
Strive
SCHOOL

/ Outlier

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.

Wikipedia

Detecting Outliers

/ Detecting Outliers (very rare values) are also important. The outliers must be removed from the data so that they do not spoil the models.

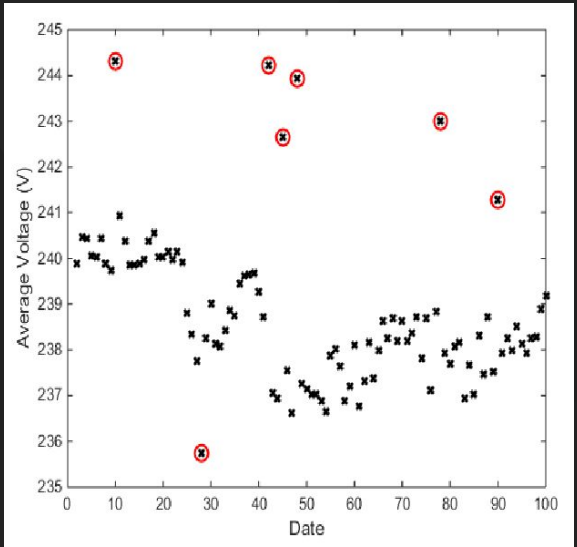
- Manual method of Outlier detection: Make a plot
  - Numerical outilers: Make a plot for showing the distribution
    - Univariate: Plot the distribution (boxplot, stripplot)
    - Bivariate: Make a scatter plot
    - Multivariate: Plot a dimensionality reduction method (TSNE, UMAP)
  - Categorical outliers: my_var.value_counts().plot.pie() , bar(), barh()
- Advanced methods of Outlier detection:
  - Robust covariance
  - One Class SVM
  - Isolation Forest
  - Local Outlier Factor

Detecting Outliers: Manual Methods

/ An image is worth a thousand words. You can think about each point individually and make a decision.

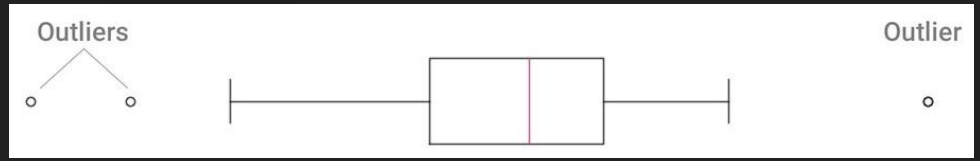
Scatter plot

Usually the variable versus Time

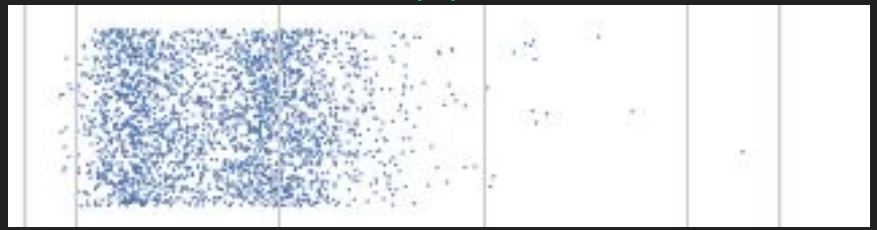


Box plot

Probably the best method



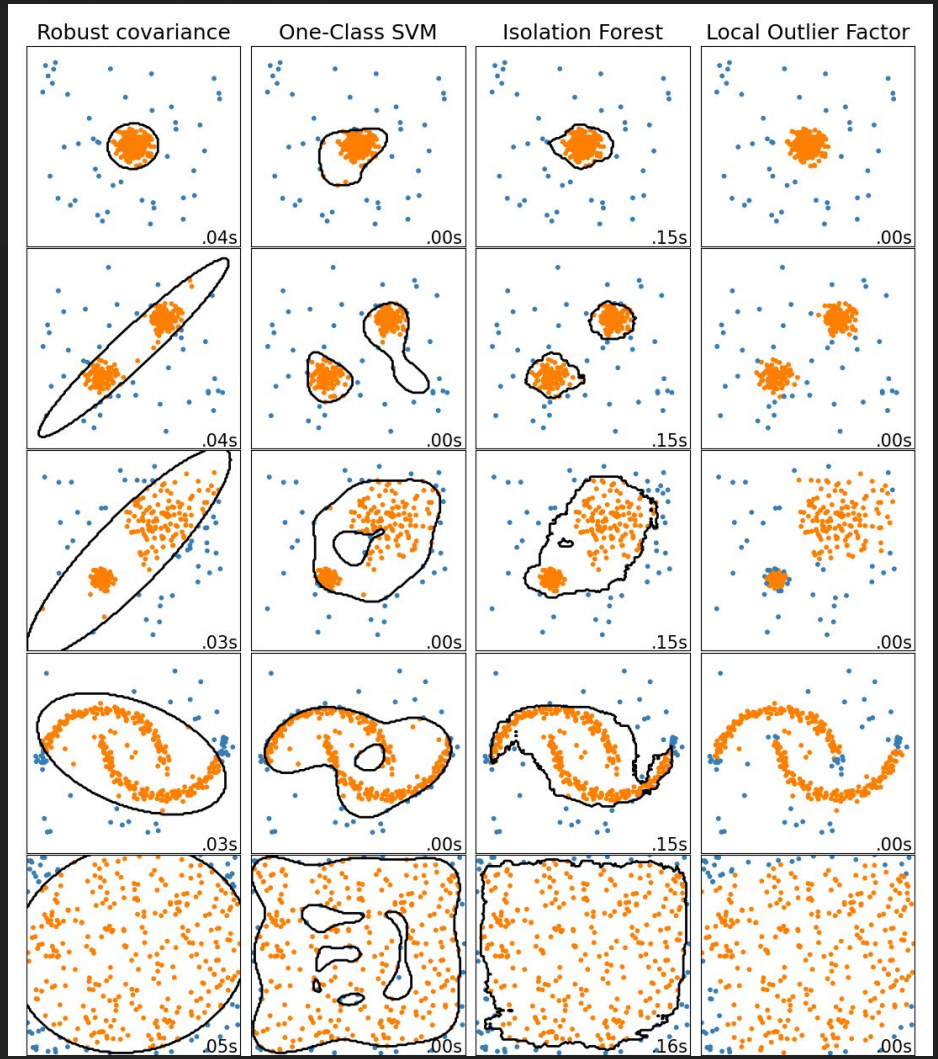
Strip plot



Detecting Outliers: Advanced Methods

(not recommended)

- Robust covariance:
  - sklearn.covariance.EllipticEnvelope
- One Class SVM
  - sklearn.svm.OneClassSVM
- Isolation Forest:
  - sklearn.ensemble.IsolationForest
- Local Outlier Factor:
  - sklearn.neighbors.LocalOutlierFactor



Typographical errors (Typos)

/ At data entry is common to introduce errors. This errors are called typos. Detect them and correct them is very important.

- Someone born in 2200 → Probably was born in 2020
- Someone born in Sapin → Probably was born in Spain

/ fuzzywuzzy is a package to find similar strings that usually are typos and errors when the data was written.

Handling Outliers

/ Once they have been detected, we have to handling them. Common Handling Outliers methods are:

- Remove them: Usually the best option (if the value is strange)
- Correct them: Best option if outlier is a “typo”
  - The max limit
  - The mean
  - Etc.
  - Other imputation method like missing imputation
- Oversample the outlier (make copies of the row)
  - You can introduce a noise for the ML model to not overfit.
    - np.random.normal(age, scale=5.0, size=1000)

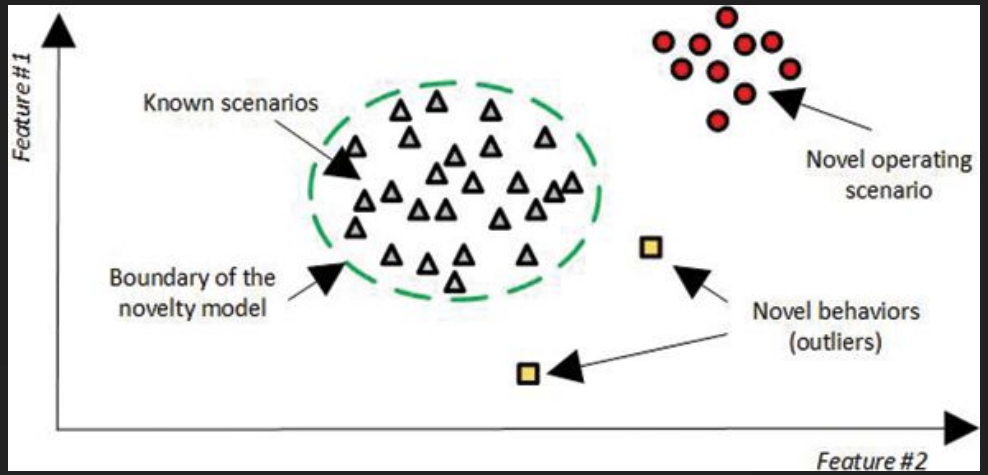
Outlier in dataset VS Outlier in the future

Outlier detection

The training data contains outliers which we are interested in detecting them.

Novelty detection

The training data does not contains outliers and we are interested in detecting whether a new observation is an outlier.



/ Q&A

What are your doubts?


Strive
SCHOOL