- Introduction to Anomaly/Novelty/Outlier Detection
- Distribution Based Anomaly Detection
  - RANSAC (Random Sample Consensus)
  - Eliptic Envelope
- Sklearn Implementation
- Isolation forest
- Disadvantages of Isolation Forest
- Sklearn Implementation

What is an Anomaly?

- Something that is not Normal
- Novel (new/never seen Before
- Unique
- Outliers

- IQR ⇒ mean and σ
  - 1 Dimension at a time

mean and σ

outlier points

Distribution Based Outlier Detection

assumption : Data follows Gaussian Distribution

$\downarrow$

$N(\bar{x}, \sigma)$

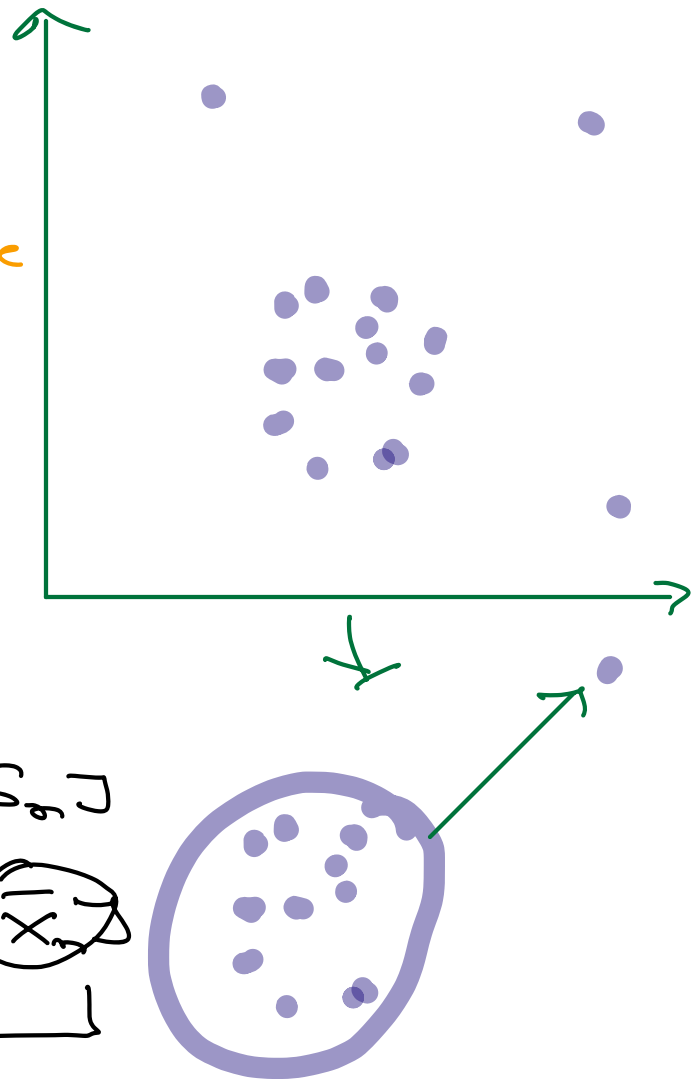
$\downarrow$

Prob-score

$\boxed{\text{prob} < 0.05 \rightarrow \text{outlier}}$

RANSAC (Random Sample Consensus)

- Take Random Sample from Dataset
- Calculate σ and mean of that Sample
- Repeat the above two step N Times



$$[\sigma_1, \sigma_2, \sigma_3 \ldots \sigma_n]$$

$$[\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_n]$$

mean of means

median of means

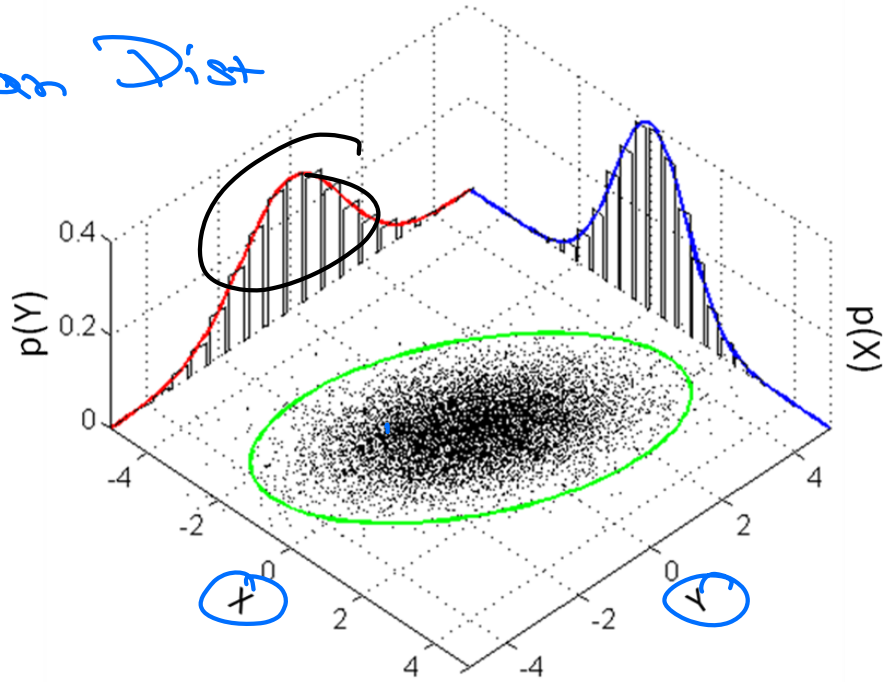
- Robust Estimation of Distribution Params
- By Doing Random Sampling Consensus we can reduce impact of outliers

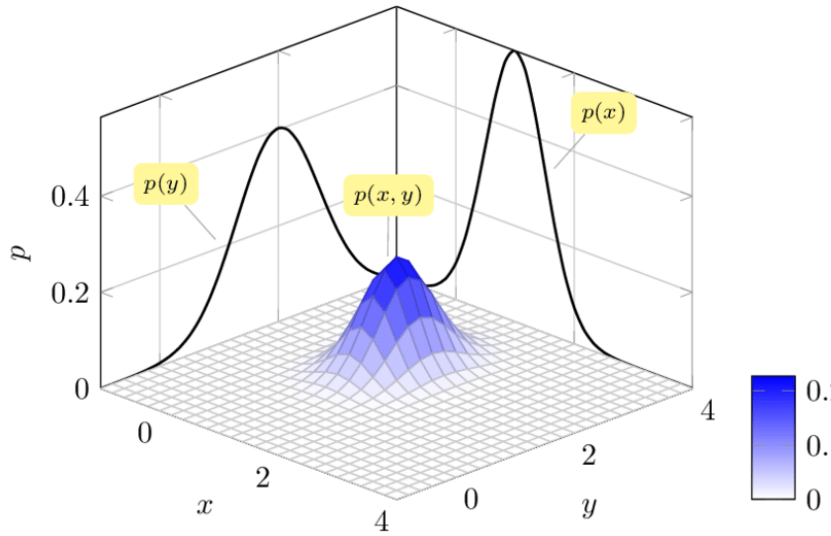
2d Gaussian Dist

- GMM
  - $\bar{x}, \bar{y}$
  - sd_x, sd_y
  - Cov_{xy}

$X \Rightarrow \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \Rightarrow$

Cov_mat $\Rightarrow \begin{bmatrix} \sigma_x^2 & Cov(x,y) \\ Cov(y,x) & \sigma_y^2 \end{bmatrix}$





Elliptic Envelope

Estimating Gaussian Dist' parameters in multi-Dimension

3d $\Rightarrow$ 1×3 $\begin{matrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{matrix}$

3×3 $\begin{bmatrix} \sigma_x^2 & \sigma(x,y) & \sigma(x,z) \\ \sigma(y,x) & \sigma_y^2 & \\ \sigma(z,x) & & \sigma_z \end{bmatrix}$

Elliptic Envelope

⊖ Estimate Params for Multi-Dim

① ⊖ Estimate parameter

↓

② ⊖ Calculate prob of every single point in Dataset

③ How Do find threshold beyond which outliers are assumed?

N = 100

P ⊖ P_1

⋮

P_{100}

↓

Contamination Factor

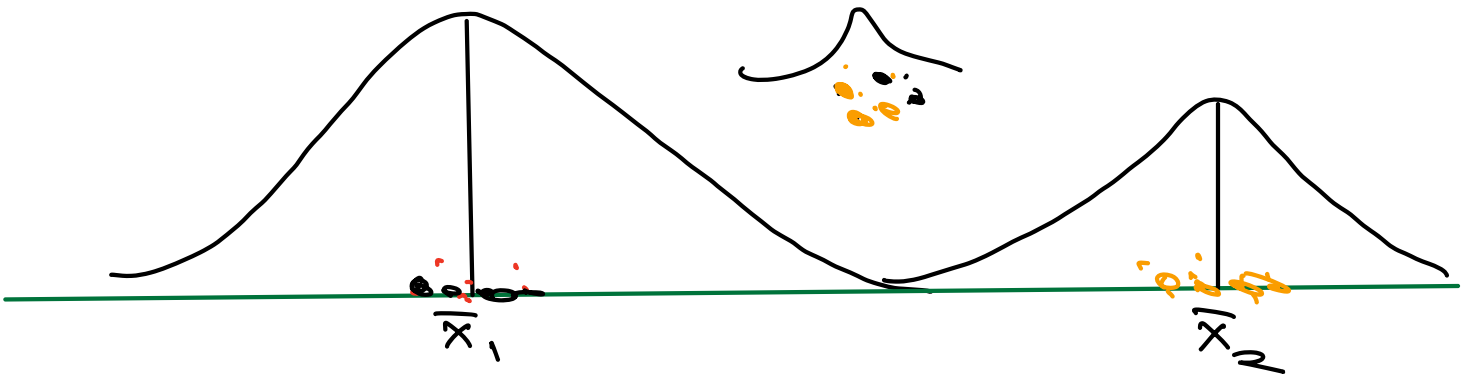
↓

$(0, 0.5] \rightarrow$ 0.10

Sort all point w.r.t prob-score and pick the lowest 10%

- 1000 Data-points 1D.
- $\sigma, \bar{x} \leftarrow$ RANSAC
- assign prob score to each point using PDF
- C.F → 0.05
- Filter 50 point with lowest prob-score → $\boxed{\frac{5}{1000} \times 100 \%}$
- Sort (prob-score, point)[0:50]

- Disadvantages
  - Strict Assumption Gaussian
  - Doesn't work with multi-modal

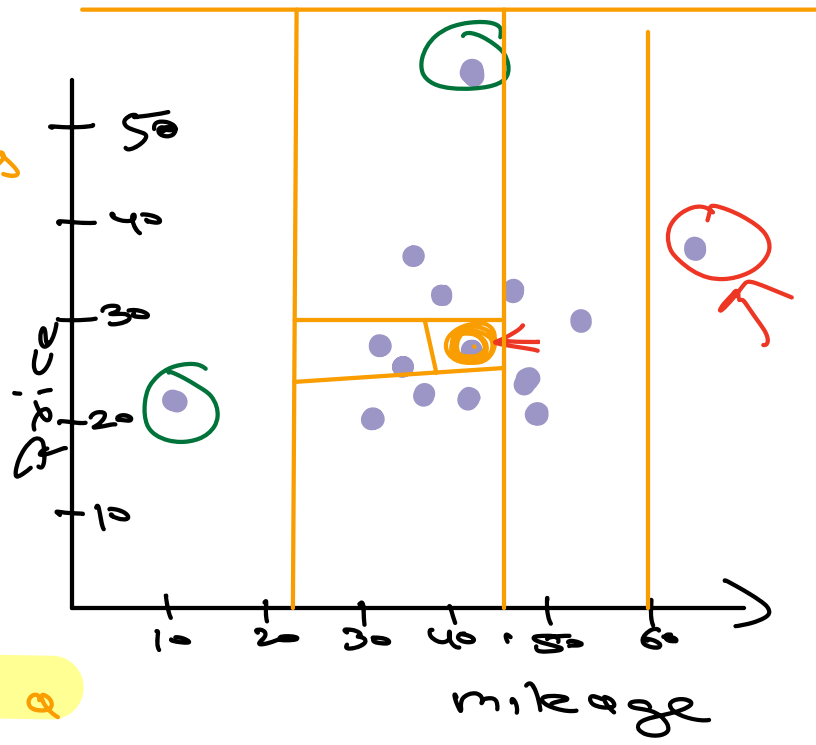


Isolation Forest

- Simple and Elegant
- Works really well for Most Use-Cases

Forest:

Build Many Forests of multiple Trees



To Build Each tree in forest

Step 1:

Randomly pick a feature

Step 2:

Randomly pick a threshold

- Repeat Step 1 and Step 2 until Every single Data-point is isolated

- Outliers points are isolated earlier than inlier points

- Calculate Splits required to isolate every point across all tree

I - tree

$x_1 \rightarrow [2, 1, 3, 2, 1, 4, 1]$

$x_2 \rightarrow [5, 9, 8, 6, 3, 2, 1]$

$x_3 \rightarrow$

x_4

$\vdots$

x_n

avg number of splits across all trees

- Define threshold to filter Data-point with less splits

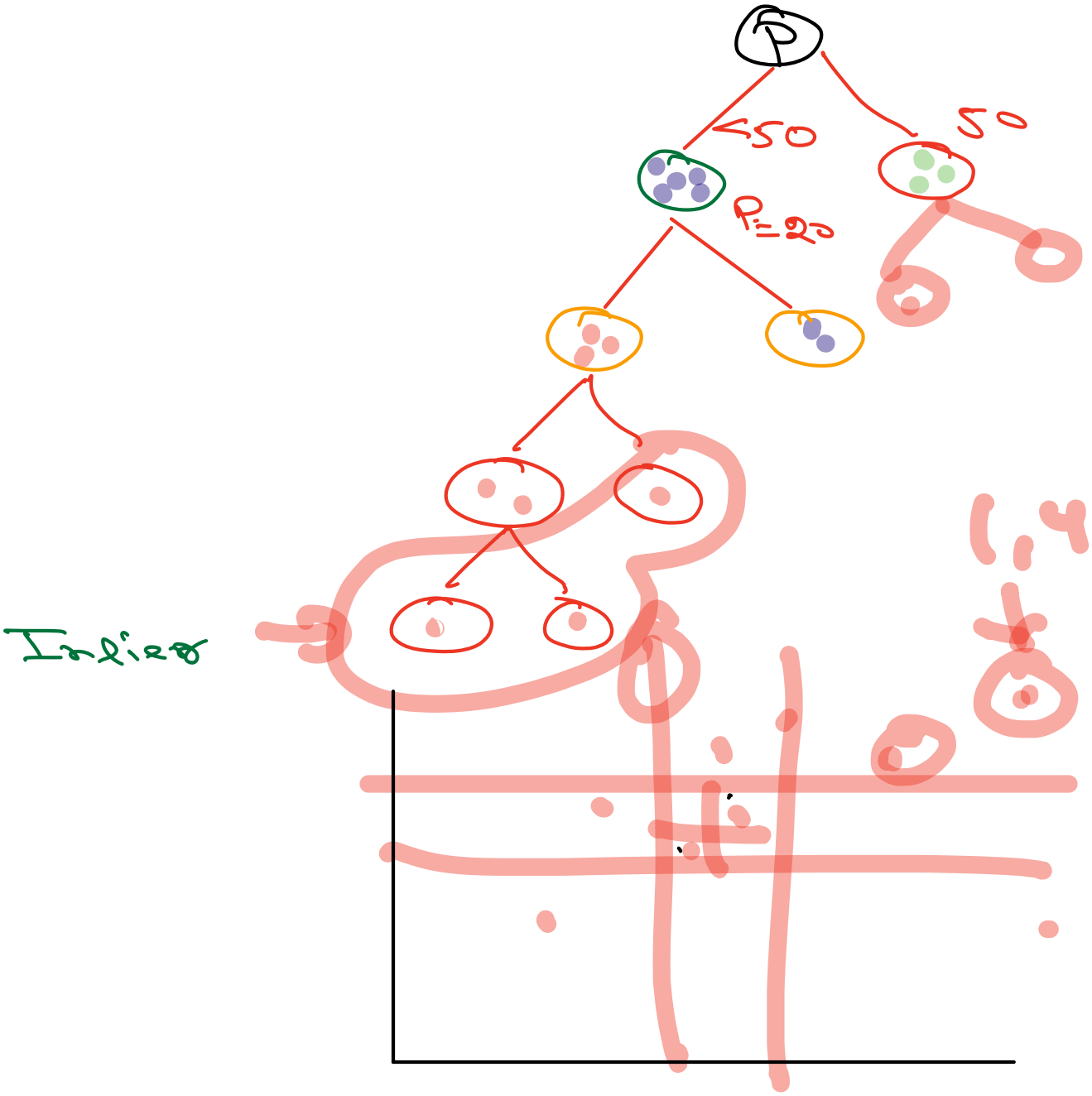
⊙ Contamination-Factor

↓

Sort Avg Splits

and

Select min n%



Disadvantage

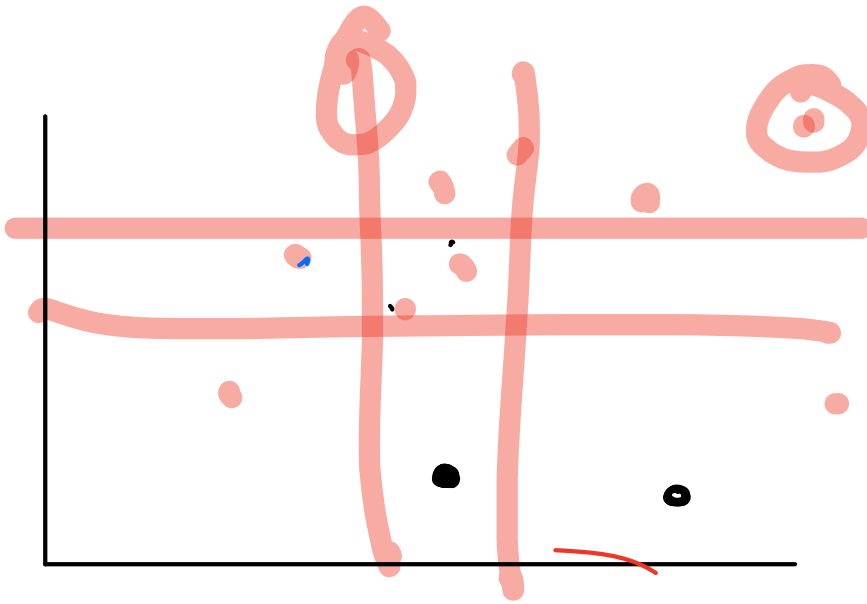
- Isolation of Every D.P is time Consuming

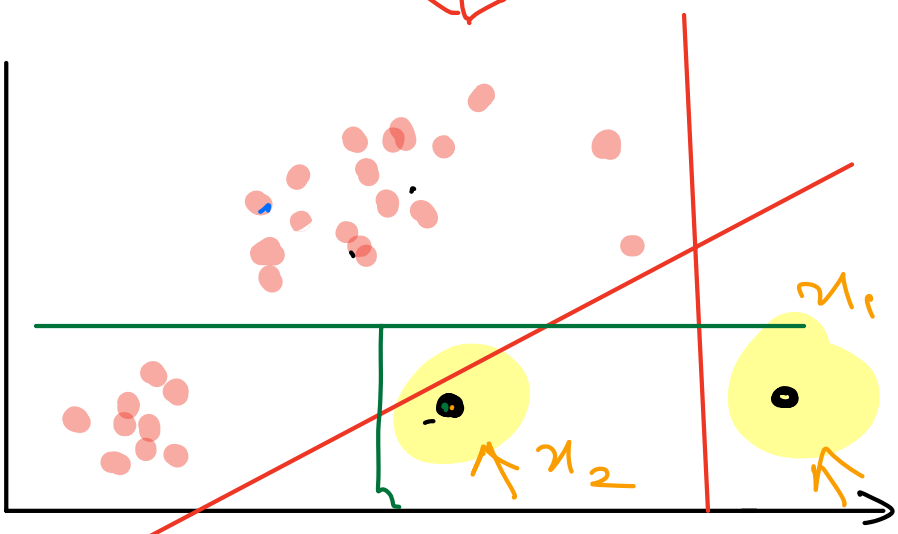
Remedy: $n^{o}_{sample} << n$

5000 d.p → 500 d.p → Tree

5000 d.p → 500 d.p → Tree2

- Biased due to axis parallel splits





✗ not possible with iforest

- One-Class Classifiers
  SVM
- LOF