MACHINE LEARNING

Outlier Detection Methods

Machine Learning

1 What is Outlier?

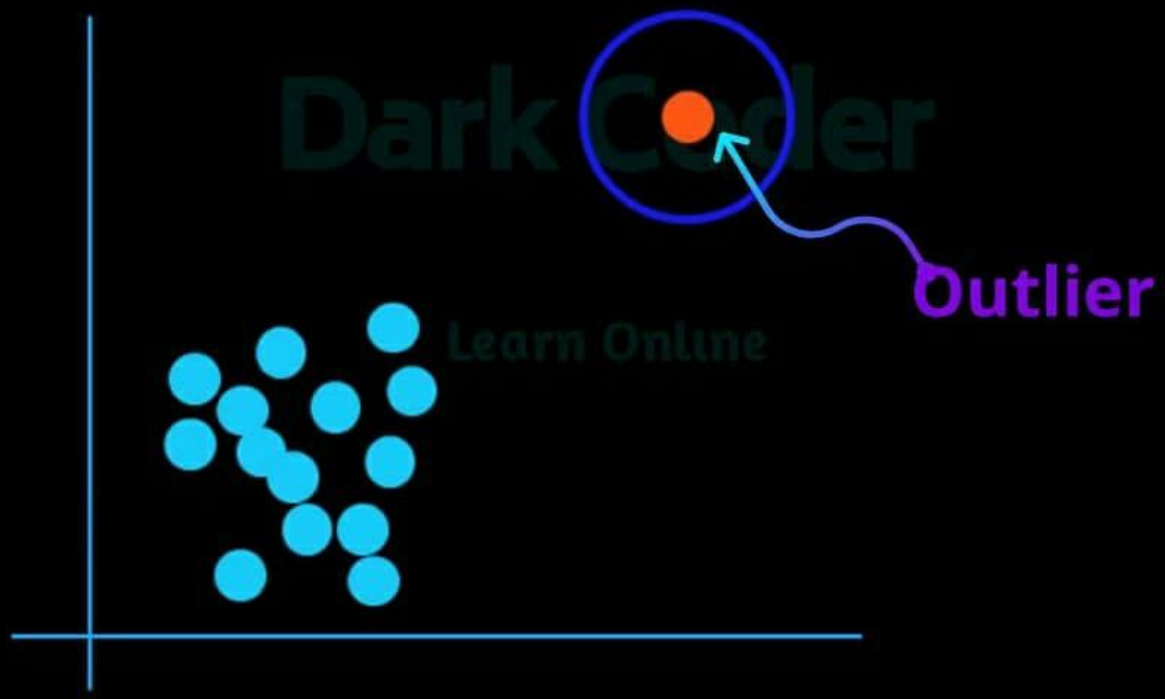
An outlier is any data point which differs greatly from the rest of the observations in a dataset.

- Outliers are of two types:
  a. Uni-variate
  b. Multivariate

- A uni-variate outlier is a data point that consists of extreme values in one variable only
- A multivariate outlier is a combined unusual score on at least two variables.

2 Criteria to identify an outlier

1. Data point that falls outside of 1.5 times of an interquartile range above the 3rd quartile and below the 1st quartile.
2. Data point that falls outside of 3 standard deviations. we can use a z score and if the z score falls outside of 2 standard deviation

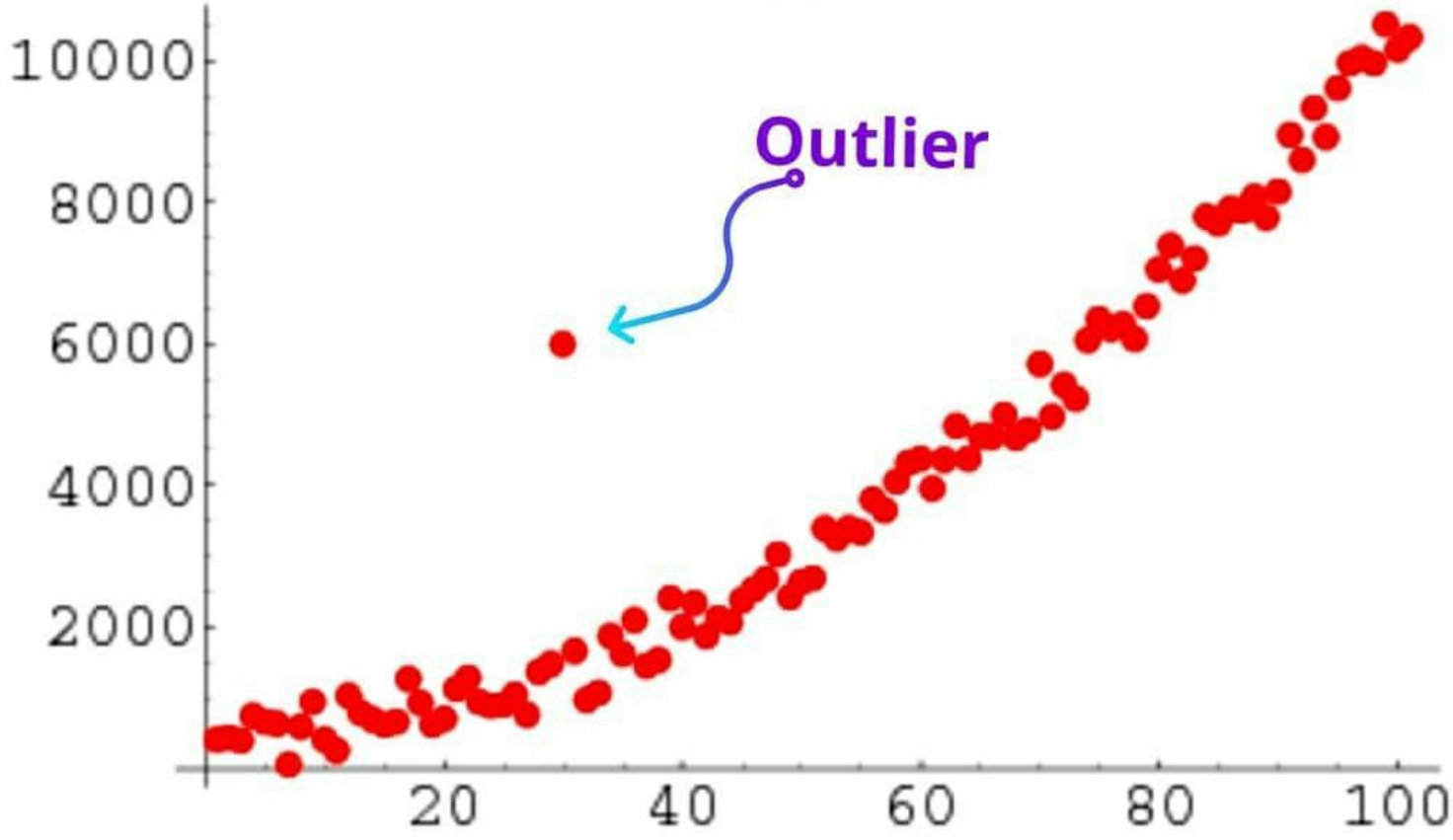


3 Outlier Detection Methods

1. Scatter plots
2. Z score
3. IQR interquartile range

4 Scatter Plot Method

We can see the scatter plot and it shows us if a data point lies outside the overall distribution of the dataset



5 Z - Score Method

Z Score = (Observation — Mean)/Std Deviation

$$z = (X - \mu) / \sigma$$

Steps :-

1. Find the mean and standard deviation of the all the data points
2. Find the z score for each of the data point in the dataset and if the z score is greater than 3 than we can classify that point as an outlier. Any point outside of 3 standard deviations would be an outlier.

6 Z - Score Method


```
import numpy as np
import pandas as pd
outliers=[]
def detect_outlier(data):
    threshold=3
    mean = np.mean(data)
    std =np.std(data)

    for y in data:
        z_score= (y - mean)/std
        if np.abs(z_score) > threshold:
            outliers.append(y)
    return outliers

dataset= [10,12,12,13,12,100,12,14,13,12,10,1,12,10,14,13,15,10]
outlier = detect_outlier(dataset)
print(outlier)
```

[100]

7 IQR interquartile range

- QR tells how spread the middle values are. It can be used to tell when a value is too far from the middle.
- An outlier is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile.

Steps:-

1. Arrange the data in increasing order
2. Calculate first(q1) and third quartile(q3)
3. Find interquartile range (q3-q1)
4. Find lower bound q1*1.5
5. Find upper bound q3*1.5
6. Anything that lies outside of lower and upper bound is an outlier

8 IQR interquartile range

```
import numpy as np
import pandas as pd
outliers=[]
def detect_outlier(data):
    sorted(data)
    q1, q3= np.percentile(data,[25,75])
    iqr = q3 - q1
    lower_bound = q1 -(1.5 * iqr)
    upper_bound = q3 +(1.5 * iqr)
    for i in data:
        if i > lower_bound and i< upper_bound:
            outliers.append(i)
            outliers_points = list(set(data) - set(outliers))
    return outliers_points

dataset= [12,13,12,100,12,14,13,12,10,1,12,10,14,13,15,10]
print(detect_outlier(dataset))
```

[1, 100]

lower_bound is 6.5 and upper bound is 18.5, so anything outside of 6.5 and 18.5 is an outlier.