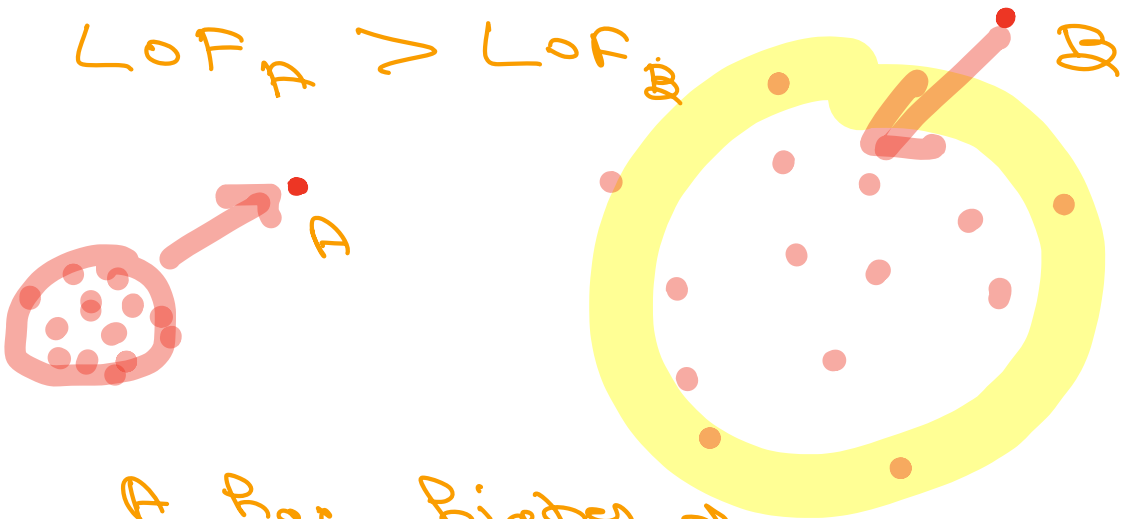
Is lower density point an Outlier?

* Just Density is Not Enough



$D_A = D_B$

$LOF_A > LOF_B$



A has higher chance of being an Outlier compared B

LOF ←

Can we compare the density of a point with Density of it's Neighbor

$$LOF_A \Rightarrow \frac{\text{Avg Density of Neighbors}}{\text{Density of Point A}}$$

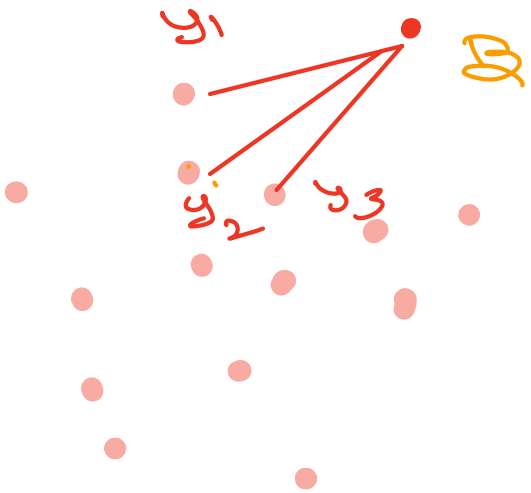
Ex:

Neighbours = 3

D_{x_1}
D_{x_2}
D_{x_3}
$\} \; avg(D_k)$

$$avg(D_k) << D_A$$

(2)



Neighbours = 3

D_{y_1}
D_{y_2}
D_{y_3}
$\} \; avg(D_k)$

$$avg(D_k^y) < D_B$$

$$LOF_k(A) = \frac{\sum_{x \in N_k(A)} D(x)}{(k) \times D_A}$$

$$LOF_k(B) = \frac{\sum_{x \in N_k(B)} D(x)}{(k) \times D_B}$$

Neighbours avg Density / points Density

if LOF == 1 :

A has same density as it's Neighbour

if LOF > 1 :

A has Lower density as it's Neighbour

if LOF >> 1 : ←

A is an outlier, since it's density is very Low Compared to neighbors

if LOF < 1 :

A has Higher density as it's Neighbour

Q: What is the ideal threshold beyond which we say the point is outlier

LOF → 10, 100, 1000

Contamination factor

Step 1: Get all LOF

Step 2: Sort them

Step 3: N%

Select Top N% LOF points

Different ways of Calculating Density

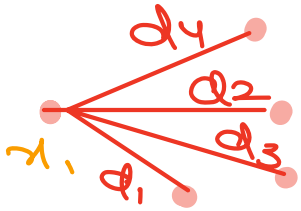
$$\frac{1}{\frac{1}{k}\sum_{B \in N_k(A)} d(A,B)}$$

$$\frac{1}{\max_{B \in N_k} [d(A,B)]}$$

① K-distance : Distance of k^{th} point

K = 3

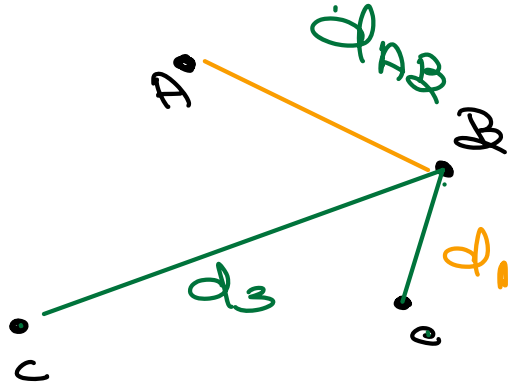
3rd → d_3



Reachability Distance

$rd_k(A, B)$

$\Rightarrow \max\{ dist(A,B), k\text{-}dist(B) \}$



If $k == 2$

$rd_2(A, B) \Rightarrow$

$\max(d_{AB}, d_{AB})$

$\Rightarrow d_{AB}$

If $k == 3$: $rd_3(A, B)$

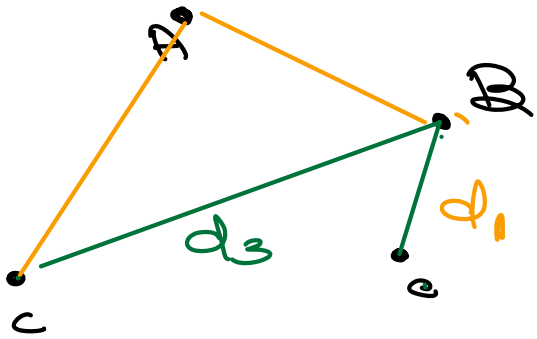
$\max(d_{AB}, d_3) \Rightarrow d_3$

Local Reachability Density

$$lrd_k(A) \Rightarrow \frac{\sum_{B \in N_k(A)} rd_k(A, B)}{k}$$

$k = 2$

$lrd_2(A)$



$rd_2(A,B) \Rightarrow \max(d_{AB}, d_2^B)$

$rd_2(A,C) \Rightarrow \max(d_{AC}, d_2^C)$

$$LRD_2(A) \Rightarrow \frac{rd_2(A,B) + rd_2(A,C)}{2}$$

$$LOF_k(A) \Rightarrow \frac{\sum_{B \in N_k(A)} lrd_k(B)}{k \times lrd_k(A)}$$

$|N_k A|$ ⇒ avg-neighbourhood

$$\frac{\text{lrd of Neighbors of A}}{\text{lrd of A}}$$

Disadvantages of LOF

1. Hyperparameter : k, threshold of LOF
2. High Time Complexity
3. Cannot Handle High Dimension Data Efficiently

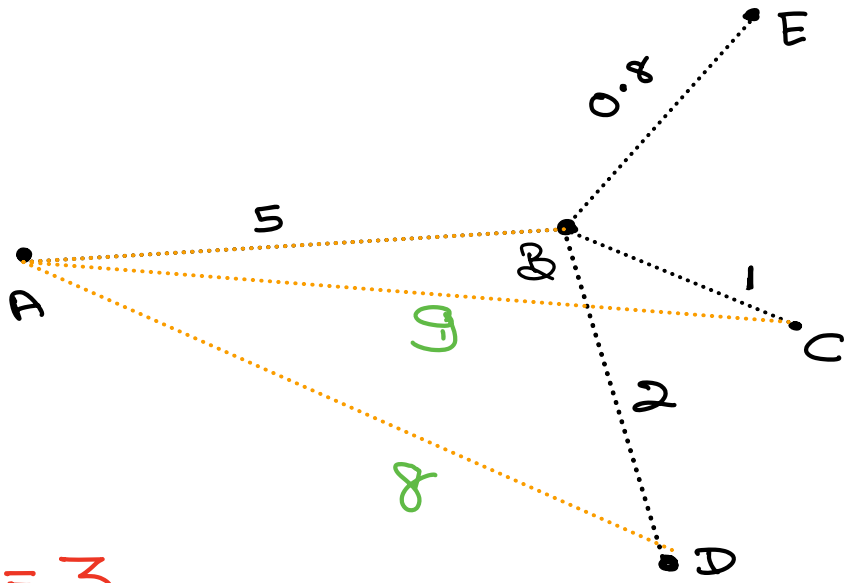
Extreme Imbalance Classification

Fraud Detect → 99.9 Val. 0.1 outliers

0.2 k

Training → Unlabelled

1. Hyperparameter → 0.05 %
2. Test model → 0.05 %



$K = 3$

Q ⇒ RD(A, B)

⇒ max (dist(A, B) , $dist_3$(B))

⇒ max (5 , 9)

⇒ 9

Q ⇒ RD (B, A)

⇒ max (dist(B, A) , $dist_3$(A))

⇒ max (5 , 9)

⇒ 9