Anomaly Detection

✡ Definition..-

Wikipedia: It is the identification of rare items or events or observations which raise suspicions from the majority of the data.

Towards Data Science: It is the technique of identification of unusual patterns that do not conform to expected behaviour called outliers.

Technical Definition: Anomaly detection is an algorithm which maps the input feature space to an "anomaly" score which is a 1-D space. Then a probabilistic threshold is defined (or trained using supervised methods) to identify anomalous data points.

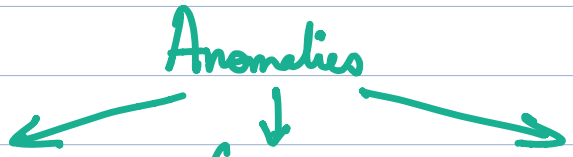
Therefore the 2 steps in anomaly detection are :-

(i) $a = g(x)$

(ii) $f(x) = P(\text{anomaly} \mid g(x))$

** This definition will be analyzed in more detail later as well.

☆ Types of Anomalies:



Point wise

If a data instance is considered as anomalous w.r.t the rest of the dataset.

Contextual

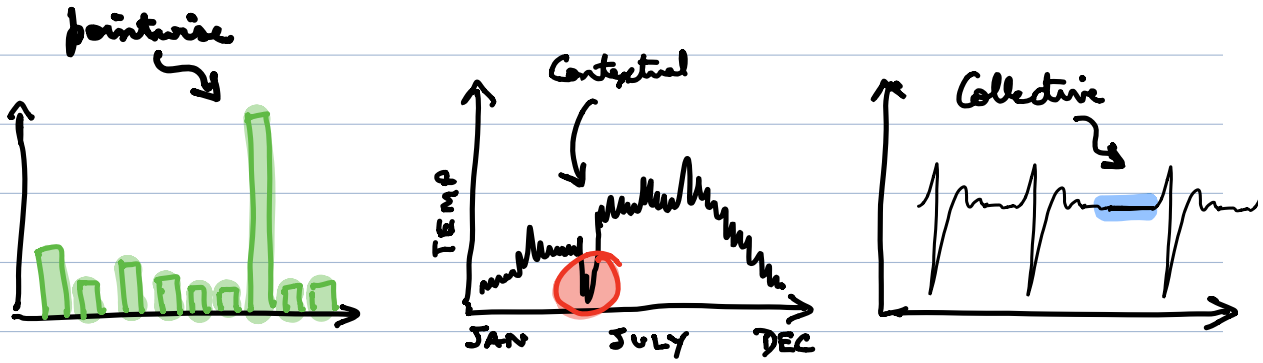
If a data instance is anomalous in a certain Context but not other-wise

Collective

If a collection of data instances are anomalous but not as individuals.

= Ordered

= Unordered.



☆ Methods of anomaly detection :

Density based – DBSCAN, LOF, Isolation forest.

Distance based – K-NN, K-MEANS, Regression hyperplane distance.

Parametric – GMM, Single class SVMs, Extreme value theory

Statical – Statistical tests and variations of Z-score.

These approaches are either supervised or unsupervised based on the respective algo.