→ Class start at 9:05 pm

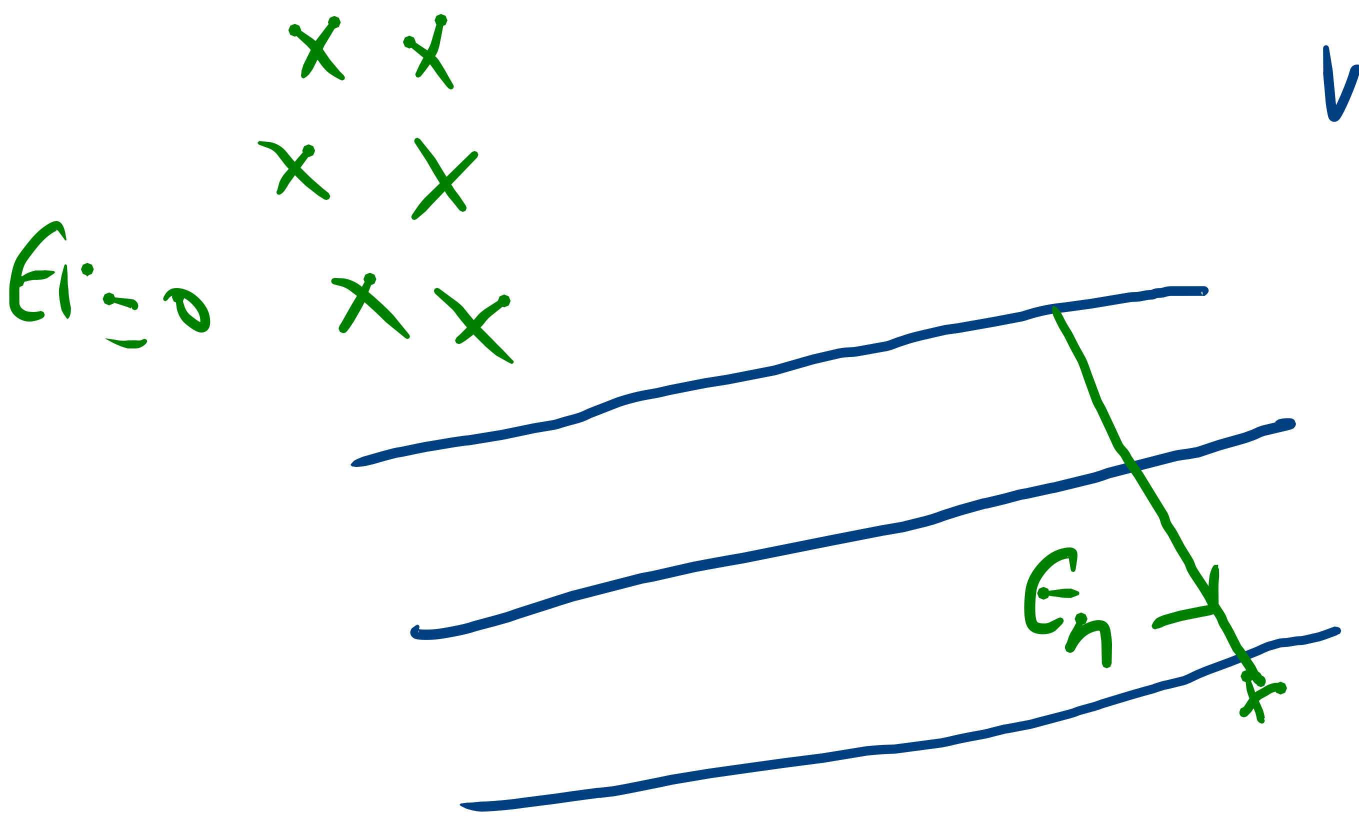
Agenda

→ One - class SVM

→ Isolation Forest

→ LOF (Local Outlier Factor)

SVM



$w^T x + b = 1$

$w^T x + b = 0$

$w^T x + b = -1$

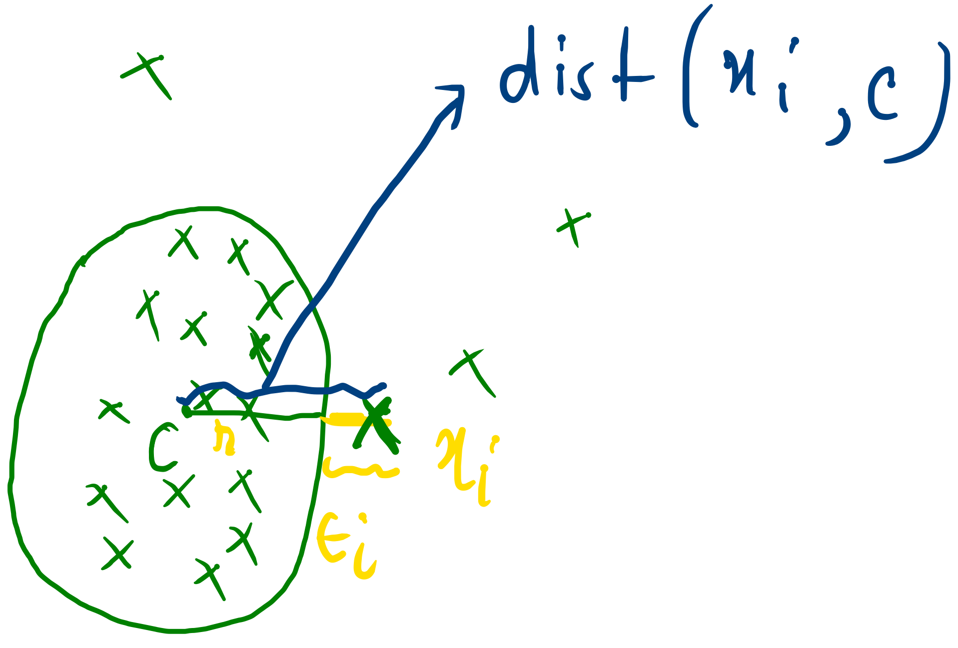
$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \epsilon_n$$

subject to $y_n(w^T x_n + b) \geq 1 - \epsilon_n$

$$\epsilon_n \geq 0$$

Oneclass SVM

Minimize the hypersphere



→ Find the centre & radius which contains most of the data point.

$$\min_{c,r} r^2$$

$$\epsilon_i = 0 \quad \text{if } dist(x_i, c) \le r$$
$$= d(x_i, c) - r \quad dist(x_i, c) > r$$

$$\begin{cases} \min\limits_{c, r, \epsilon_i} r^2 + \lambda \sum_{i=1}^{n} \epsilon_i \\ \text{s.t } dist(x_i, c) < r^2 + \epsilon_i^2 \quad \forall i \\ \epsilon_i > 0 \end{cases}$$

Disadvantage

1. As the no of data point ↑, time complexity ↑
2. Kernel selection
3. SVM complexity

Original data → 2D

RBF - Kernal

Hypersphere in in high dimension space

Isolation Forest

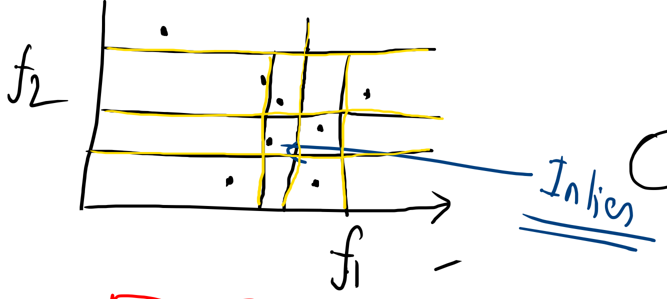
Randomly sampled datapoints

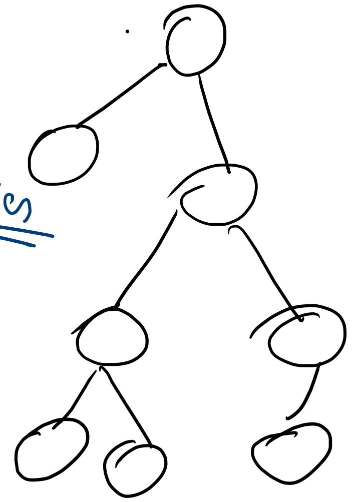
1M
↳ 1 lakh sample

→ Building many trees (like Random Forest)

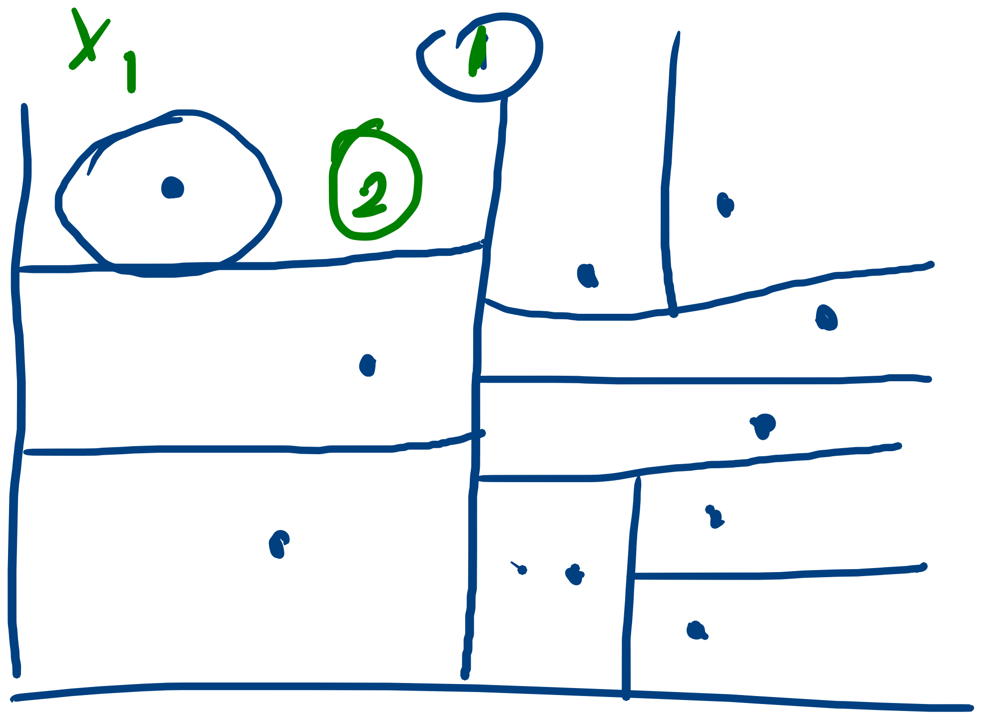
→ Each tree → randomly pick a feature
↳ randomly threshold it

→ Build each tree till the leaf node contain one datapoint

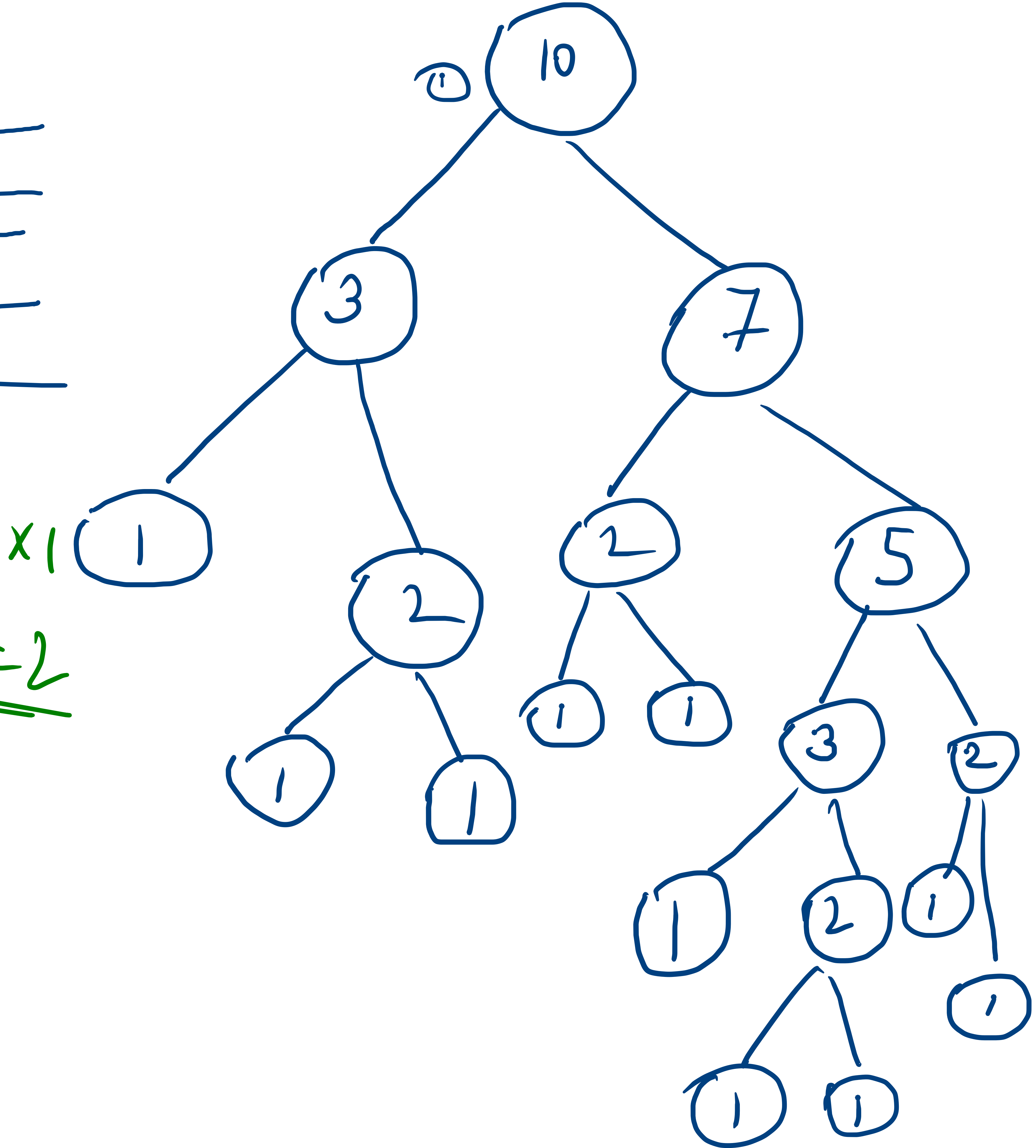


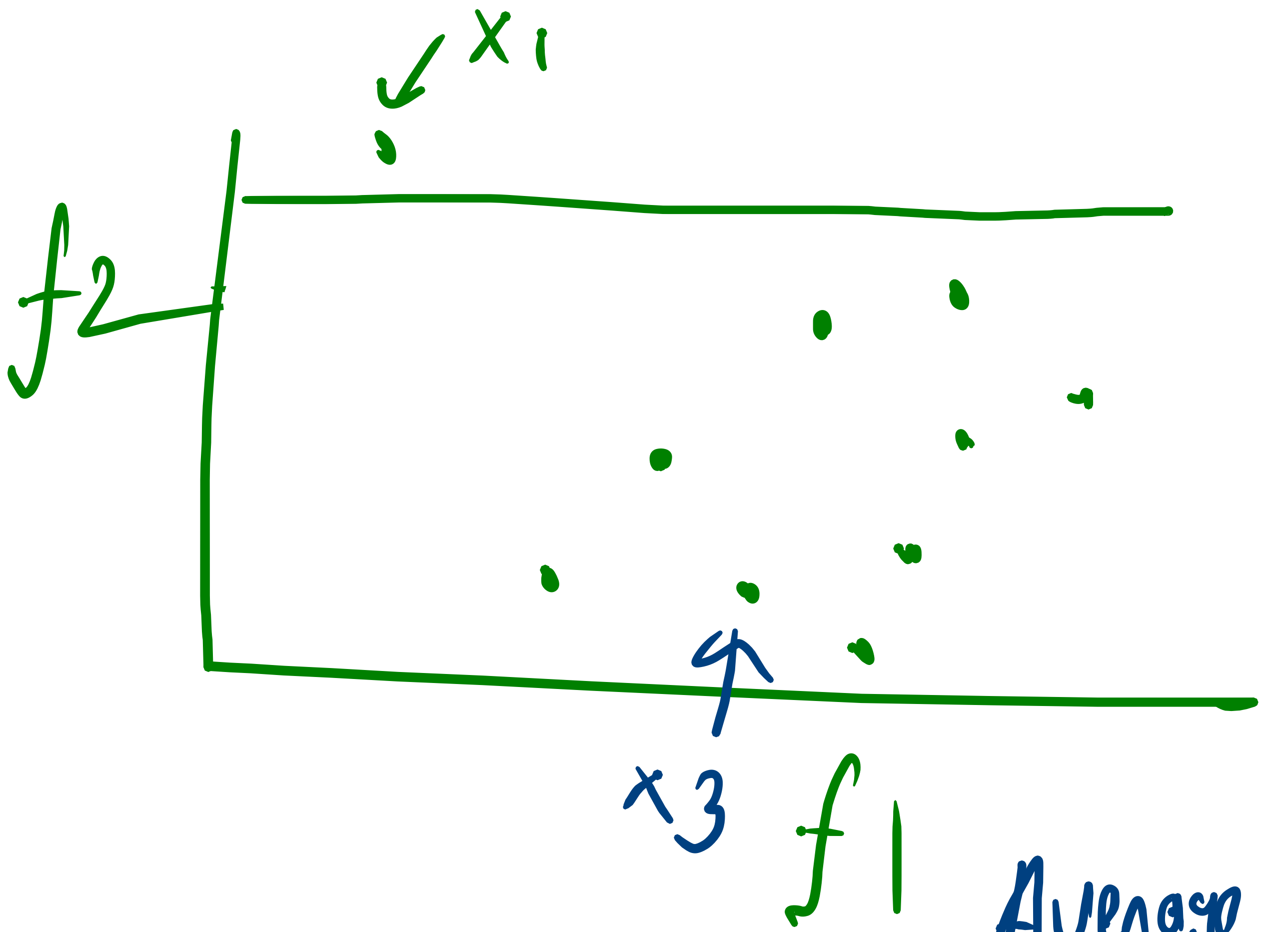


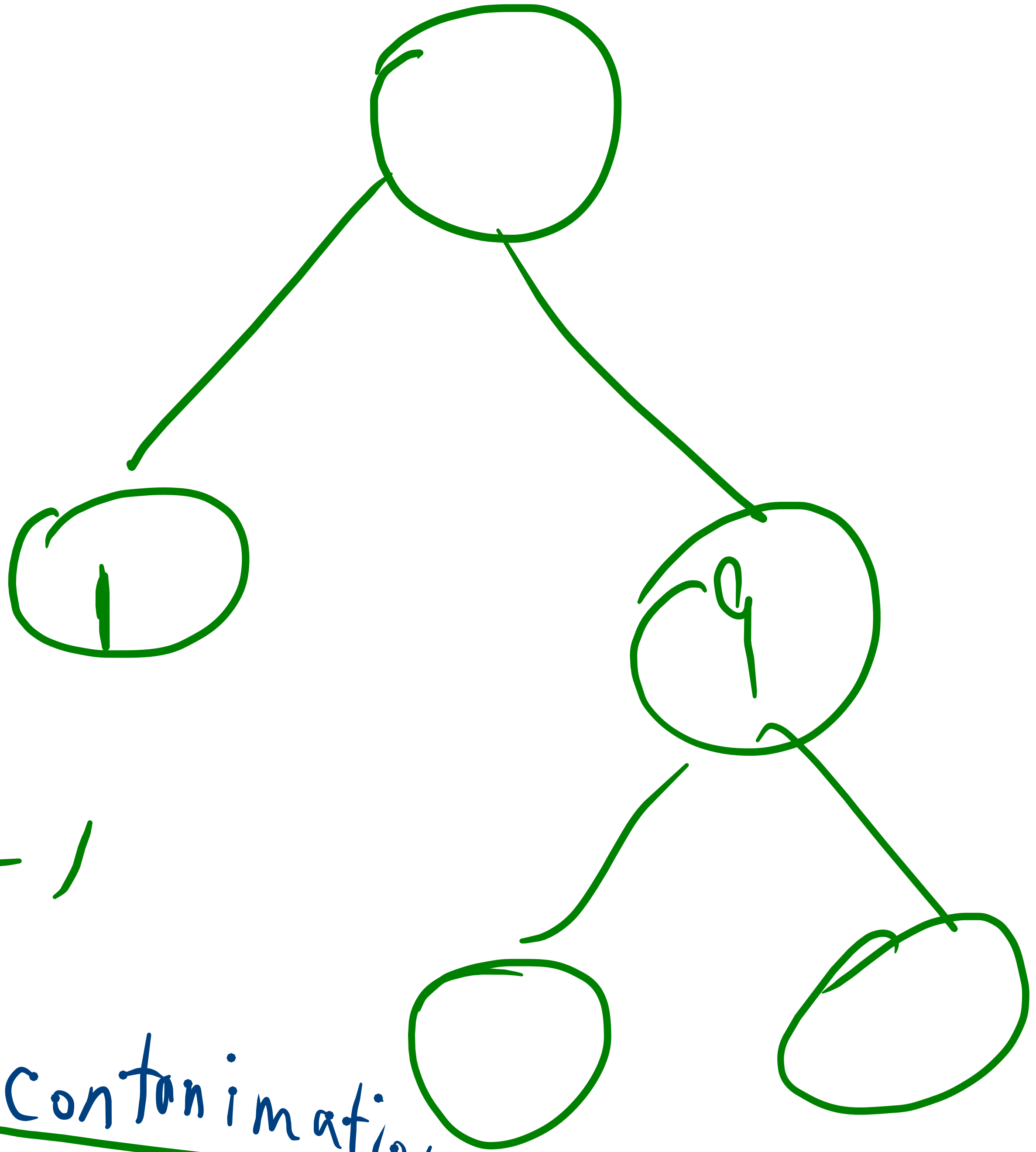
→ More depth = Inlier
Less depth = Outlier



Depth = 2







Average depth x_1

Depth = 1

x_1	2, 1	1.5
x_2		
x_3	6 7	6.5
⋮	2, 3	2.5
x_7	5, 8	6.5
x_{10}		

Contamination

10% of the data

x_1, x_2

→ LOF

→ Break until 10:10pm

DBSCAN : Noise point (Neither core point
non boundary point)

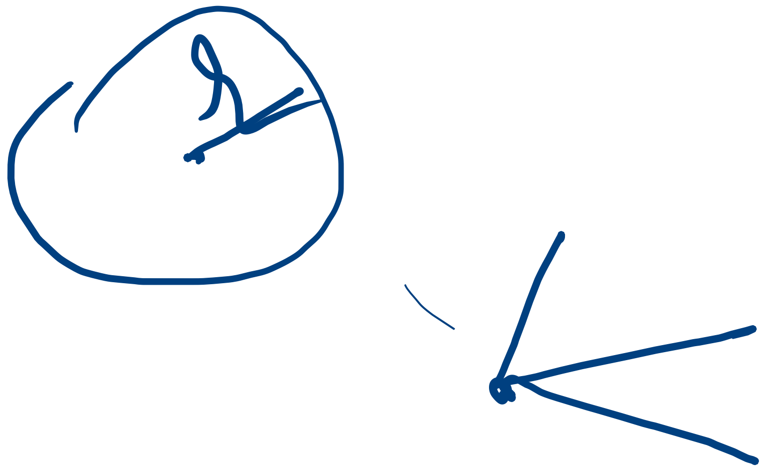
Elliptic Envelope : Probability is very low
as points are far from centre μ_x, μ_y

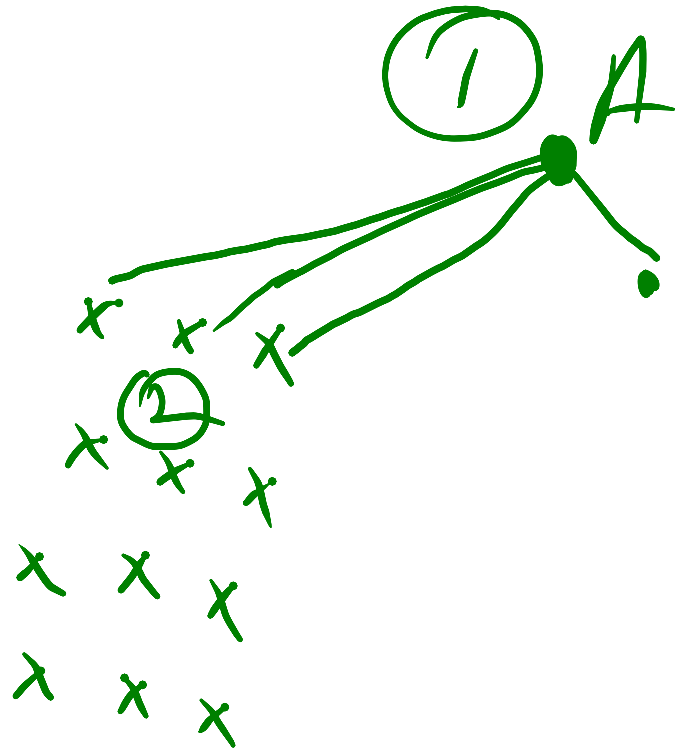
Isolation Forest : Outlier are at less tree
depth

LOF : Outlier density is lower w.r.t
density of its neighbors

LOF

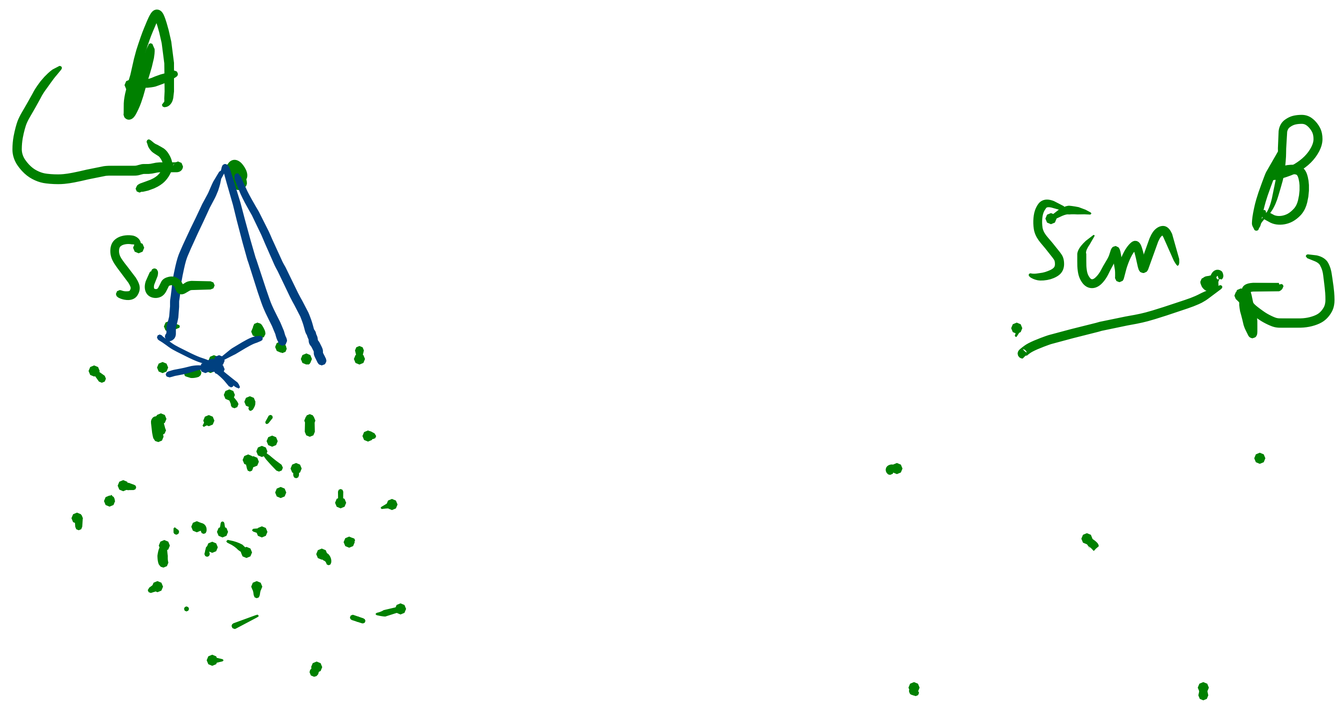
→ Concept of Density & K-NN





Density $\propto \frac{1}{\text{dist}}$

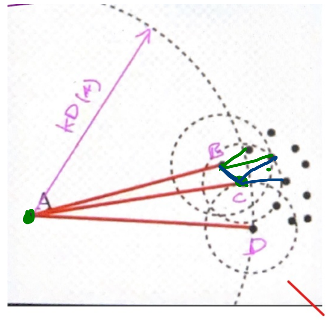
$$\text{LOF} = \frac{\text{Avg density of Neighbours of A}}{\text{Density of A}}$$



Which one is more probable to be outlier A or B ?

$$LOF = \frac{\text{density}\uparrow}{\text{density}\downarrow} = \uparrow \qquad \frac{\text{density}\downarrow}{\text{density}\downarrow} \approx 1$$

$$LOF \gg 1 \longrightarrow \underline{\underline{\text{outlier} \longrightarrow \underline{\underline{A}}}}$$



1 K-distance

K = 3

→ dist(A, D)

→ dist of point A to it K th nearest neighbor

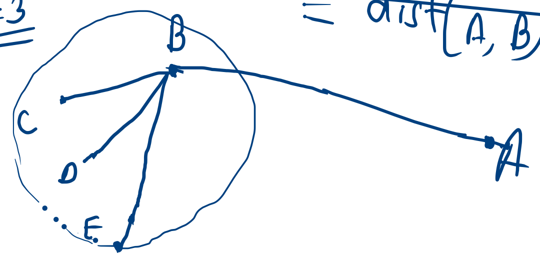
2

Reachabily dist

$$\boxed{Reach(A,B) = max(dist(A,B), K\text{-}dist(B))}$$

= dist(A, B)

K = 3



Reach (C, B) = max (dist(C,B), K-dist(B))

= K_dist(B)

Is Reach (A, B) == Reach (B, A) Symetric ??

↳ max (dist (A,B) / K_dist (B))

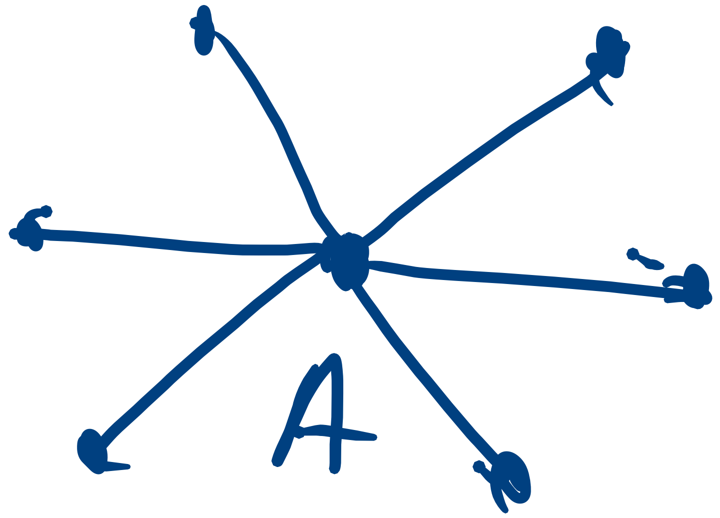
max (dist (B, A), K_dist (A))

||

dist(A B)

Local Reachabity density

$$\mathrm{lrd}(A) = \frac{1}{\left(\dfrac{\sum_{B \in N_k} \mathrm{Reach\ dist}(A,B)}{N_k} \right)}$$



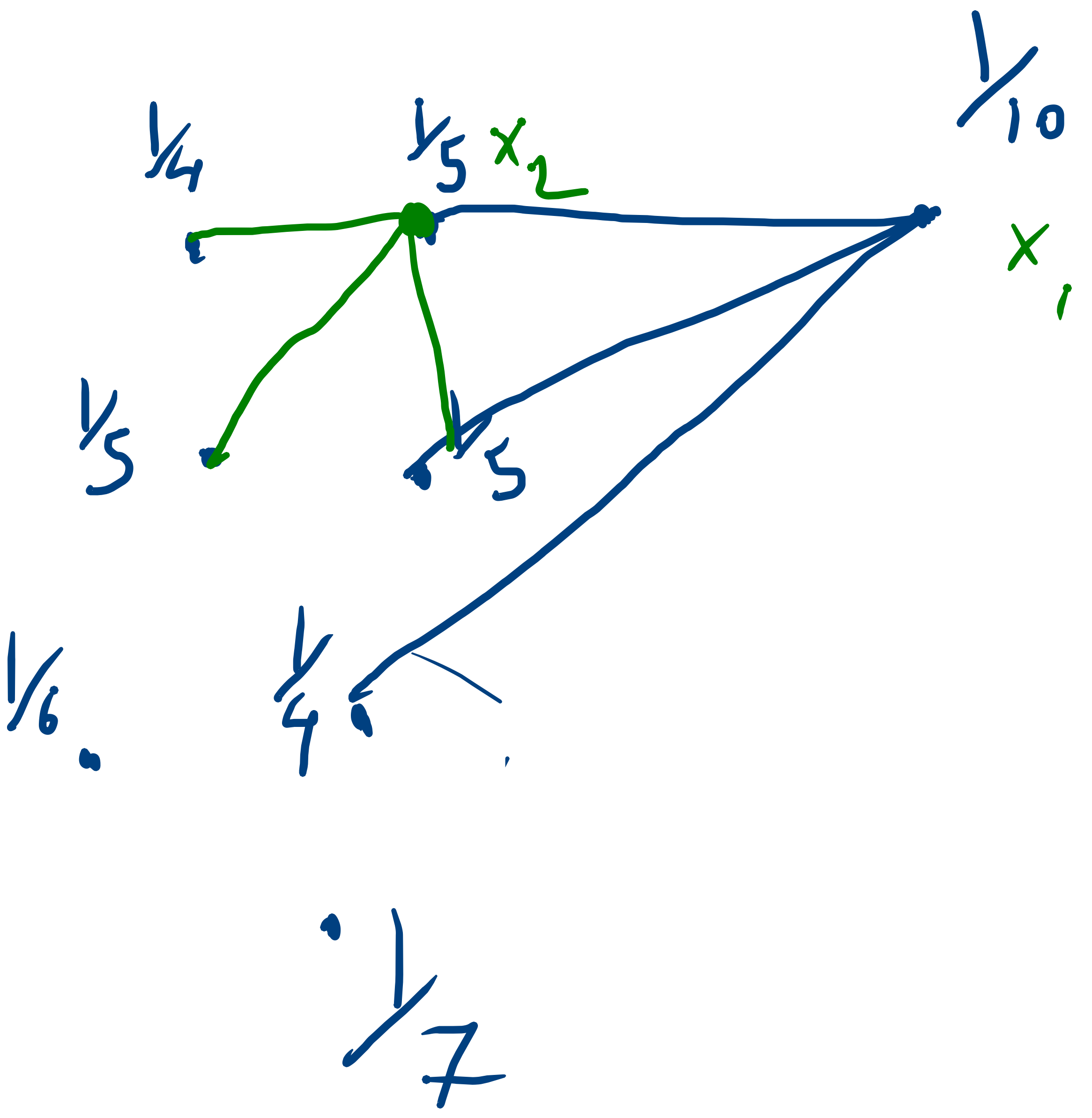
Local outlier factor

$$LOF_A = \frac{\text{Avg density (lrd) of neigh of } A}{\text{Density (lrd) of } A}$$

$$= \frac{\sum_{B \in N_k} lrd_k(B)}{|N(A)|\, lrd_k(A)}$$

For outlier LOF

(A) $= 1$ (B) < 1 (C) $\gg 1$



$$\mathrm{lof}_{x_1} = \frac{\frac{1}{3}\left(\frac{1}{5} + \frac{1}{5} + \frac{1}{4}\right)}{\frac{1}{10}} \approx 1.5$$

$$\mathrm{lof}_{x_2} = \frac{\frac{1}{3}\left(\frac{1}{4} + \frac{1}{5} + \frac{1}{5}\right)}{\frac{1}{5}} = \frac{2.5}{3} \leq 1$$

Disadvantage of LOF

→ Find optimal K (K-NN)

→ High dimension calcuting K-NN is expensive

→ Dependent on contamination value

LOF

$x_1 \rightarrow 10$

$x_2 \rightarrow 5$

$x_3 \rightarrow 1$

$\vdots$

$x_{10} \rightarrow$

$$\frac{\text{LOF} >> 1}{}$$

$\rightarrow$ Outlier

Contam = 0.1

$\downarrow$

10%