

④ ISOLATION FOREST :

→ '1-D' dataset

'series' / 'columns'

Finance :

Age Salary

20 10000

25 15000

26 14000

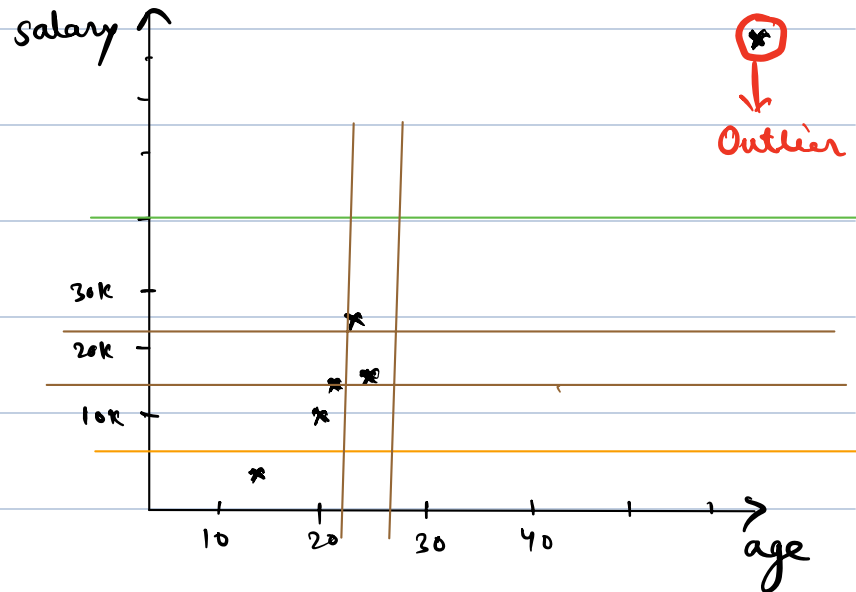
23 12000

72 1,00000

15 5000

→ IQR ✓

→ Z-score ✓

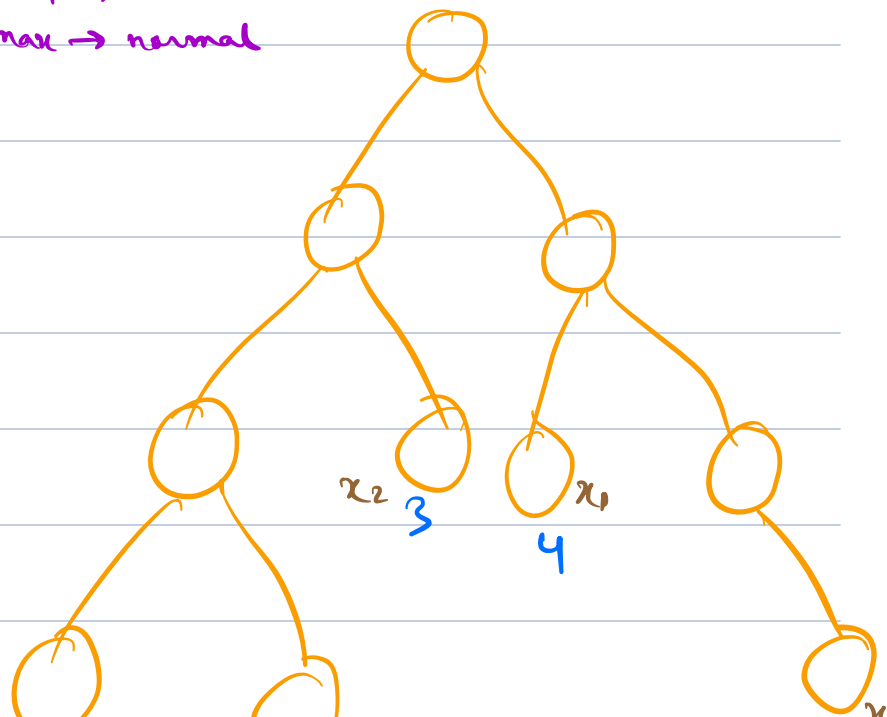
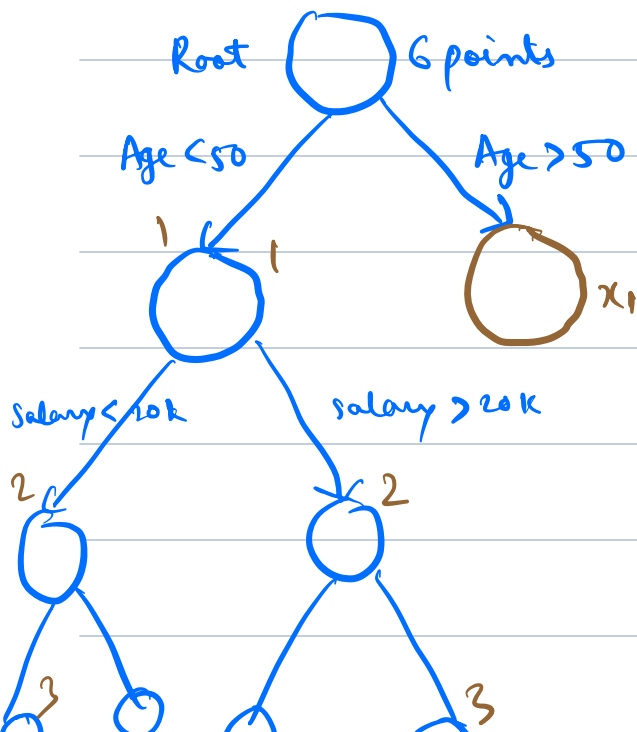


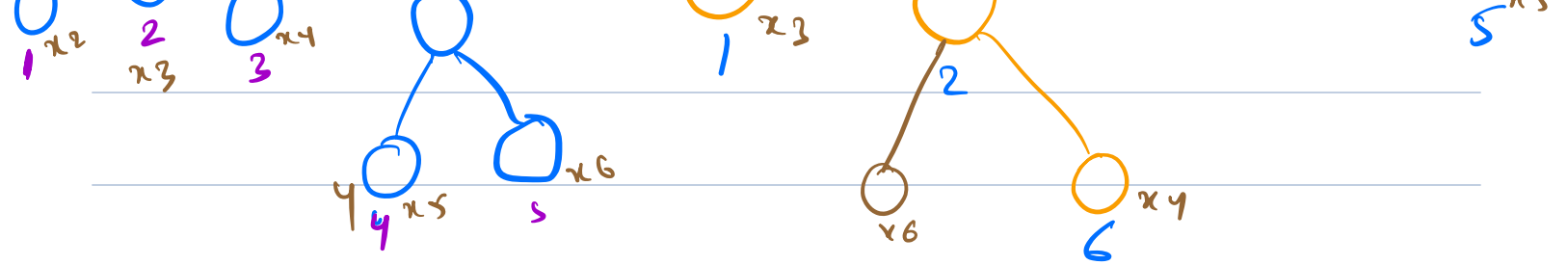
FOREST → "Combination of Random Trees"

→ Random splits

Main Idea: Tries to isolate a point by doing random splits

min → outlier
max → normal





We get 'n' trees \rightarrow and it gives depth of each data point

	x_1	x_2	x_3	x_4	x_5	x_6
3 \leftarrow Tree 1	1	3	3	3	4	4
3 \leftarrow Tree 2	2	3	3	2	3	3
\vdots	$E(h(x)) = \frac{1 + 2 + \dots}{n}$					
\vdots						
Tree 'n'	$c(m) = \frac{3 + 3 + \dots}{n}$					

Imp Mathematical Intuition:

\rightarrow Calculate 'Anomaly Score' for every data point.

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

$m \rightarrow$ Total Data points

$x \rightarrow$ Random data point

$E(h(x)) \rightarrow$ Average path length of 'x' data point in a tree

$c(m) \rightarrow$ Average depth of a tree.

Conditions to check:

① $E(h(x)) << c(m) \rightarrow S(x, m) \simeq 1$

② $E(h(x)) >> c(m) \rightarrow S(x, m) \simeq 0$

③ $E(h(x)) \simeq c(m) \rightarrow S(x, m) \simeq 0.5$

If $S(x, m) > 0.5$ and close to 1 \rightarrow Outlier

If $S(x, m) < 0.5$ and close to 0 \rightarrow Normal Data Point

Contamination \rightarrow percentage / proportion of outliers in data

Data $\rightarrow 100\%$

$\hookrightarrow 5\%$ outliers

contamination $= 0.05$

$\hookrightarrow 10\%$ outliers

$= 0.1$

⑤ LOF - Local Outlier Factor

\hookrightarrow is based on : ① k-nn ✓

② Density ✓

density (x_i) \ll density of 'k' points

↳ x_i as an 'outlier'

• Mathematical Intuition:

$$LOF_k(A) = \frac{\text{Avg. neighborhood density of } A}{\text{Density of } A}$$

$$= \frac{\sum_{B \in N_k(A)} lsd_k(B)}{|N_k(A)| \cdot lsd(A)}$$

Conditions to check:

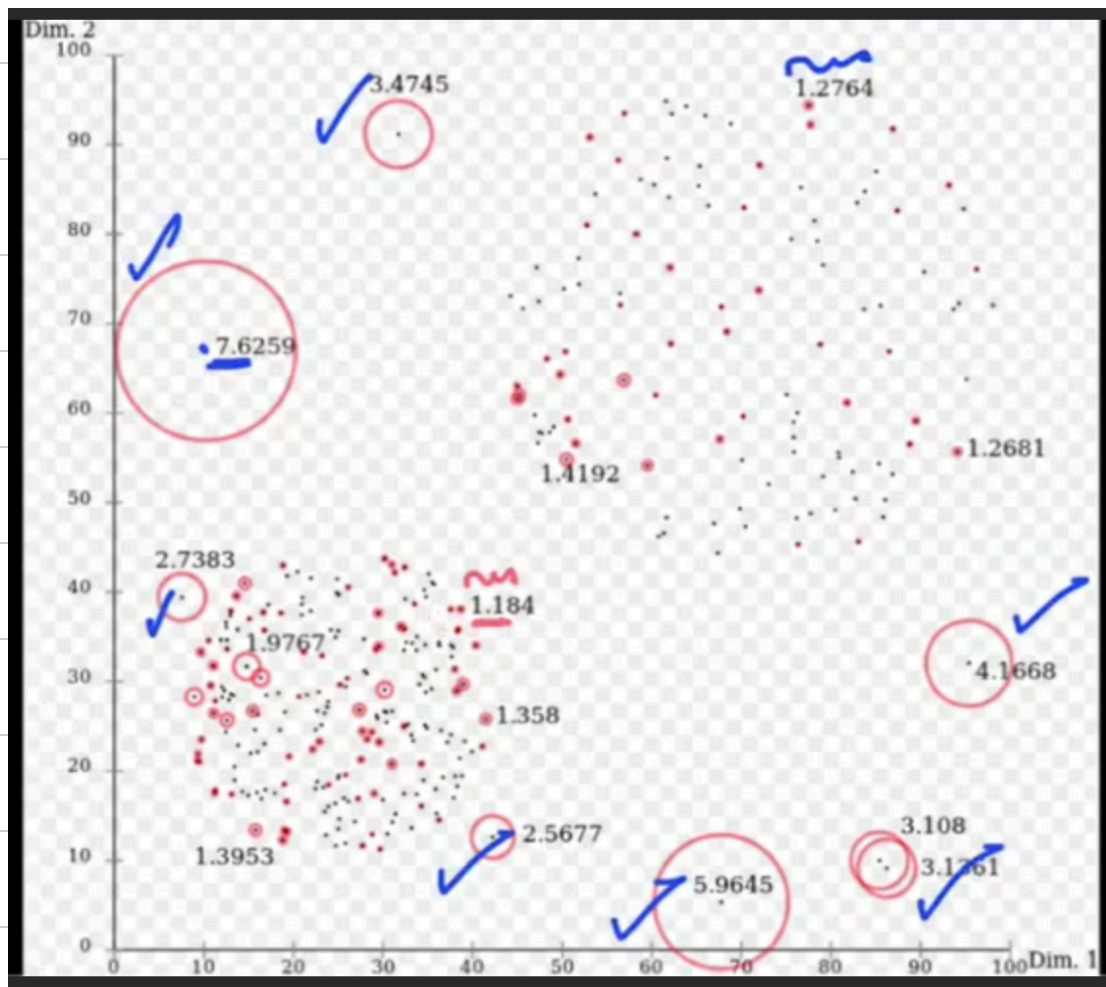
① $LOF(A) = 1 \rightarrow A$ has same density as 'k' neighbors. \rightarrow Normal

② $LOF(A) > 1 \rightarrow A$ has lower density compared to 'k' neighbors

• It may or may not be an outlier

• $LOF(A) \gg 1 \rightarrow A$ is an outlier

③ $LOF(A) < 1 \rightarrow A$ has more density compared to 'k' neighbors
 \rightarrow Normal



Parameters:

- ① Contamination
- ② n-neighbors

Announcements:

- ✓ ① Dimensionality Reduction
- ✓ ② Clustering
- ✓ ③ Anomaly Detection

Next 2 classes
"End-to-End Project"