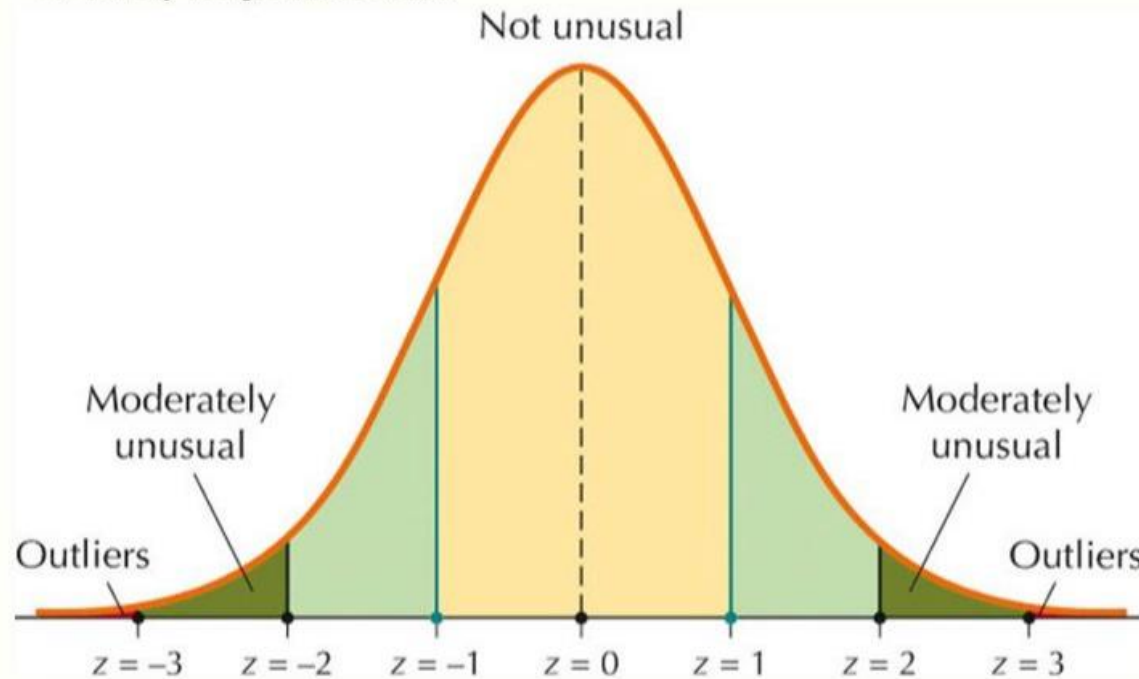


Outliers in Machine Learning

Z score for Outlier Detection

Detecting Outliers with z-Scores

An **outlier** is an extremely large or extremely small data value relative to the rest of the data set. It may represent a data entry error, or it may be genuine data.

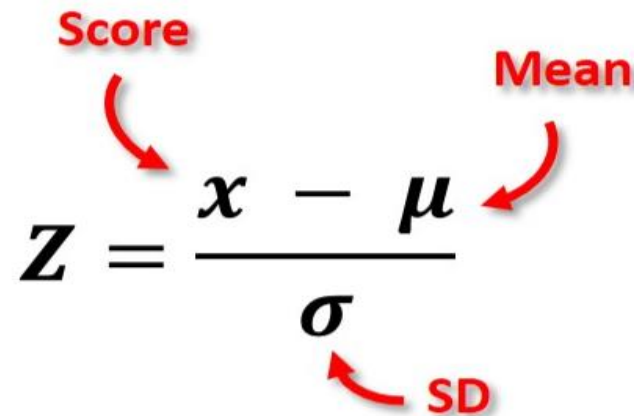


Syed Afroz Ali Data Scientist (Kaggle Grandmaster)

<https://www.kaggle.com/pythonafroz>

<https://www.linkedin.com/in/syed-afroz-70939914/>

Z score is an important concept in statistics. Z score is also called standard score. This score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean.

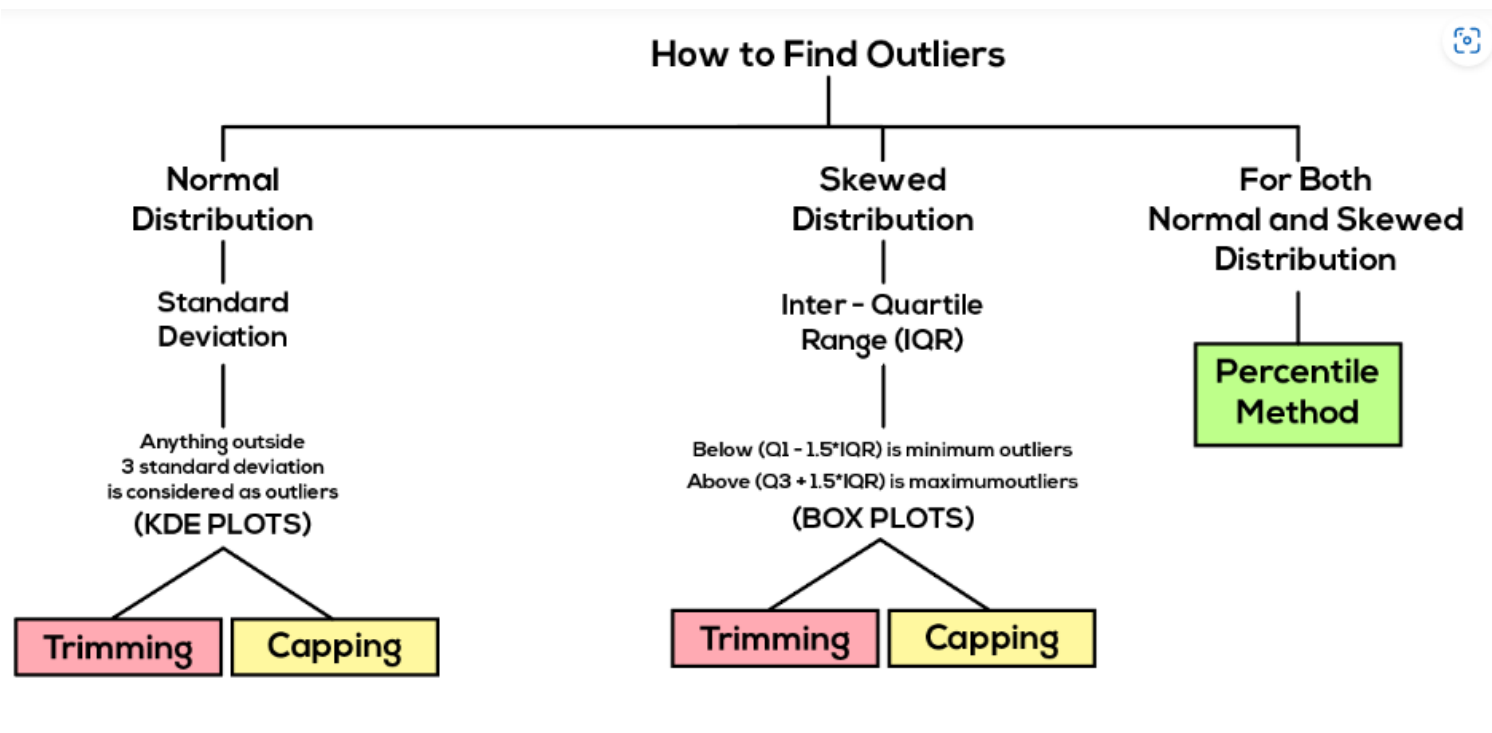
$$Z = \frac{x - \mu}{\sigma}$$


A normal distribution is shown below and it is estimated that 68% of the data points lie between +/- 1 standard deviation. 95% of the data points lay between +/- 2 standard deviation 99.7% of the data points lie between +/- 3 standard deviations

Z score and Outliers: If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier. For example, in a survey, it was asked how many children a person had. Suppose the data obtained from people.

Why do we need "Anomaly Detection"?

There are several reasons why someone would consider deleting few examples of their dataset, even when the dataset is small and we need every bit of information we can get. Outliers can be destructive to our model and our perception of reality. We want our model to predict the most probable label and not be affected by some random value in our dataset. The best way is to remove as little as possible, but make the models robust so that it can ignore or emulate their effect on our prediction.



- Trimming is the method of cutting off or getting rid of the outliers at the end of the dataset. This is easier than all the other methods.
- Capping is setting a limit for the feature and set the value of all the outliers exceeding the limit to the value of the limit. So in the student example, we will set a limit of score and change the score of the 2% student to that. For example, 75 are the max score limit that we set. The score of 2% outlier students will be set to 75.
- Percentile method is equal percentile on both the sides to detect outliers. Once you detect these outliers you can use either trimming or capping to get rid of them.

What is the definition of Outliers and why do we have them?

Outliers are abnormal observations that diverge from other groups. They can have negative effects on our perception of the data and the construction of our model.

We might have outliers because of: Data entry or human errors, damaged or not qualified measurement instruments, data manipulation, dummies made to test detection methods or to add noise, and finally novelties in data.

Even when you generate random numbers from a distribution(e.g. Gaussian), there will be some rare values that stand far away from the mean of all other examples. These are the ones we want to get rid of (or analyse in the real world to know why they are there).

Should we always remove them?

Well interesting question, No! More precisely, not always! They are there for a reason. Why do we have them? Are the measuring tools not working correctly? Are we observing a pattern in the time and place they occur? Maybe there is a bottleneck in our system that is generating them.

Univariate and Multivariate

1. Univariate outliers: When we look at the values in a single feature space.
2. Multivariate outliers: When we look at an n-dimensional space with each dimension representing one feature. In this case because we have too many features to take into account, we cannot simply plot the data and detect which point is away from the normal groups, therefore we use models to do this detection for us.

Cause for outliers

- Data Entry Errors:- Human errors such as errors caused during data collection, recording, or entry can cause outliers in data.
- Data Entry Errors:- Human errors such as errors caused during data collection, recording, or entry can cause outliers in data.
- Measurement Error:- It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty.
- Natural Outlier:- When an outlier is not artificial (due to error), it is a natural outlier. Most of real world data belong to this category.

Different outlier detection techniques

1. Hypothesis Testing
2. Z-score method
3. Robust Z-score
4. I.Q.R method
5. Winsorization method(Percentile Capping)
6. DBSCAN Clustering
7. Isolation Forest
8. Visualizing the data

Syed Afroz Ali Data Scientist (Kaggle Grandmaster)

<https://www.kaggle.com/pythonafroz>

<https://www.linkedin.com/in/syed-afroz-70939914/>

1. Hypothesis testing (grubbs test)

Grubbs' test is defined for the hypothesis:

H_0 : There are no outliers in the data set

H_1 : There is exactly one outlier in the data set

The Grubbs' test statistic is defined as:

$$G_{\text{calculated}} = \frac{\max |X_i - \bar{X}|}{SD}$$

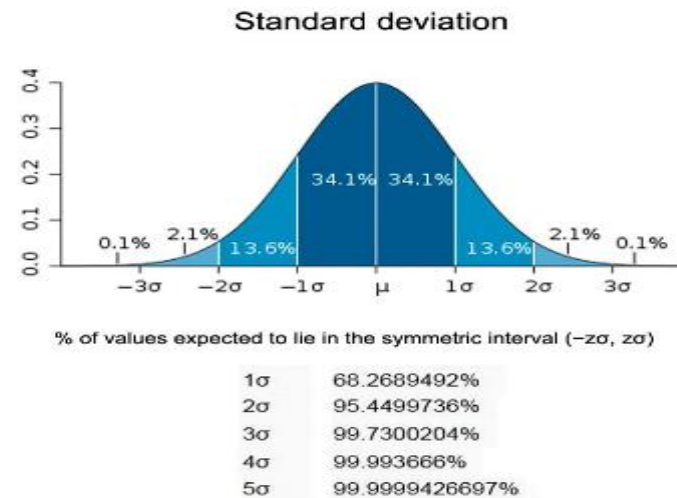
with \bar{X} and SD denoting the sample mean and standard deviation, respectively.

$$G_{\text{critical}} = \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{(t_{\alpha/(2N), N-2})^2}{N-2 + (t_{\alpha/(2N), N-2})^2}}$$

If the calculated value is greater than critical, you can reject the null hypothesis and conclude that one of the values is an outlier

2. Z-score method

Using Z score method, we can find out how many standard deviations value away from the mean.



Syed Afroz Ali Data Scientist (Kaggle Grandmaster)

<https://www.kaggle.com/pythonafroz>

<https://www.linkedin.com/in/syed-afroz-70939914/>

Figure in the left shows area under normal curve and how much area that standard deviation covers.

- * 68% of the data points lie between + or - 1 standard deviation.
- * 95% of the data points lie between + or - 2 standard deviation
- * 99.7% of the data points lie between + or - 3 standard deviation

$$Zscore = \frac{X - Mean}{Standard Deviation}$$

If the z score of a data point is more than 3 (because it cover 99.7% of area), it indicates that the data value is quite different from the other values. It is taken as outliers.

3. Robust Z-score

It is also called as Median absolute deviation method. It is similar to Z-score method with some changes in parameters. Since mean and standard deviations are heavily influenced by outliers, alter to this parameters we use median and absolute deviation from median.

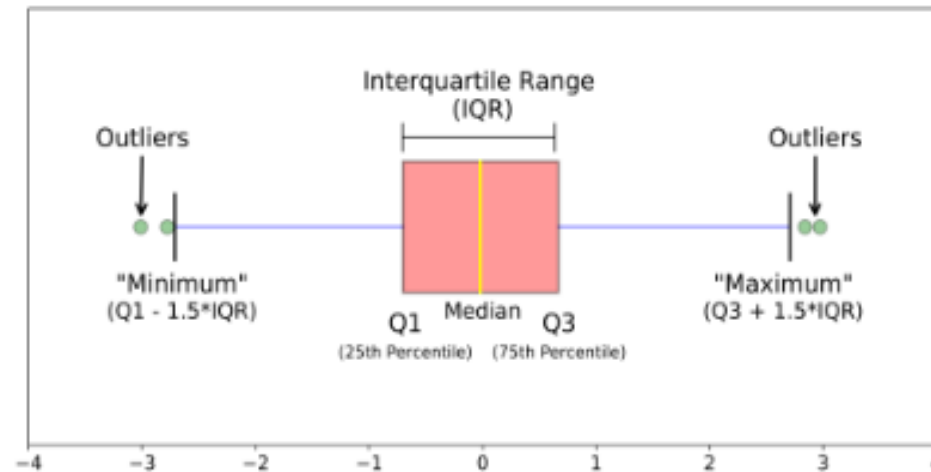
$$R. Z. score = \frac{0.6745 * (X_i - Median)}{MAD}$$

Where $MAD = \text{median}(|X - \text{median}|)$

Suppose x follows a standard normal distribution. The MAD will converge to the median of the half normal distribution, which is the 75% percentile of a normal distribution, and $N(0.75)$

4. IQR method

In this method by using Inter Quartile Range (IQR), we detect outliers. IQR tells us the variation in the data set. Any value, which is beyond the range of $-1.5 \times \text{IQR}$ to $1.5 \times \text{IQR}$ treated as outliers



Number Summary

1. Minimum
2. First Quartile (Q1)
3. Median
4. Third Quartile (Q3)
5. Maximum

Syed Afroz Ali Data Scientist (Kaggle Grandmaster)

<https://www.kaggle.com/pythonafroz>

<https://www.linkedin.com/in/syed-afroz-70939914/>

- Q1 represents the 1st quartile/25th percentile of the data.
- Q2 represents the 2nd quartile/median/50th percentile of the data.
- Q3 represents the 3rd quartile/75th percentile of the data.
- $(Q1 - 1.5 * IQR)$ represent the smallest value in the data set and $(Q3 + 1.5 * IQR)$ represent the largest value in the data set.

Percentiles and Quartiles: {find outliers}

Percentile is a value below which a certain percentage of observation lies.

Dataset : {2,2,3,4,5,5,5,6,7,8,8,8,8,8,9,9,10,11,11,12}

What is the percentile range of 10?

$n = 20$

$$\begin{aligned} \text{Percentile rank of } x &= \frac{(\# \text{ of values below } x)}{x} * 100 \\ &= \frac{16}{20} * 100 = 80\% \end{aligned}$$

What value exists at percentile ranking of 25%?

$$\begin{aligned}
 \text{Value} &= \frac{\text{Percentile}}{100} (n + 1) \\
 &= \frac{25}{100} (21) = 5.25 \text{ index} \\
 &\approx \frac{5+5}{2} = 5
 \end{aligned}$$

Removing Outliers

{1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27}

$$Q1 = \frac{25}{100} (20) = 5\text{th index} = 3$$

$$Q3 = \frac{75}{100} (20) = 15 \text{ index} = 8$$

Lower fence = $Q1 - 1.5(IQR) = -4.5$

Higher fence = $Q3 + 1.5(IQR) = 15.5$

Inter Quartile Range (IQR) = $Q3 - Q1 = 8 - 3 = 5$

Remaining data : {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9}

Minimum: 1

First Quartile: 3

Median: 5

Third Quartile: 8

Maximum: 9

Syed Afroz Ali Data Scientist (Kaggle Grandmaster)

<https://www.kaggle.com/pythonafroz>

<https://www.linkedin.com/in/syed-afroz-70939914/>

5. Winsorization Method (Percentile Capping)

This method is similar to IQR method. If a value exceeds the value of the 99th percentile and below the 1st percentile of given values are treated as outliers.

6. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density based clustering algorithm that divides a dataset into subgroups of high density regions and identifies high density regions cluster as outliers. Here cluster -1 indicates that the cluster contains outlier and rest of clusters have no outliers. This approach is similar to the K-mean clustering. There are two parameters required for DBSCAN. DBSCAN give best result for multivariate outlier detection.

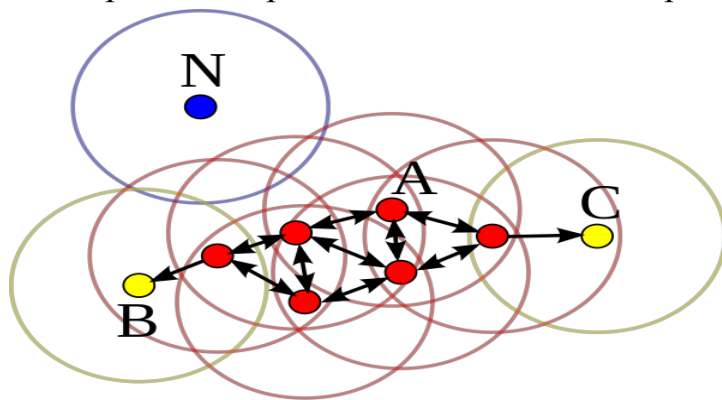
1. Epsilon: a distance parameter that defines the radius to search for nearby neighbours.
2. Minimum amount of points required to form a cluster.

Using epsilon and minPts, we can classify each data point as:

Core point → a point that has at least a minimum number of other points (minPts) within its radius.

Border point → a point is within the radius of a core point but has less than the minimum number of other points (minPts) within its own radius.

Noise point → a point that is neither a core point or a border point



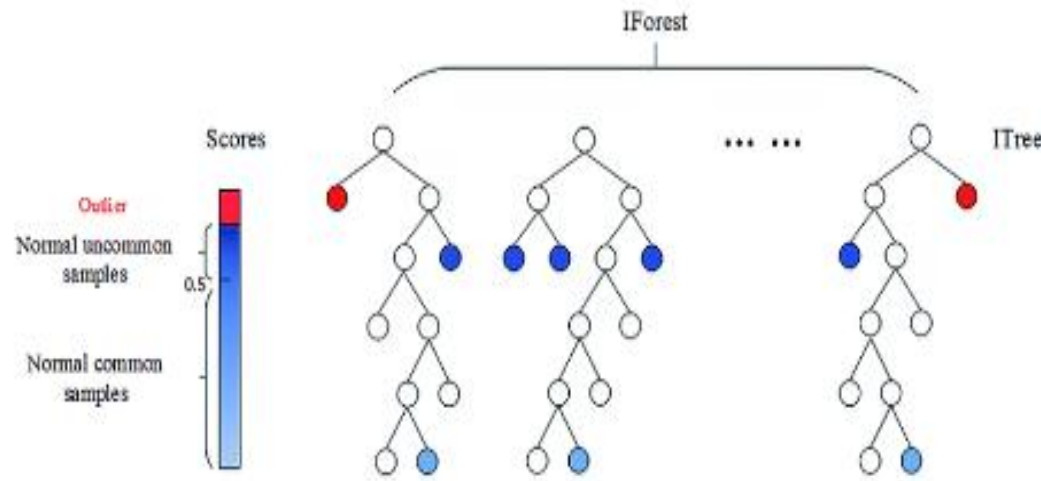
Syed Afroz Ali Data Scientist (Kaggle Grandmaster)

<https://www.kaggle.com/pythonafroz>

<https://www.linkedin.com/in/syed-afroz-70939914/>

7. Isolation Forest

It is a clustering algorithm that belongs to the ensemble decision trees family and is similar in principle to Random Forest.



1. It classifies the data point to outlier and not outliers and works great with very high dimensional data.
2. It works based on decision tree and it isolates the outliers.
3. If the result is -1, it means that this specific data point is an outlier. If the result is 1, then it means that the data point is not an outlier.

8. Visualizing the data

Data visualization is useful for data cleaning, exploring data, detecting outliers and unusual groups, identifying trends and clusters etc. Here the list of data visualization plots to spot the outliers.

1. Box and whisker plot (box plot).
2. Scatter plot.
3. Histogram.
4. Distribution Plot.
5. QQ plot.