

④ ISOLATION FOREST:

→ "1-D" dataset

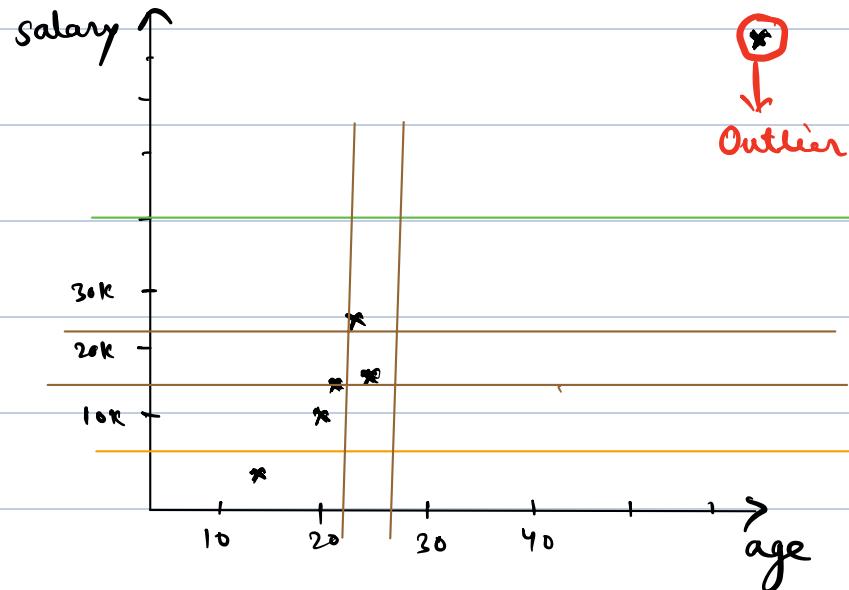
finance:

Age	Salary
20	10000
25	15000
26	140000
23	120000
72	1,000000
15	5000

'series' / 'columns'

→ IQR ✓

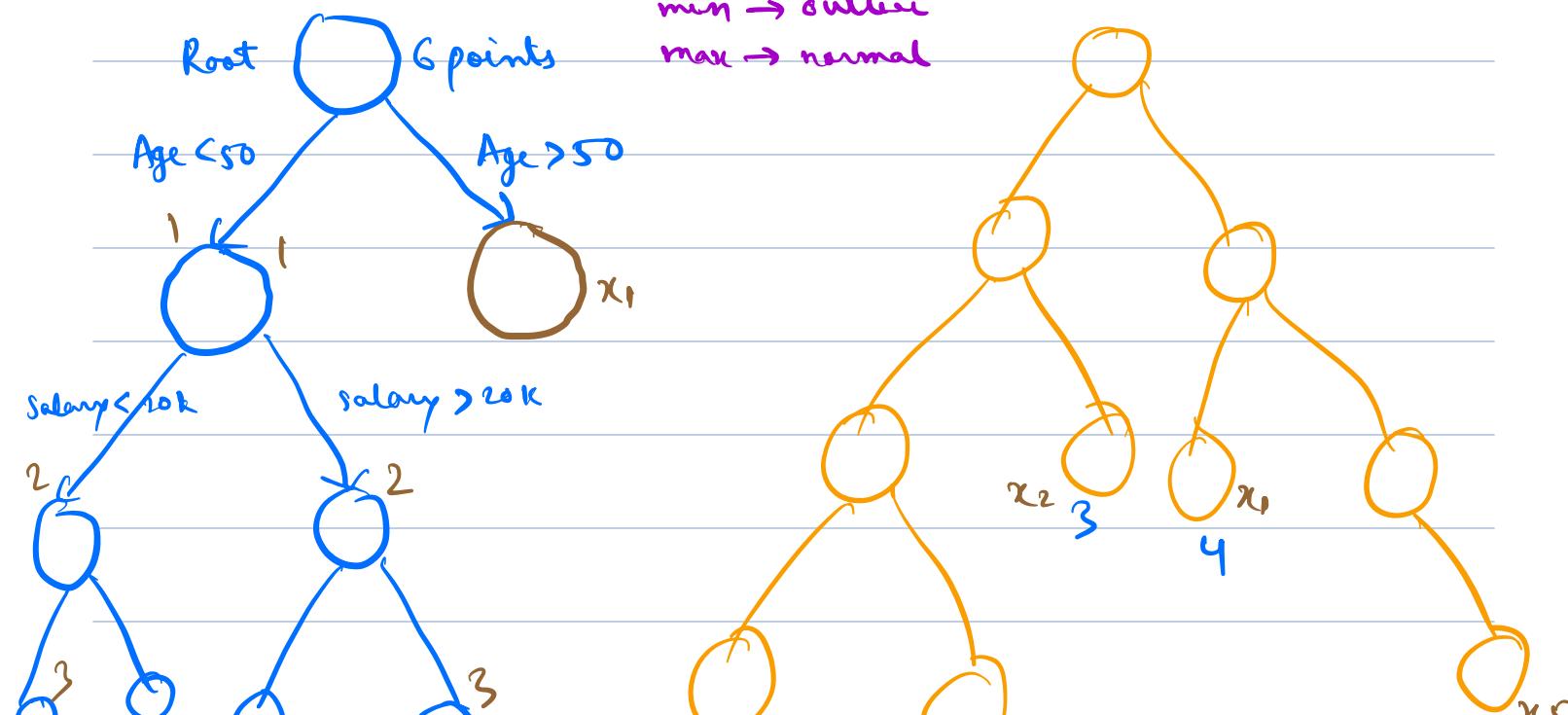
→ Z-score ✓

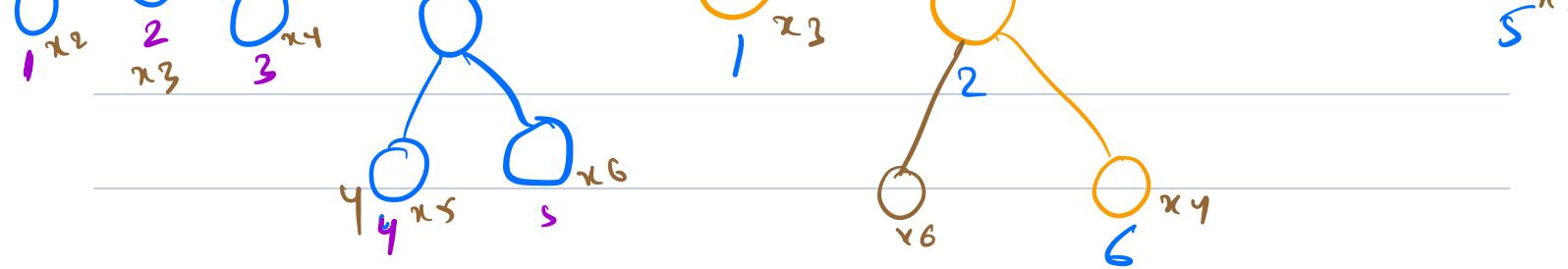


FOREST → "Combination of Random Trees"

Random splits

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





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
$z \leftarrow \text{Tree 1}$	1	3	3	3	4	4
$z \leftarrow \text{Tree 2}$	2	3	3	2	3	3
\vdots	$E(n(x)) = 1 + 2 + \dots$					
\vdots	$c(m) = 3 + 3 + \dots$					
$\text{Tree } n$						

~~Imp~~ Mathematical Intuition:

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

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

$m \rightarrow$ 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 $E(h(x)) \ll c(m) \rightarrow s(x, m) \approx 1$

② $E(h(x)) \gg c(m) \rightarrow s(x, m) \approx 0$

③ $E(h(x)) \approx c(m) \rightarrow s(x, m) = 0.5$

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

$$\text{contamination} = 0.05$$

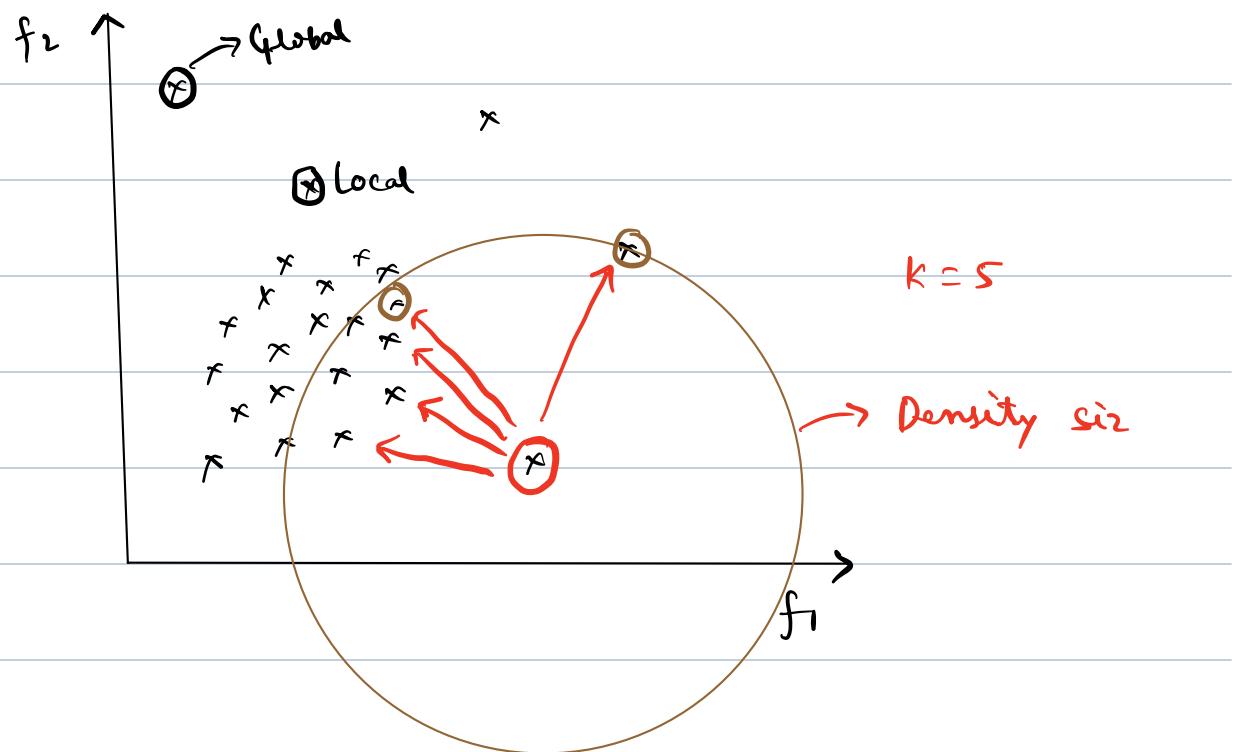
$\hookrightarrow 10\%$ outliers

$$= 0.1$$

⑤ LOF - Local Outlier Factor

\hookrightarrow is based on : ① k-nn ✓

② Density ✓



Density size of x_i :

Avg. - density size for k -neighbors

$\text{Size}(x_i) \gg \text{Size}(k)$

$\text{Density}(x_i) \ll \text{Density}(k)$

$x_i \rightarrow \text{Outlier}$

→ The core idea of LOF is to compare:

- density of a point with its neighbor's density

density (x_i) << density of 'k' points

↳ x_i as an 'Outlier'

• Mathematical Intuition:

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

$$= \frac{\sum_{B \in N_k(A)} \text{ld}_k(B)}{|N_k(A)| \cdot \text{ld}(A)}$$

Conditions to check :

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

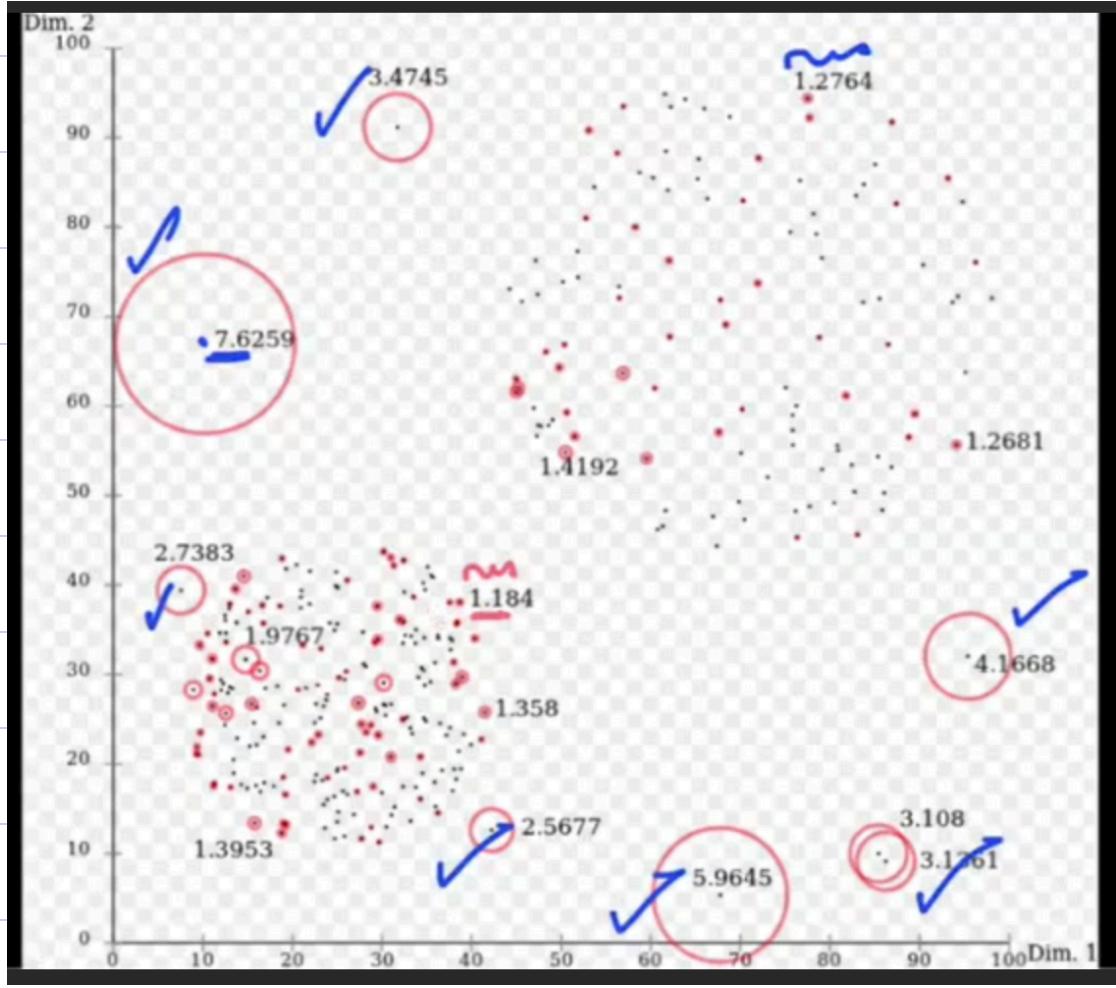
② $\text{LOF}(A) > 1 \rightarrow A$ has lower density compared to 'k' neighbors

- It may or may not be an outlier

- $\text{LOF}(A) \gg 1 \rightarrow A$ is an outlier

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

↳ Normal



Parameters:

- ① Contamination
- ② n-neighbors

Announcements:

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

Next 2 classes
 "End-to-End Project"