Feature Engineering
↓
Feature Transformation

# OUTLIERS

Date

**Agenda:**

1.) What are outliers?
2.) When is outlier dangerous?
3.) Effect of outliers on Machine Learning Algorithms
4.) How to ~~test~~ treat outliers?
5.) How to detect outliers?
6.) Techniques of outlier Detection &

**(1) :** What are outliers?

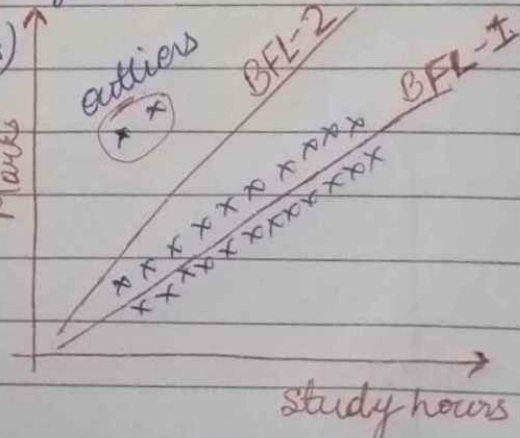
Sharma Ji ka Beta 😊

So, a data point / observation which behaves very different from rest of data points.

**ex:** like there are few Businessman sitting in class & we wanna calculate their salary (avg.) say it comes in lakhs, but if Elon Musk or Ambani is added to that class & now if we calculate average salary, it turns out to be in Crore, so which doesn't looks like avg. salary of all persons, so here Elon & Ambani acts as outliers.

So, here outliers are dangerous.

BFL-1 (in absence of outliers)
BFL-2 (in presence of ")




**(2) :** Dangerous / not?

**Note:** But outliers are not always dangerous, in some cases they could be very useful.

**ex:** Credit Card fault detection, Cancer detection. (in these cases, we are actually looking for outliers.)

} in these cases what we ~~do~~ remove outliers then there won't be any mean left with data.

STUDY BUDDIES

Date

ex: Anomaly detection algorithms

→ So, it depends upon data & business problem whether outliers are important or not important.

→ Its easy to detect outliers, but what to do with it (remove, keep or make changes) depends on business understanding.

(3) :- Impact of outliers on ML-Algorithms :

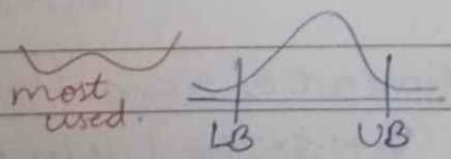
| effects very much ↓ | not much impact ↓ |
|---|---|
| Linear Regression, Logistic Regression, Adaboost, Deep Learning Alg. | Tree based: Decision Trees, Random Forest, XG-Boost etc. |

↳ because, ~~Hot~~ here we calculate weights.

(4) :- How to treat outliers ?

① Trimming
↳ thin data
↳ fast

② Capping (apply limit lower/upper)




③ Missing values (treat outliers as NaN)

④ Discretization (divide data into range like, 0-10, 10-20 & so on...)

→ How to treat outliers depends upon business problem & from where outliers occured in data.

STUDY BUDDIES

#_____

Date

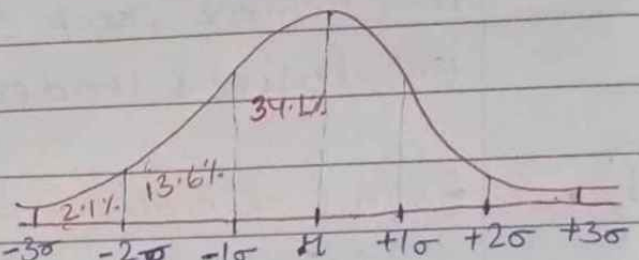
(5) How to detect outliers?

3- most used methods (∵ many methods are there)

① Normal distribution:

assumption: if feature is normally or sort of normally distributed.

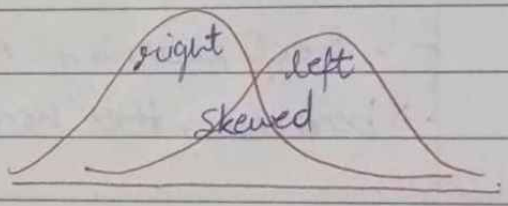
empirical formula:
- Iσ (68.2%) observation
- IIσ (95.3%) ,,
- IIIσ (99.5%) ,,

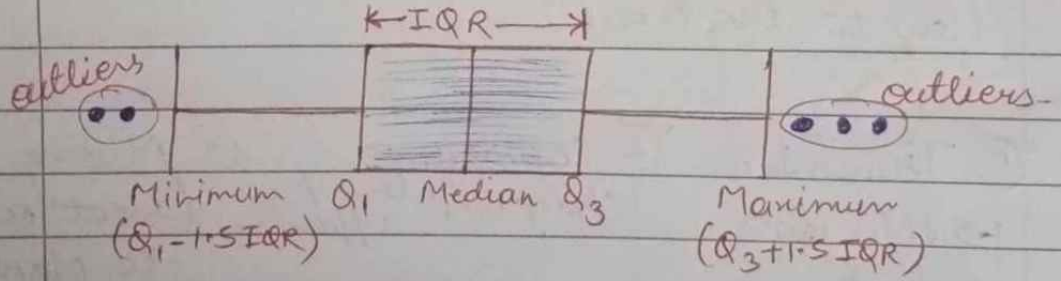



So, any observation out of range $[\mu-3\sigma, \mu+3\sigma]$ are treated as outliers.

② Skewed distribution:

5- point summary (Box-plot)

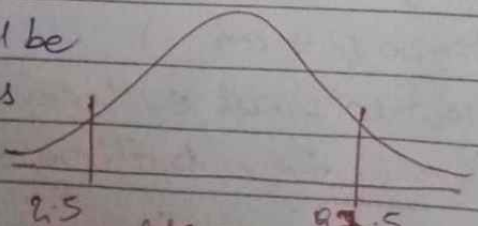



So, any observation out of rang [minimum, maximum] or $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$, they are treated outliers.

③ other distributions:

Any point out of selected percentile range will be treated outlier & this percentile range is selected based on business problem.




STUDY BUDDIES

Date

⑥ Techniques for outlier detection & removal :

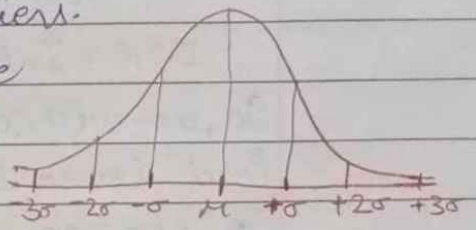
- Z-score treatment
- IQR based filtering
- Percentile
- winsorization

① Z-score treatment (for Normal or sort of Normal dis$^n$)

out of $[\mu-3\sigma, \mu+3\sigma]$ ① are outliers.

This technique is called z-score because,

$$\left\{ z\text{-score} = \frac{x_i - \mu}{\sigma} \right\} \text{---} ①$$




$$\Rightarrow \left\{ x_i = \sigma \times z + \mu \right\} \text{---} ②$$ (for all $x_i$'s we calculate z-score)

compare eq$^n$ ① with ②, so basically we calculate z-score of all values & see if it lies b/w (-3,3) then its not an outlier, otherwise it is.

So, that's how we detect outlier for Normal Dis$^n$.

outlier treatment :

| Trimming | Capping |
|---|---|
| • remove all outliers<br>• but if there are too many outliers then data size will reduce too much, so in that case we use capping. | • let, we got $[\mu-3\sigma, \mu+3\sigma] = [5, 80]$ & we got outliers, 3, 85, 90.<br>then, 3 -> 5 (set to lower bound)<br>85, 90 -> 80 (" " upper ")<br>• Basically, we are putting caps on data, if they go below or above the bounds, then we apply caps on them |

Note: assumption is data should be normally or ~~abo~~ sort of normally distributed.

STUDY BUDDIES

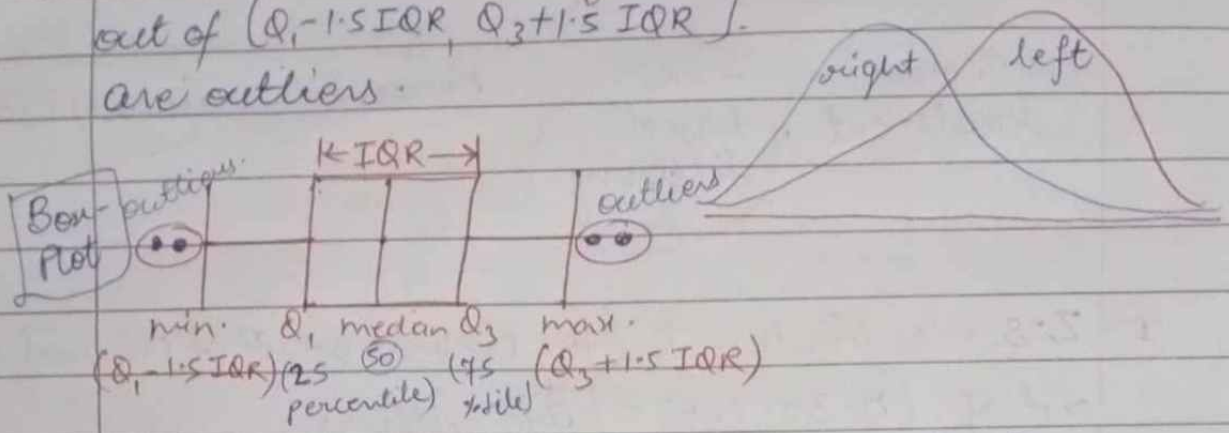
Date

IQR Proximity Rule :

(2) Outlier Detection using IQR : (for skewed distribution)

out of $(Q_1 - 1.5\,IQR,\ Q_3 + 1.5\,IQR)$.

are outliers.




IQR = Inter-Quartile Range = $Q_3 - Q_1$.

So, we will calculate $Q_1$, $Q_3$, then IQR, and then find lower bound $(Q_1 - 1.5\,IQR)$ & upper bound $(Q_3 + 1.5\,IQR)$ & the values out of these bounds will be treated as outliers.

outlier treatment

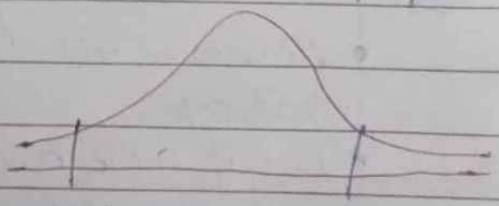
trimming capping

(depending upon number of outliers detected)

(3) Percentile :

(suppose 9 received 99 %tile marks, means 99% people are have less than mine).

Set percentile range based on business understanding.



(∵ Highest marks = 95/100 → 100 percentile
lowest marks = 10 → 0 percentile).

(Percent depends upon total marks, percentile defen is relative quantity means if 9 have 95 %tile means 95% students are below me & 5% are above me).

STUDY BUDDIES

#

Date

Here, we decide %.tile threshold depending on business problem.

As, 5%.tile threshold range is [5%.tile, 95%.tile]
mostly used " " " is [1%.tile, 99%.tile].

After deciding this %.tile range & finding values corresponding to these %.tiles (min, max.) range, we detect outliers (which are out of this range).

outlier treatment

- trimming or removing
- capping (In this technique capping is called winsorization.)