Anomaly Detection

What is an Anomaly?

- Anomaly is synonymous with an outlier. These terms are often interchanged and may be called Novelty depending on the context.

What's the difference?

- Anomaly means something which is not a part of the normal behavior
- Novelty means something unique, or something that you haven't seen before(novel)

1. Distribution Based

- The simplest way to detect an outlier would be to use distribution parameters (mean and standard deviation).

Problem with this approach:

- While we know the distribution, the parameter estimates of the distribution are often corrupted by the noise/outlier
- Hence, we need to robustly estimate the parameters of the distribution.

2. z-score

Calculate the z-score for each data point: The z-score of a data point (x) is calculated using the formula:

$$z = \frac{x-\mu}{\sigma}$$

Identify outliers:

- you can identify outliers based on a chosen threshold(τ)

- $|z|>\tau$, the point is an outlier

This method is particularly useful when dealing with normally distributed data.

A common threshold τ=2,3
- If ($|z|>2$), the data point is often considered a mild outlier.
- If ($|z|>3$), the data point is often considered a significant outlier.

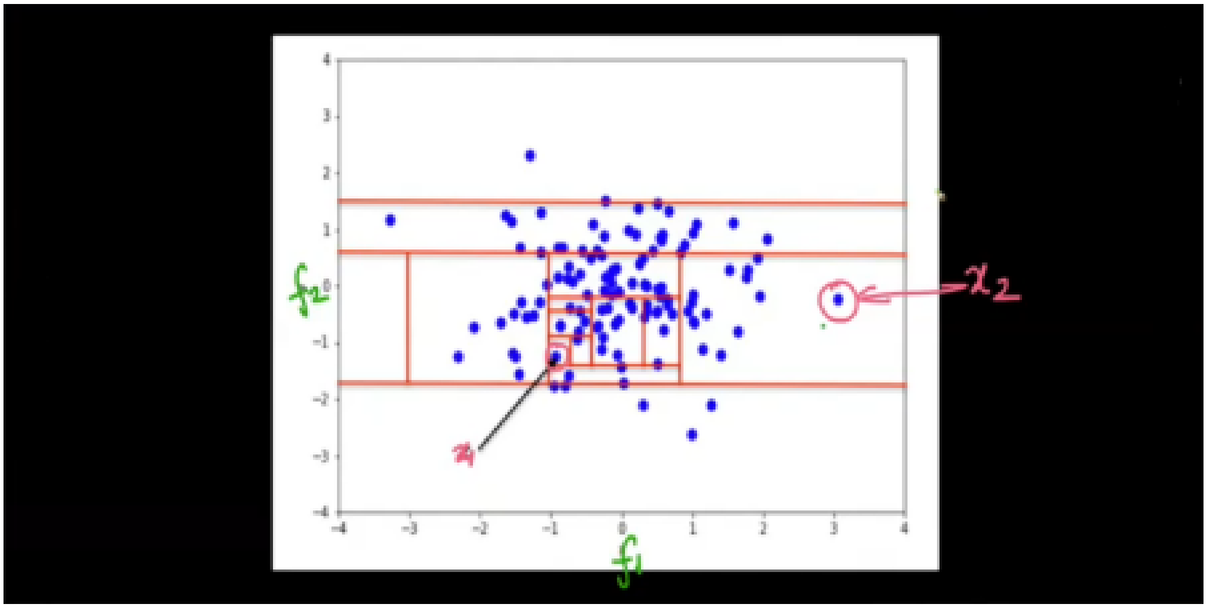
3. DBSCAN

It randomly picks a point and checks its neighborhood using eps
- if num of points ≥ min_num_pts → core point
- if num of points < min_num_pts, but contains at least one core point → Boarder point
- if num of points < min_num_pts, but not a border point → Noise point

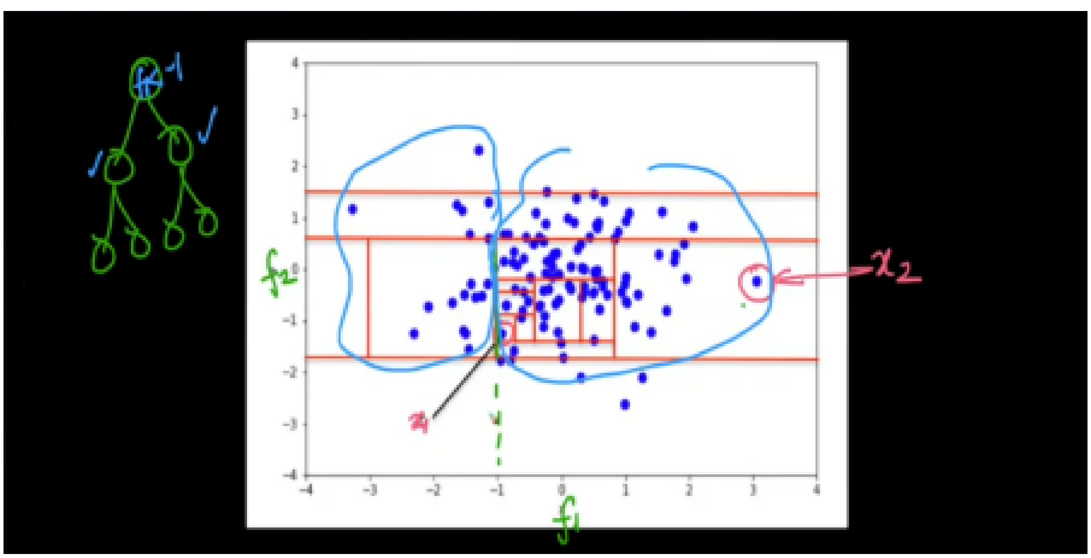
4. Isolation Forests (iForests)

- Consider a dataset D which contains data points x_1, x_2,...., x_n. Just like Random forests, Isolation Forests build many trees.

- Following are the steps involved in Isolation Forest:
    - Build many trees like random forests
    - For each tree:
        - Randomly pick a feature
        - Randomly threshold that features
        - Build each tree until the leaf consists of only one datapoint



- In isolation forests, we are building random trees. So if we pick feature f_1 and put a threshold there will be a vertical bar.
- Similarly, if we pick feature f_2 and put a threshold there will be a horizontal bar.
- For example, if we pick feature f_1 and select threshold as $f_1 < 1$, then our first root node will be based on this condition



- Based on the diagram above,
  - The node containing x_1 will be at more depth.
  - Observe that the point x_1 is in a dense region, and point x_2 is far away

- That is because, to break point x_1 from all the other points, more and more splits will be required and that will increase the depth of the node containing point x_1.

- So, to sum it up, the idea behind Isolation Forest is:
    - On average outliers have lower depth in the random trees
    - On average, inliers have more depth in the random trees

Evaluation of Isolation Forest

- Imagine, we have to build random trees.
    - For each point x_i in the dataset, we can get an average depth.

- We use this average depth to convert it into a metric.

- The basic intuition is that the lesser the average depth, the higher the likelihood is there that it is an outlier

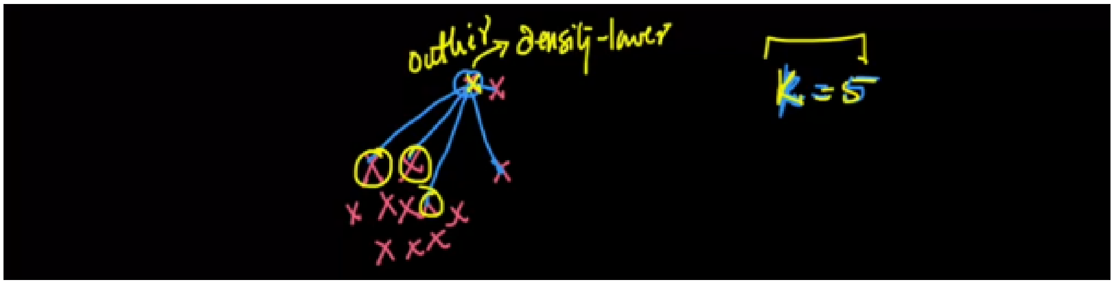
Disadvantages

- They are biased towards axis parallel splits.
    - Because of this, the boundary will not be smoothened.
    - Because the model is biased towards the axis, it will classify the point as an inlier and as an outlier

Local Outlier Factor (LOF)

- Core idea: to compare the density of a point with its neighbors' density
- If the density of a point is less than the density of its neighbors, we flag that point as an outlier

- Imagine a bunch of datapoints as shown below



- We compute the density of a point based on average distance.
- If the average distance between a point and its **K** nearest neighbors is large, it is more likely that the point will be an outlier
- Also, the larger the value of **K**, the more confident are the results.

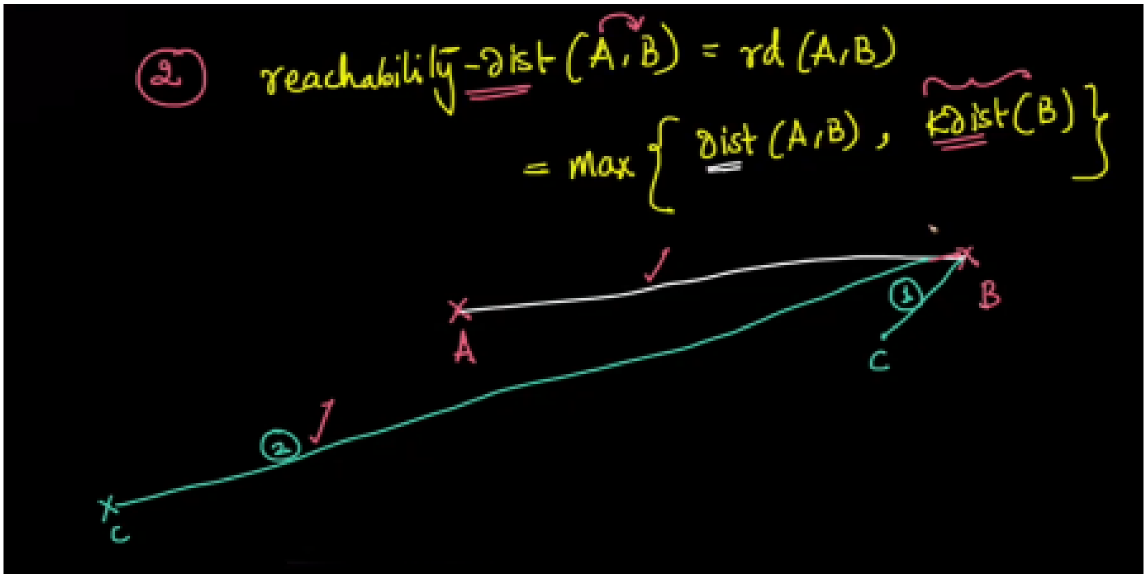
1(a) K-distance

- We define the K-distance of point **A** as the distance of point **A** to its **K**th nearest neighbor
- In general, the larger the value of k-distance is, the farther away the point is from other data points

1(b) Set: Nk (A): It is a set of k-nearest neighbors of point **A**.

2. Reachability distance

- From point **A** to point **B**, we define reachability distance as
  - a maximum of the distance from point **A** to point **B** and the maximum k-distance of point **B**
- Consider point **B** with some **k** nearest neighbors shown in the diagram below.



- There is a possibility that some neighbors might be close(condition 1) and some neighbors might be very far away(condition 2)

- In this case, there is a neighbor of point **B** whose k-distance is greater than the distance between point **A** and **B**, and hence, it is considered as its reachability distance.

3. Local Reachability Density

- It is often represented as **lrd_k(A)**, which tells the local reachability density of point **A.**

- It is defined as the average reachability distance between point **A** and **k** neighbors

$$\text{So, } lrd_k(A) = \frac{\sum_{B \in N_k(A)} rd_k(A,B)\$}{N_k(A)}$$

- The summation in the above equation represents the sum of reachability distances from a point **A** and a set of neighbors **B** as **B** Є **N_k(A)**

- We define the Local Outlier Factor of the point as follows:

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} lrd_k(B)}{|N_k(A)|.ldr_k(A)}$$

- **lrd_k(A)** is the density of point **A**

- The expression $\frac{\sum_{B \in N_k(A)} lrd_k(B)}{|N_k(A)|}$ is the average neighborhood density
- So, LOF of point **A** is nothing but the average neighborhood density(lrd) of point **A** divided by the density of **A**

Interpretation of LOF

- If **LOF(A) = 1**, then we can say that the point has the same density(lrd) as its **k** nearest neighbors
- If **LOF(A) > 1**, then the **k** neighbors of point **A** have a higher density than point **A**.
  - That does not mean point **A** is an outlier. It may or may not be.
  - But if **LOF(A) >>> 1**, then the point is an outlier.
- If **LOF(A) < 1**, then the point has more density than its nearest neighbors.

Disadvantages of LOF

- Finding optimal K
- Finding threshold.
  - If LOF(A) >> 1, what is the threshold??
- Cannot handle high dimensional data efficiently
- High Time Complexity