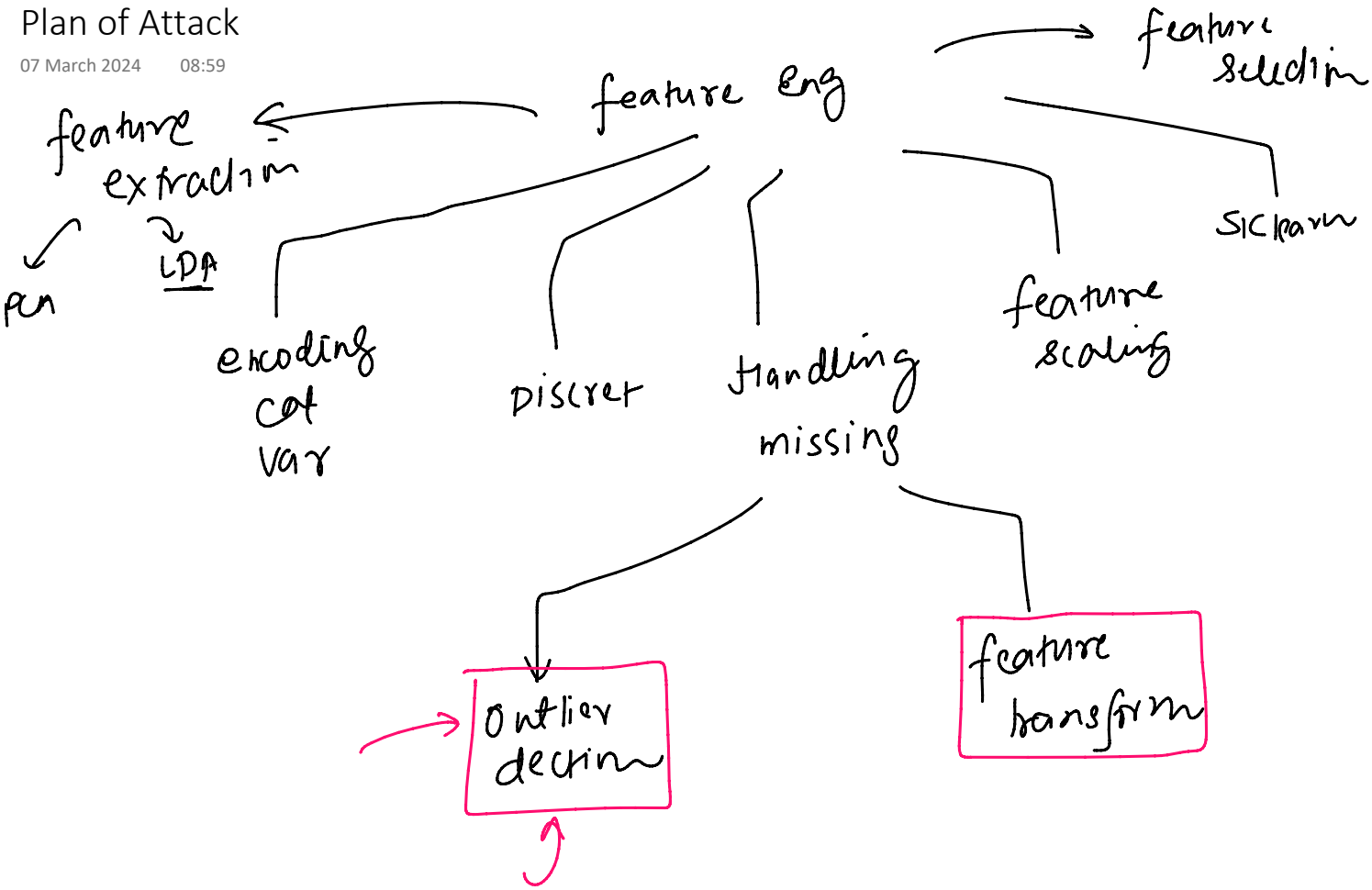
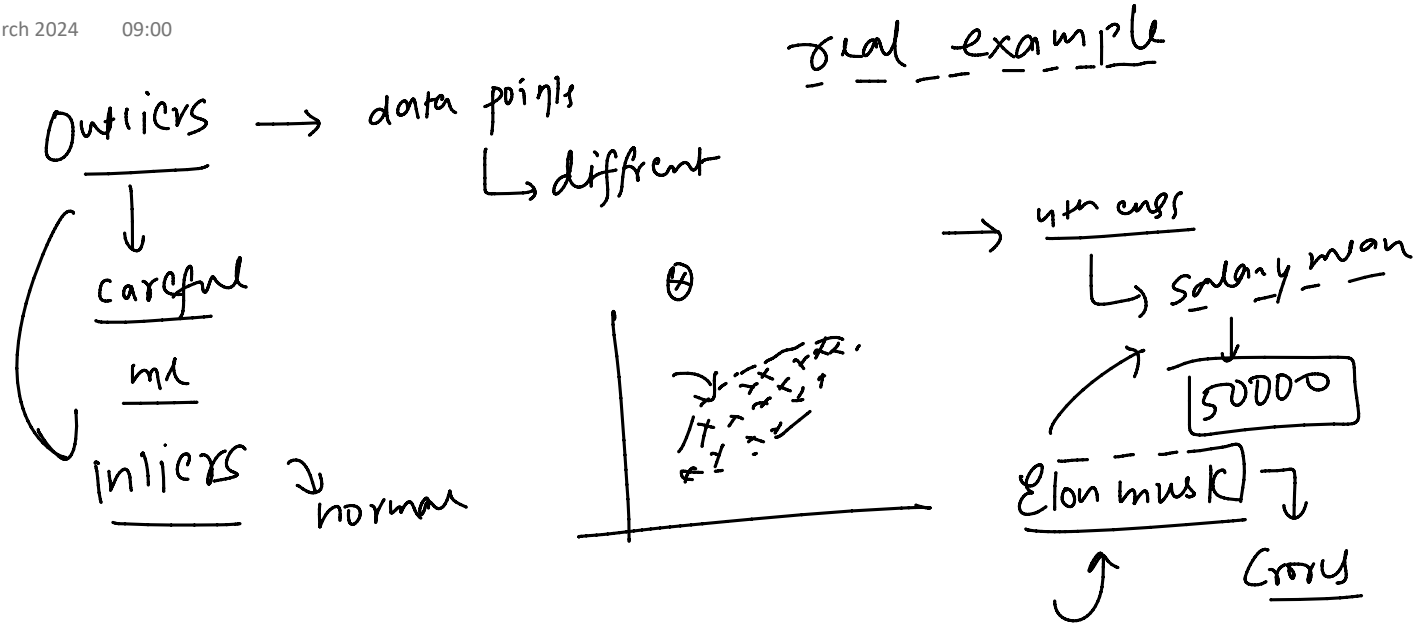
Plan of Attack

07 March 2024 08:59



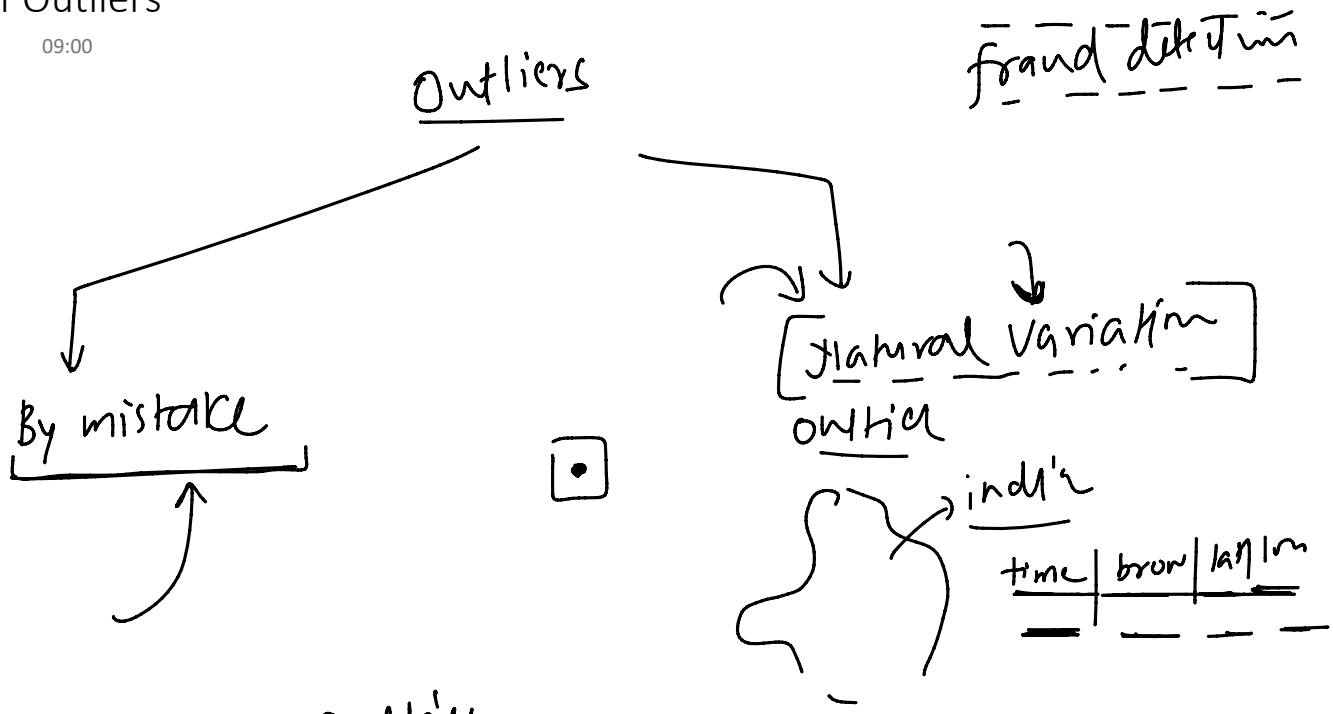
What are Outliers

07 March 2024 09:00



Types of Outliers

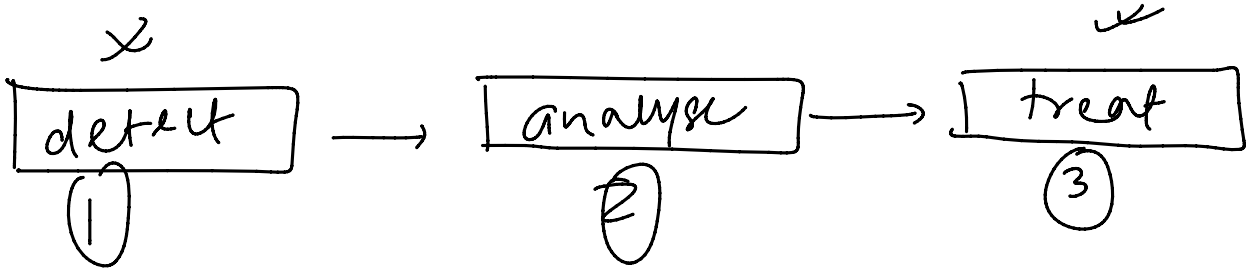
07 March 2024 09:00



Outlier

1) Outlier detection →
2) By mistake / natural var
3) outlier treatment

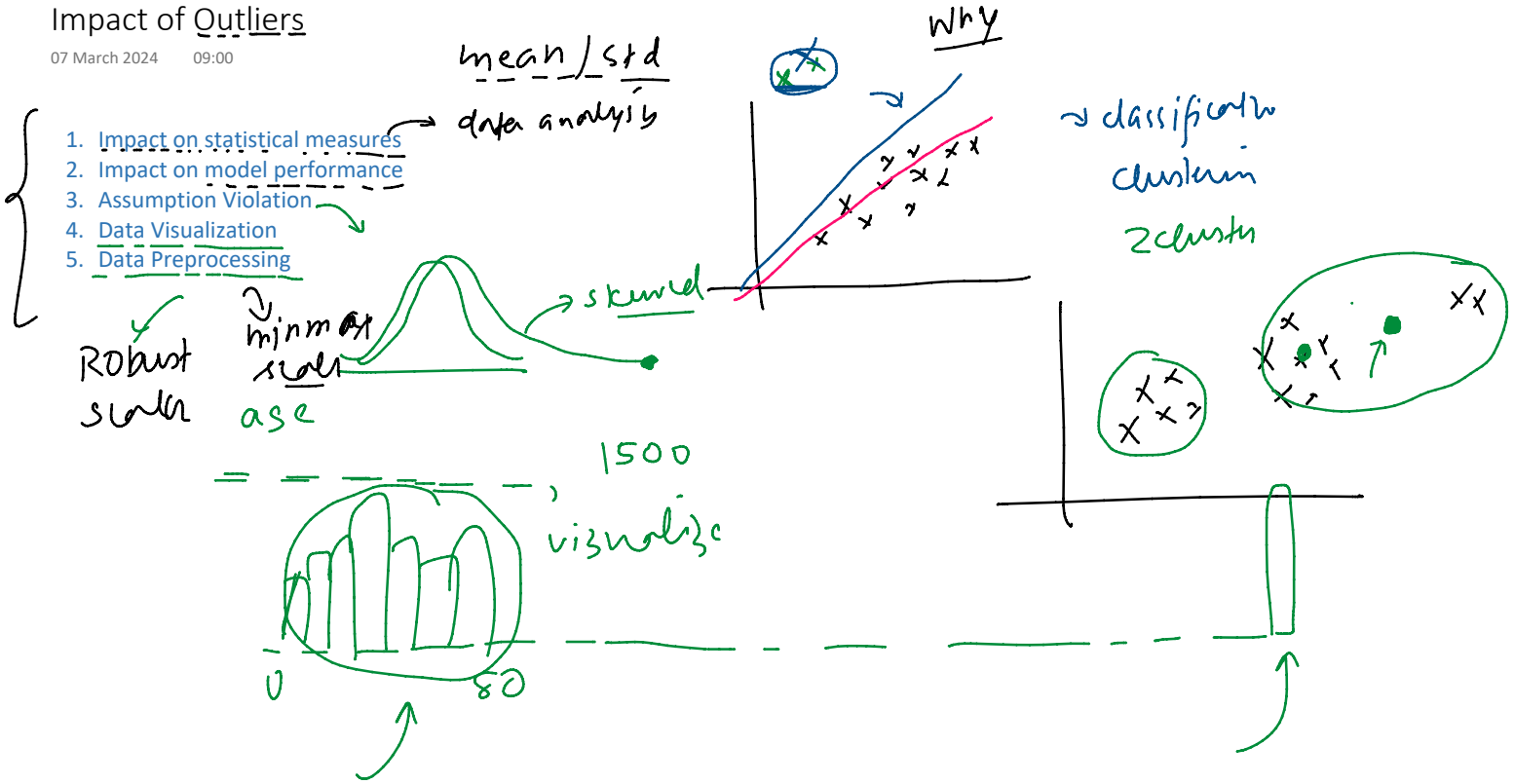
flow outlier



Impact of Outliers

07 March 2024 09:00

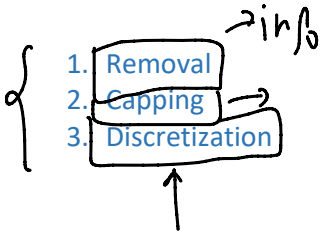
1. Impact on statistical measures
2. Impact on model performance
3. Assumption Violation
4. Data Visualization
5. Data Preprocessing

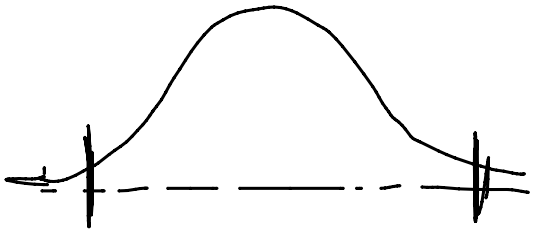


How to deal with outliers

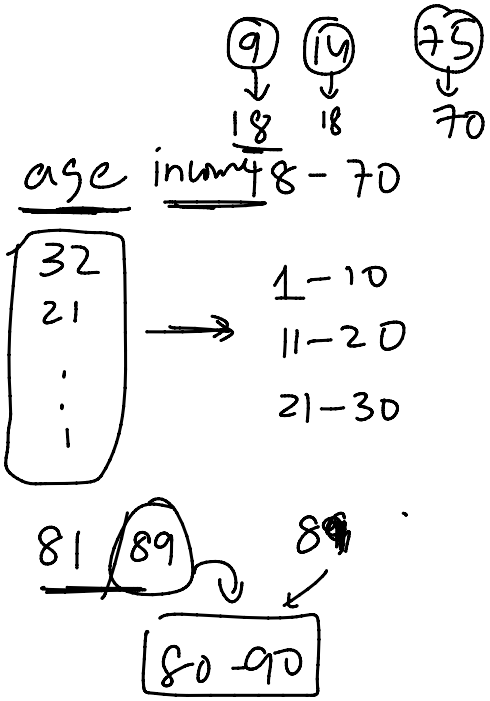
07 March 2024 10:25

detect



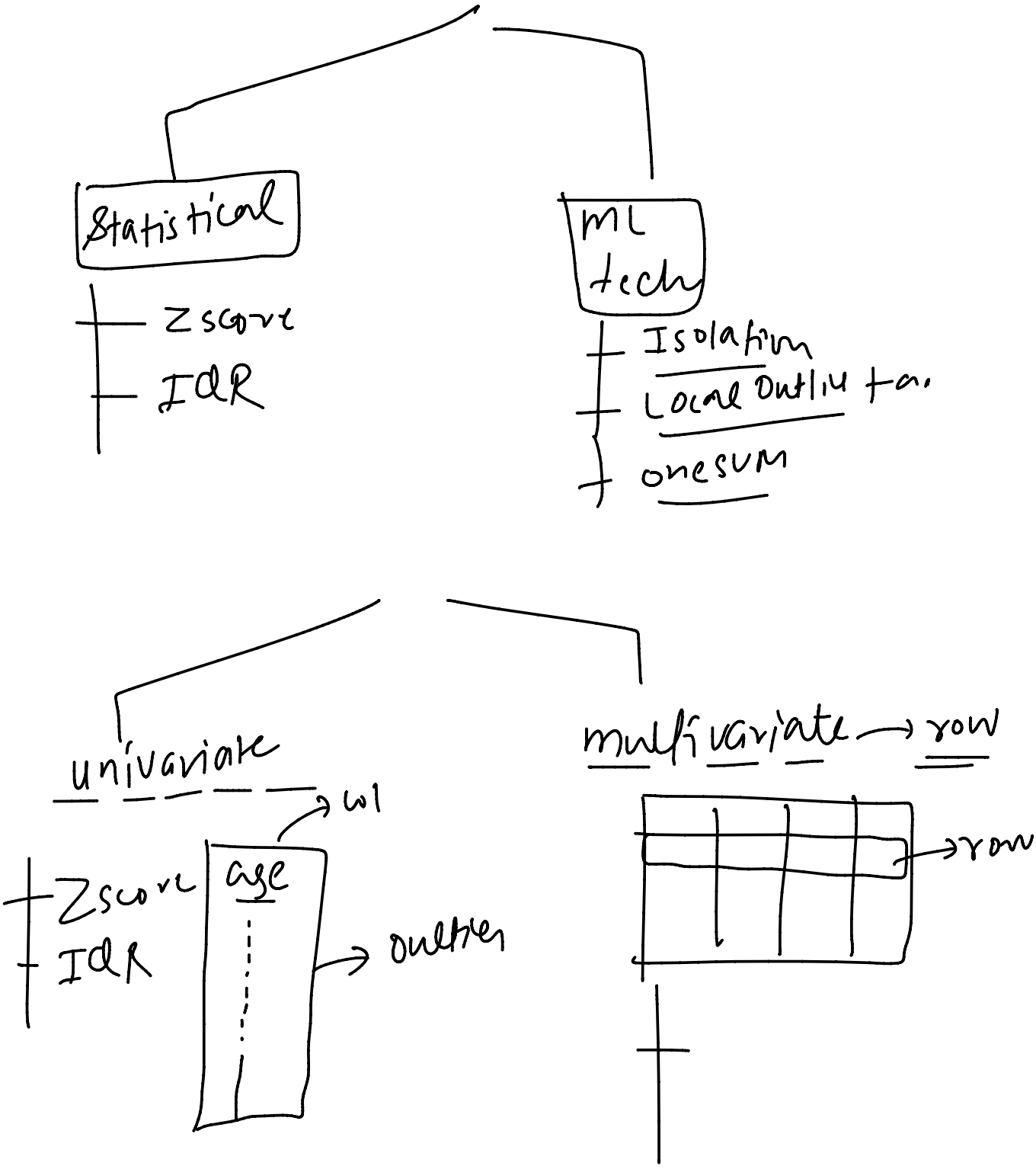


num → bin → cat



Outlier Detection Techniques

07 March 2024 09:41



Z-Score → univariate outlier det

07 March 2024 09:41

age	Income	purchase
32	32000	0
64	150000	1

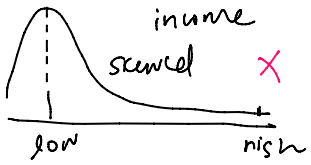
Z score

w Z score

$\mu \rightarrow$
$\sigma \rightarrow$

Big assumption

→ Normal / Normal like



$\frac{x_i - \mu}{\sigma}$ ←

[-3 to 3]

$[\mu + 3\sigma \;/\; \mu - 3\sigma]$

80

6

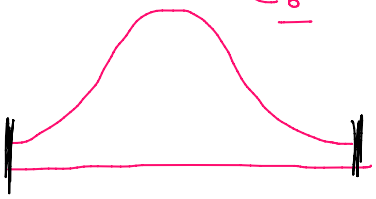
[62 1, 3]

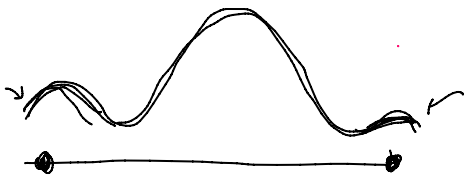
outliers

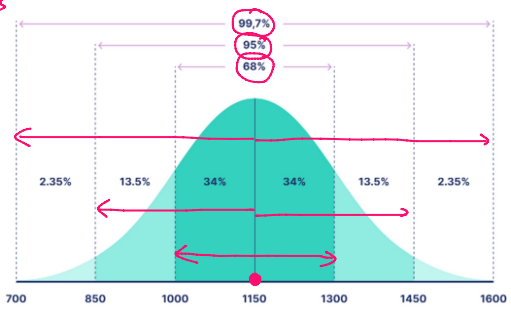
$\mu - 3\sigma$

$\mu + 3\sigma$

0.03 % outside trial





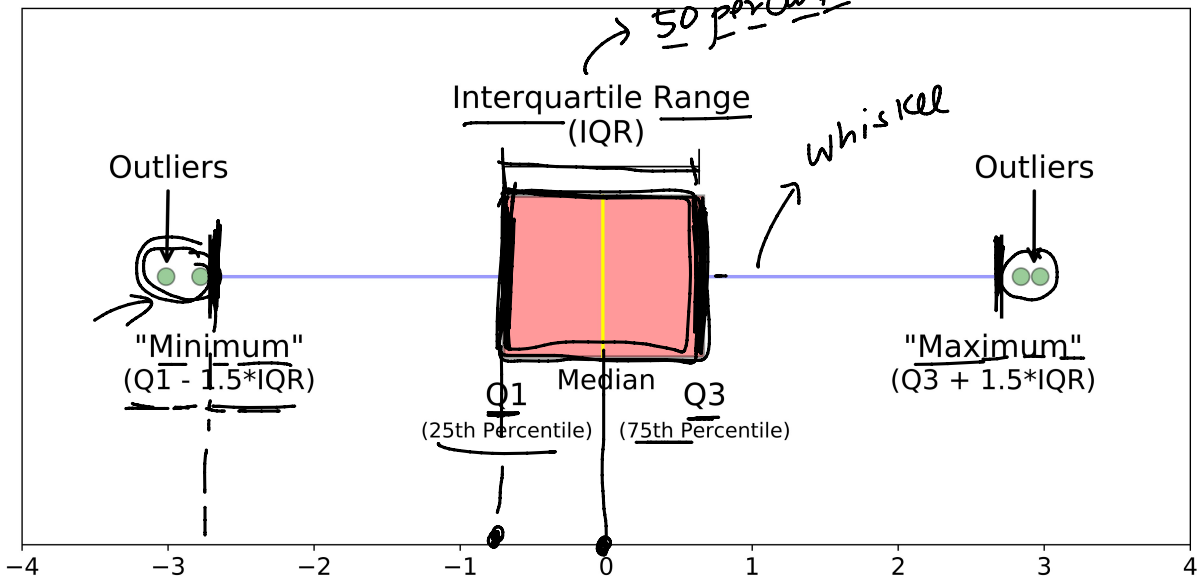


IQR and BoxPlot

07 March 2024 09:41

normal → col → skewed

numerical col



boxplot

percentile

25th

50th pt

75th pt

median

5 pt → 10

Q3 + 1.5 IQR

99% → 360

Q1 − 1.5 IQR

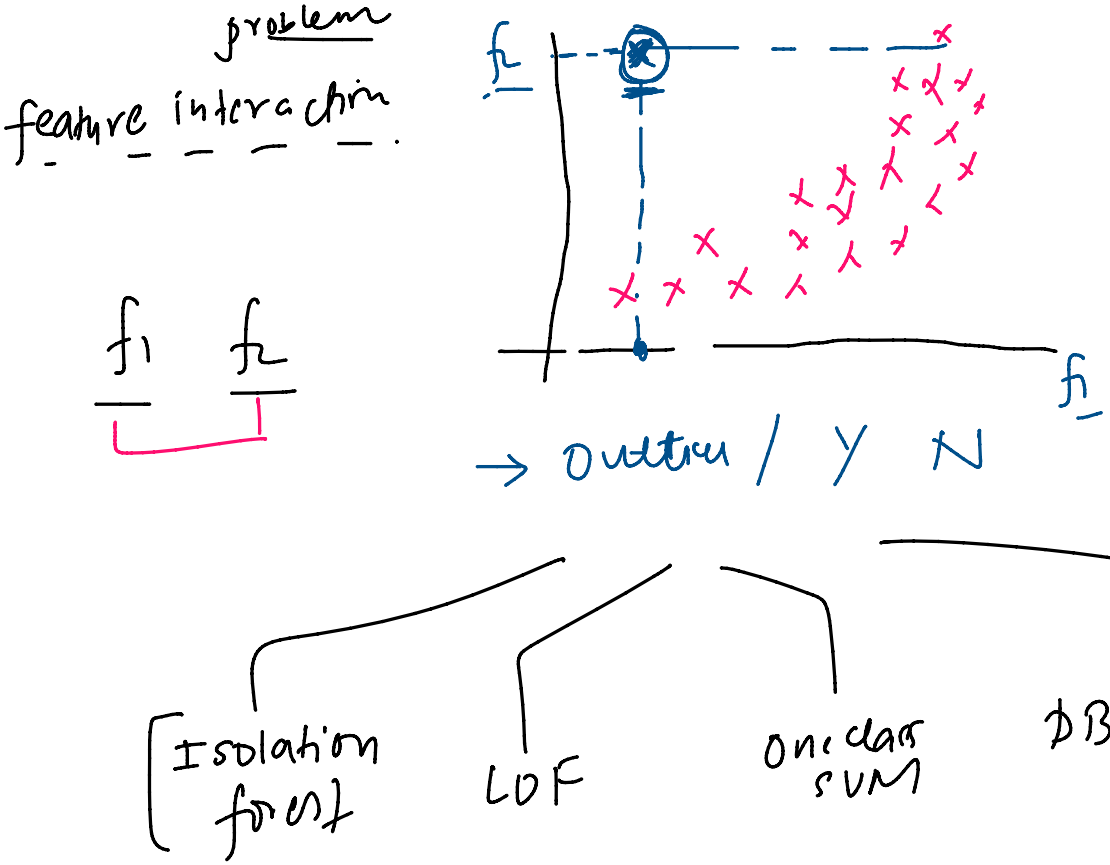
Problem with Univariate Techniques

07 March 2024 16:21

problem

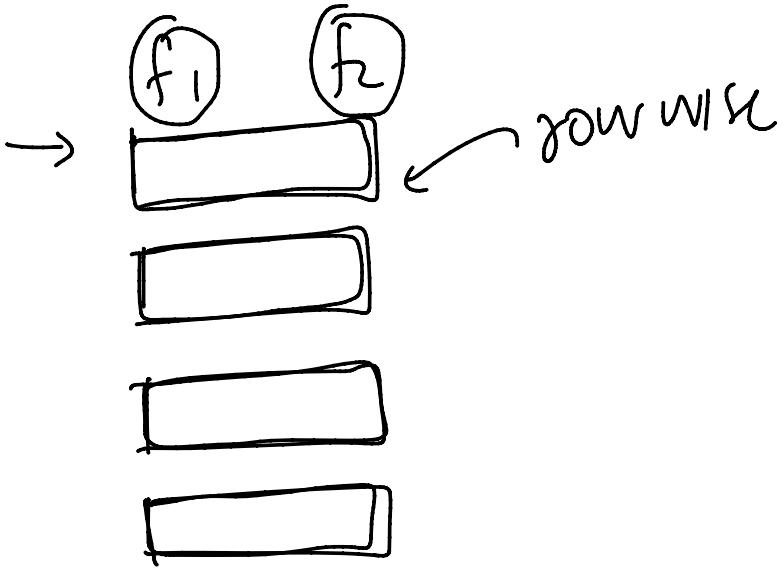
feature interaction

Why?



Multivariate Outlier Detection

07 March 2024 16:38



Isolation Forest
07 March 2024 09:42

Anomaly detection
Random for
decision trees

What → How

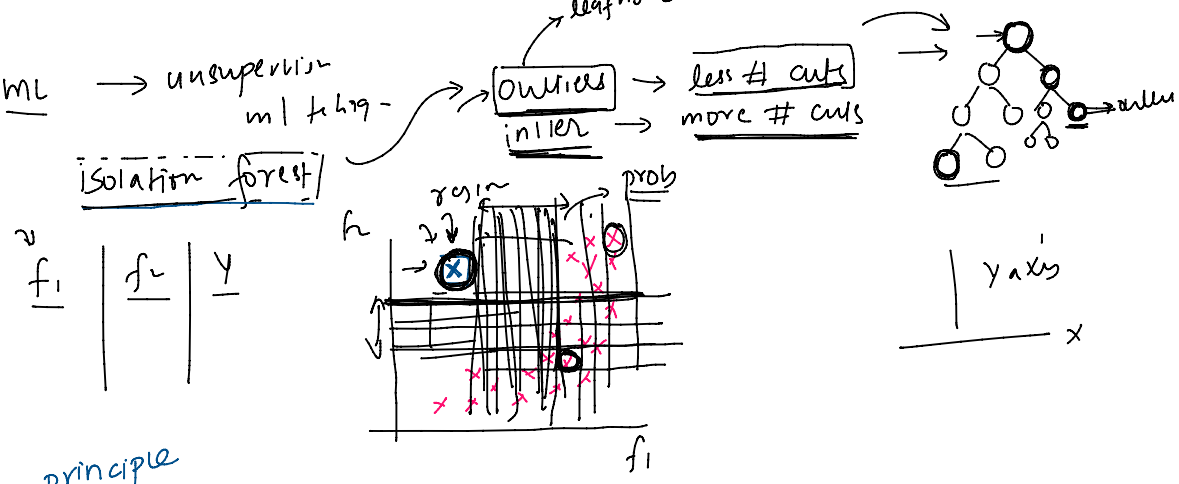
(Y) → tree building entropy, gini

ML → unsupervised ml thing

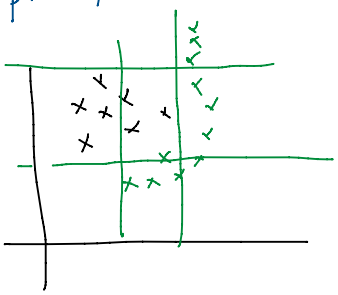
upar / neeche
leaf node

outliers → less # cuts
inlier → more # cuts

isolation forest

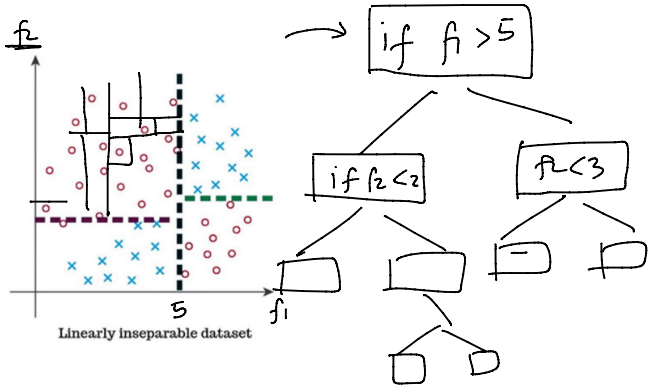


principle

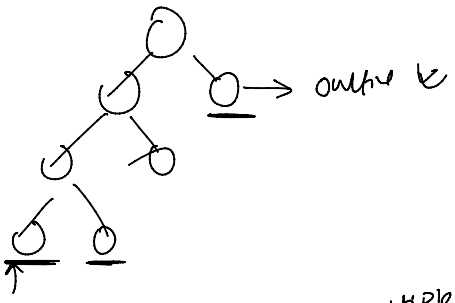


Extra trees

if

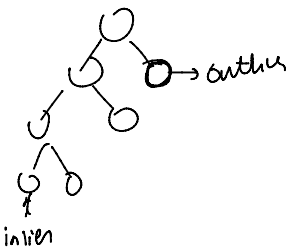


decis → fully grow

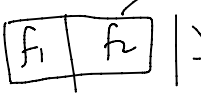


How → isolation forest → multiple trees involved

extra trees → 10 trees

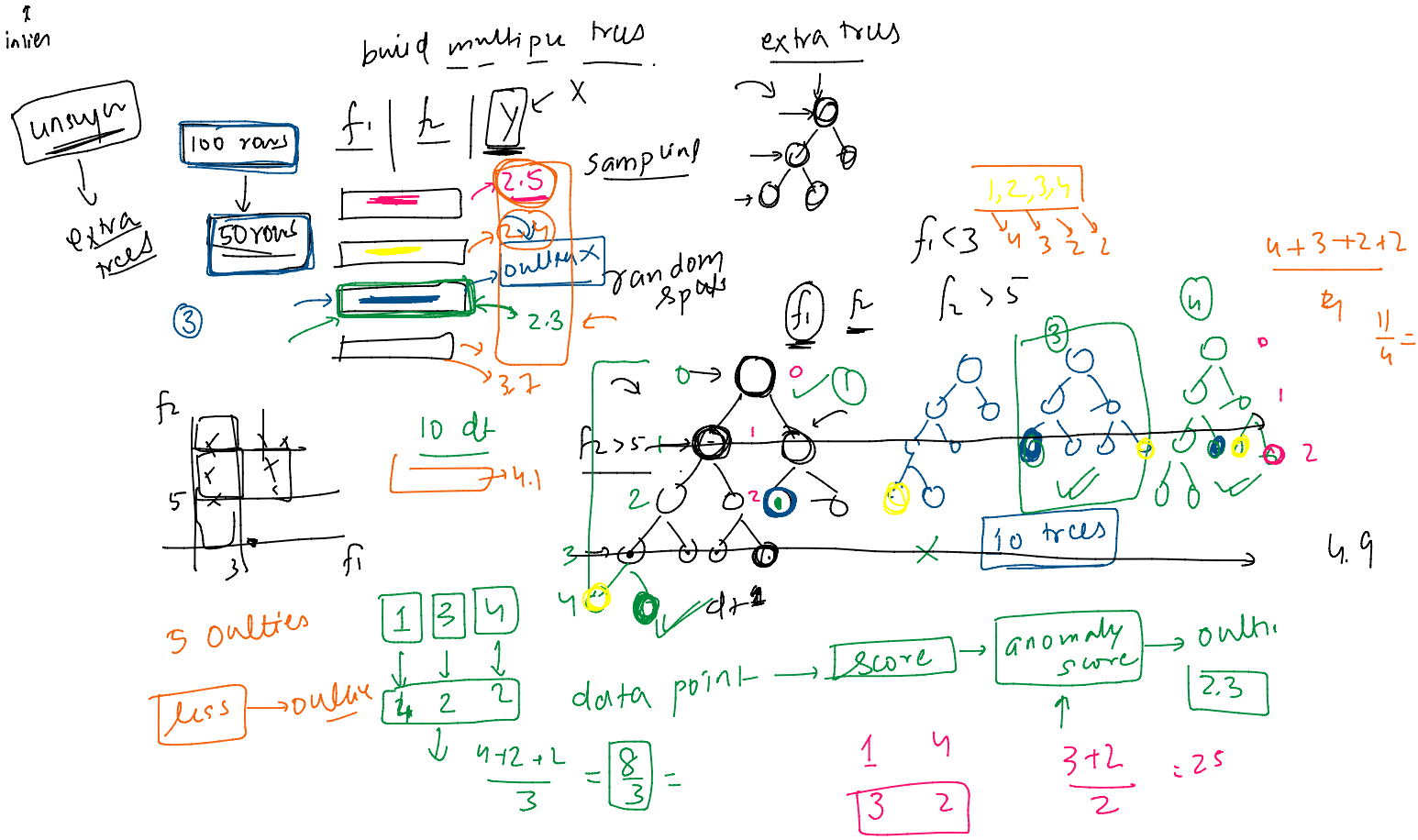


10 trees



build multiple trees

extra trees

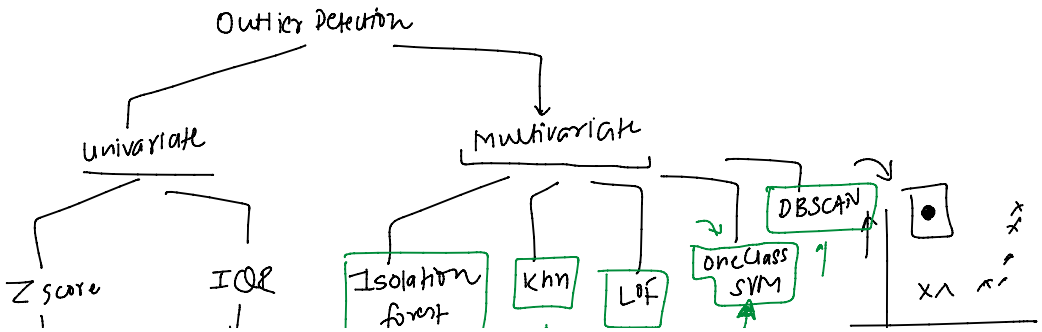


{anomaly score} →

→ decide # extra trees
→ max sample
→ build all trees
→ for every row calculate the avg position of leaf in the tree in which they are present

5 outlier

1 anomaly / row



Z score → Normal

IQR

Isolation forest

knn

LOF

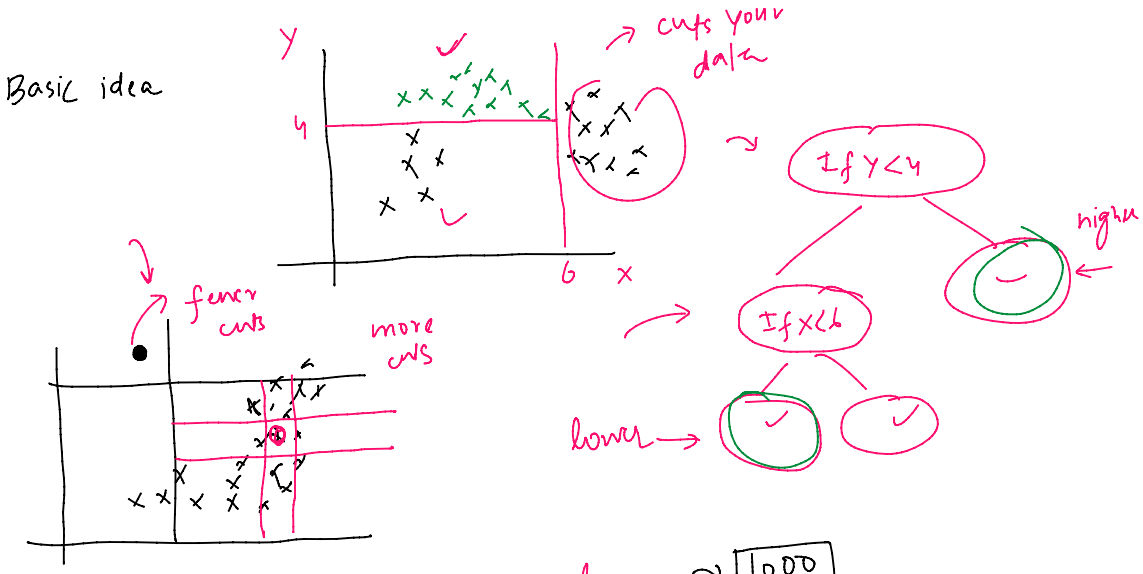
OneClass SVM → trial sessions

→ comparison

Quick Revision

Isolation forest

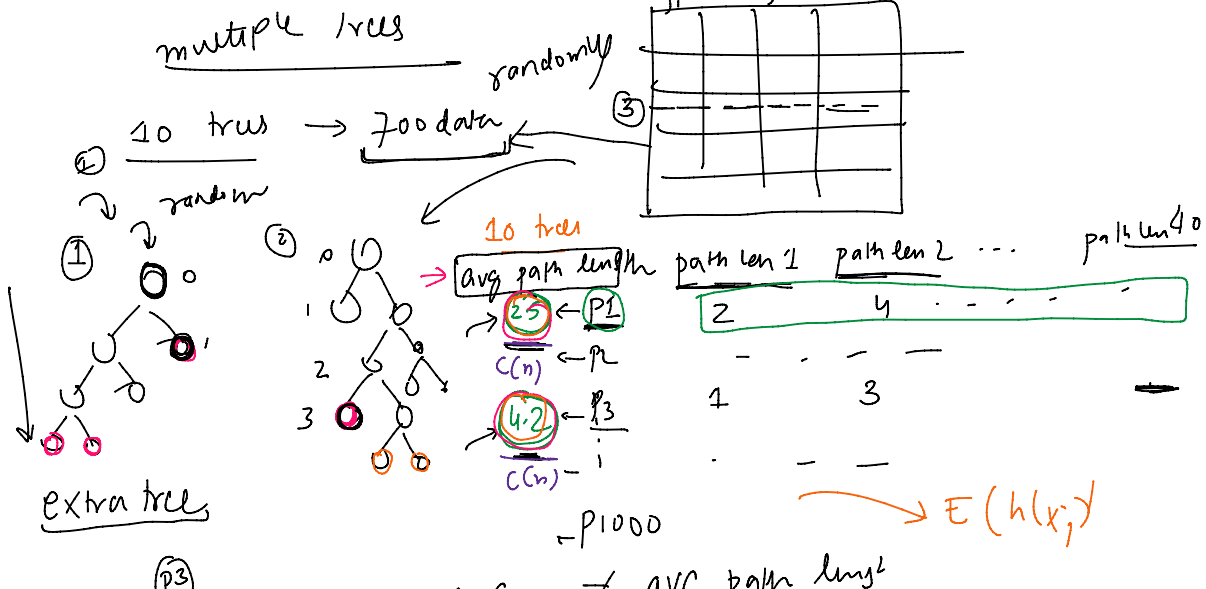
Basic idea



isolation → unsupervised

multiple trees

10 trees → 700 data → randomly → 1000

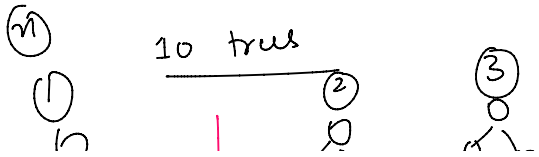


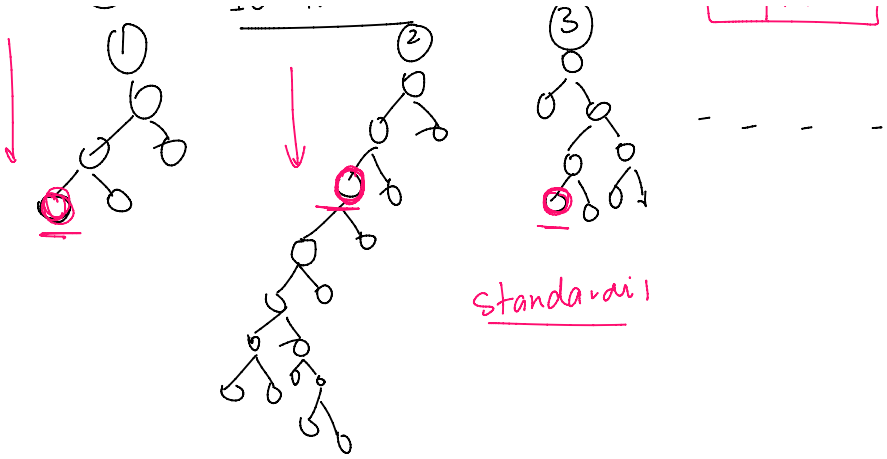
extra tree

P3

anomaly score ≠ avg path length

depth = 2





standard

Calculation of Anomaly Score

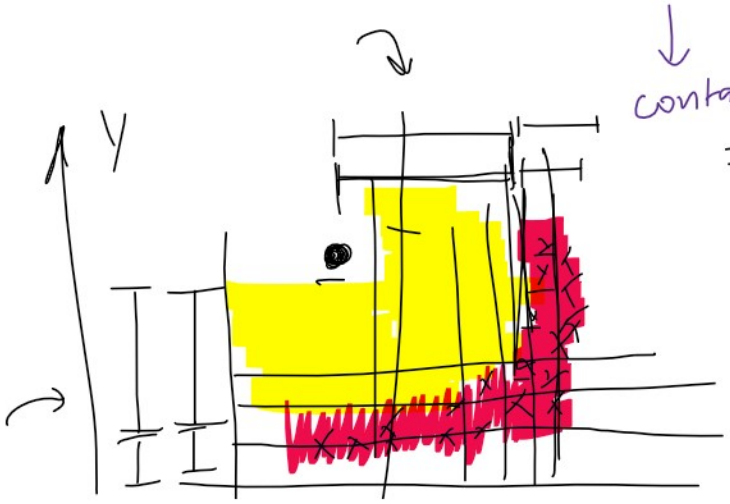
07 March 2024 17:50

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

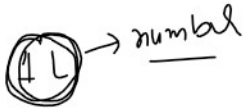
$$E(h(x)) = \frac{1}{N}\sum_{i=1}^{N} h_i(x)$$

$$c(n) = 2(\ln(n-1) + 0.5772156649) - \frac{2(n-1)}{n}$$

It's a baseline measure of how deep we would expect to go in a tree of n samples if the points were uniformly distributed (i.e., no anomalies).



1- 5



[1 - 100]

1) can't say

2) X > Y

3) X = Y

4) X < Y

X + Y = 1 Lakh

(1 - 90) (91 - 100)

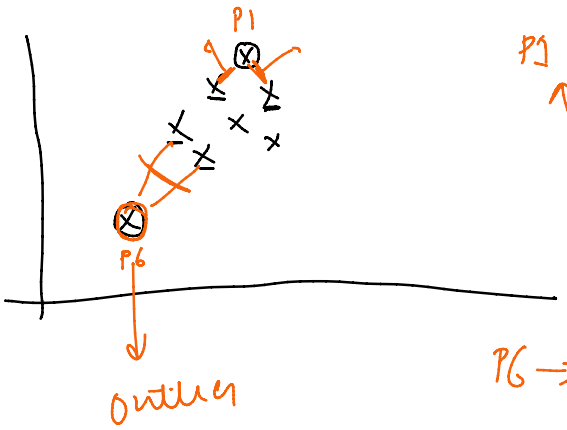
X Y

Outlier Detection using KNN

12 March 2024 08:24

Disadvantages

K=2

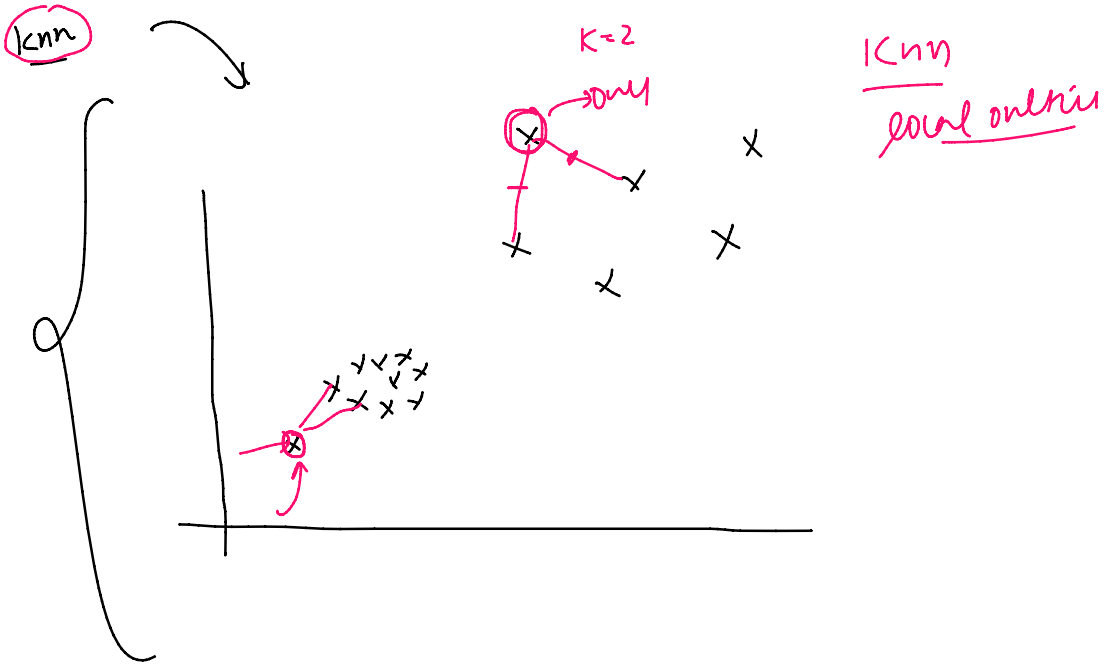


P1 → (3+5)/2 = 4 → avg neighbor distance

P6 → (8+10)/2 = 9 → avg nei. dis

too simple

↳ there is some problem

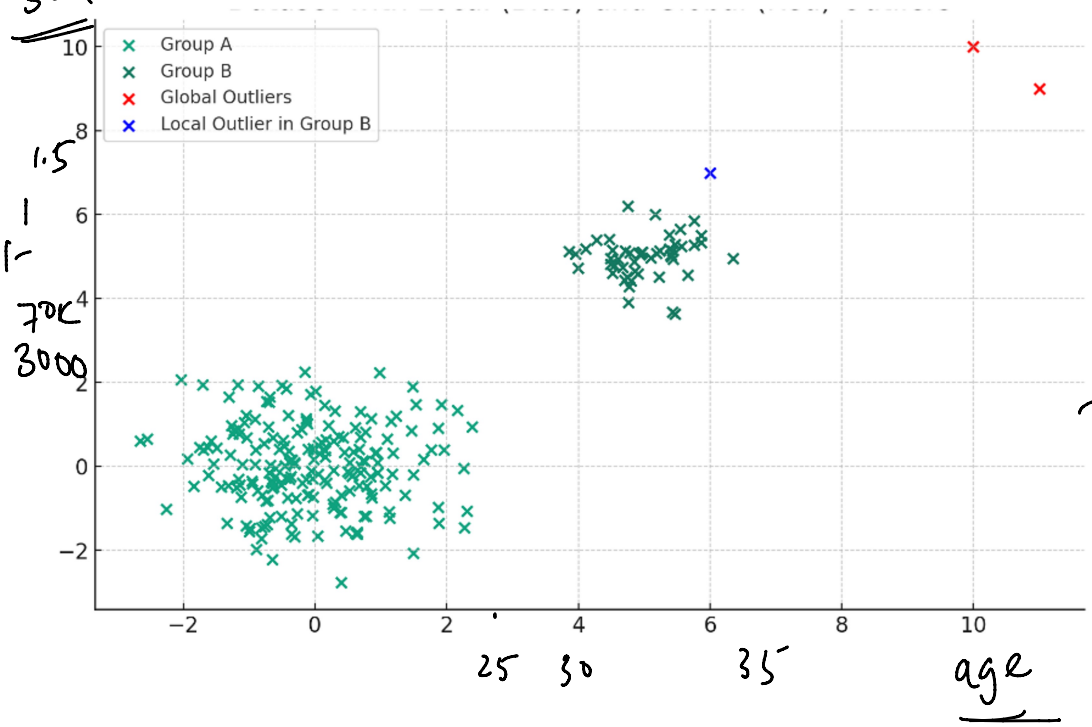


LOF

↳ local outlier → knn

Local vs Global Outliers

12 March 2024 08:35



sal

1.5

1

70K

3000

25 30 35

age

global

local

local outlier

knn → good algo

students

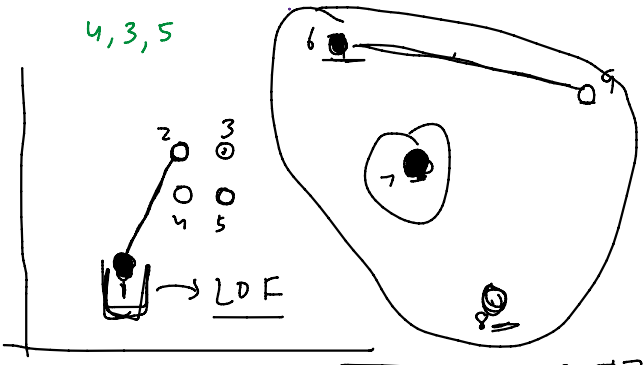
promise

1.5 L

Local Outlier Factor

07 March 2024 09:42

7 → LRD → 0.06
6 → LRD → .0.075
8 → LRD → 0.075
9 → LRD → 0.1



Step 1: Calculate k-distance for each point

- For every point p in the dataset:
  - Calculate the distance from p to every other point.
  - Determine the k-th smallest distance; this is the k-distance for point p.

Step 2: Compute reachability distance

- For every point p and each of its neighbors o:
  - The reachability distance from p to o is the maximum of the k-distance of o and the distance between p and o:

$$\text{reachability distance}_k(p, o) = \max(\text{k-distance}(o), d(p, o))$$

Step 3: Calculate local reachability density (LRD)

- For each point p:
  - Compute the LRD of p as the inverse of the average reachability distance from p to its k-nearest neighbors:

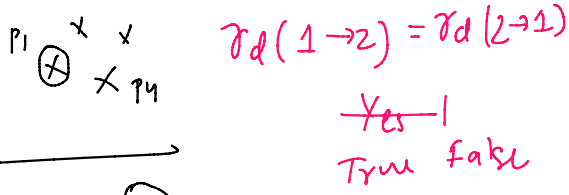
$$\text{LRD}_k(p) = \frac{1}{\frac{\sum_{o \in k\text{-neighbors}(p)} \text{reachability distance}_k(p,o)}{|k\text{-neighbors}(p)|}}$$

$$\frac{0.075 + 0.075 + 0.1}{3}$$

Step 4: Compute the Local Outlier Factor (LOF)

- For each point p:
  - Calculate the LOF by averaging the ratios of the LRD of p's neighbors to the LRD of p:

$$\text{LOF}_k(p) = \frac{\sum_{o \in k\text{-neighbors}(p)} \frac{\text{LRD}_k(o)}{\text{LRD}_k(p)}}{|k\text{-neighbors}(p)|}$$

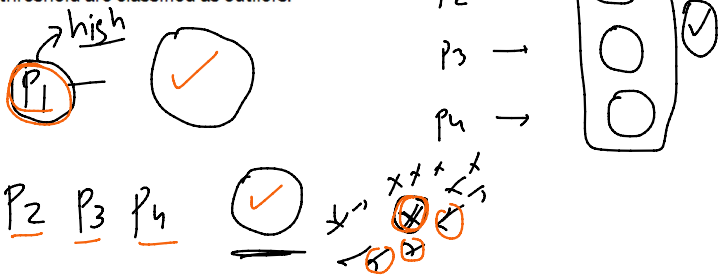


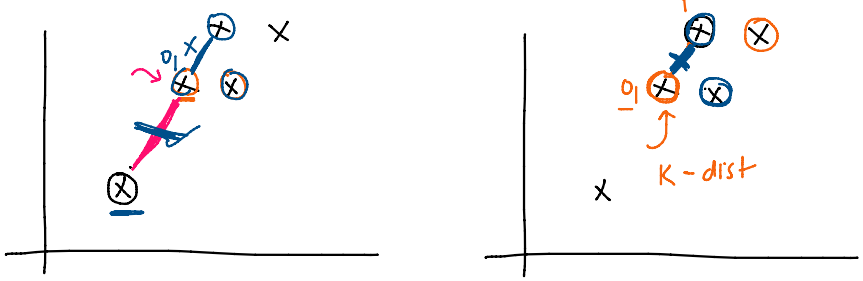
Step 5: Interpret the LOF scores

- Points with an LOF score around 1 are similar in density to their neighbors and likely not outliers.
- Points with higher LOF scores have lower local density compared to their neighbors and are considered outliers.

Step 6: Identify outliers

- Choose a threshold LOF score. Points exceeding this threshold are classified as outliers.





reach = max

reach =

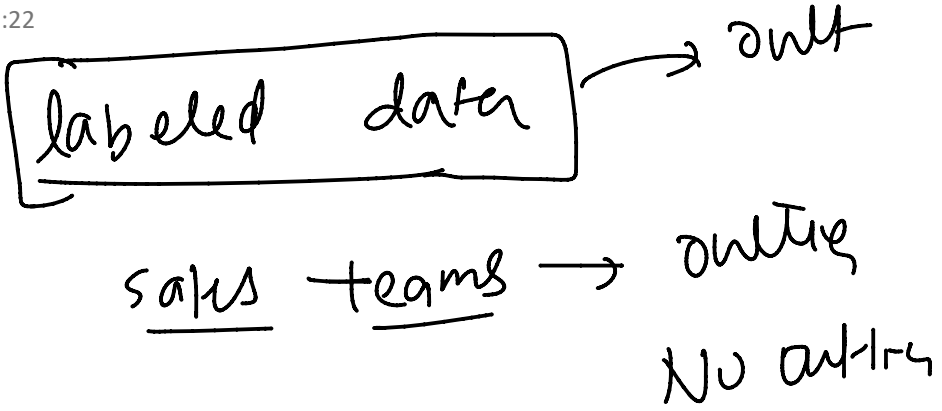
$$reach(p, o) = \max(kd(o), dist(p, o))$$

DBSCAN

07 March 2024 09:42

How to access the accuracy?

12 March 2024 15:22



When to use which algo?

12 March 2024 16:41

1. **Z-Score Technique**:
   - Use when your data is approximately normally distributed.
   - Ideal for univariate data analysis where you want to find outliers based on a single feature.
   - Suitable for quick, preliminary outlier detection to identify extreme values.
2. **IQR and Box Plot Method**:
   - Use for univariate data that may not be normally distributed or when the data is skewed or has outliers.
   - Effective for exploratory data analysis to visually identify outliers.
   - Suitable when you need a robust method that is less affected by extreme values.
3. **Isolation Forest**: → high dim
   - Use for high-dimensional datasets where other algorithms might struggle due to the curse of dimensionality.
   - Ideal when you have a mix of normal and anomalous instances, and the anomalies are 'few and different', which makes them more susceptible to isolation.
   - Suitable for situations where you do not have a clear definition of what constitutes an outlier but want to isolate points that appear to be different.
4. **KNN (k-Nearest Neighbors)**:
   - Use for datasets where the notion of density and local neighborhoods is meaningful.
   - Ideal for moderate-sized datasets; computational complexity can be a concern for very large datasets.
   - Effective when outliers are expected to be points that deviate significantly from their local neighborhoods.
5. **LOF (Local Outlier Factor)**:
   - Use for datasets where you want to detect outliers that are not just globally anomalous but also locally anomalous in varying density regions.
   - Ideal when it's important to consider the relative density of a point's neighborhood, not just the absolute distance to the nearest neighbors.
   - Suitable for datasets where clusters have different densities, and you want to identify outliers relative to these local densities.
6. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**:
   - Use for spatial data or when you want to identify outliers as part of a clustering process.
   - Ideal when you have data that forms natural clusters of varying shapes and sizes, and you want to identify points that do not belong to these clusters.
   - Suitable for datasets where you have some intuition about the density or the distance that defines clusters and noise (outliers).