Clustering and Fitting
Ismail

Purpose of clustering and fitting in Data science

In Data Science, we can use clustering analysis to identify what categories the data points fall into when we use a clustering algorithm, which might provide some insightful information from our data. From the usage of fitting types which are Under fitting and Over fitting, one can get to know about the accuracy of data. If the data is accurate or with less error than the fitting modal will be the best fit. So, this way clustering and fitting plays important role in Data Science

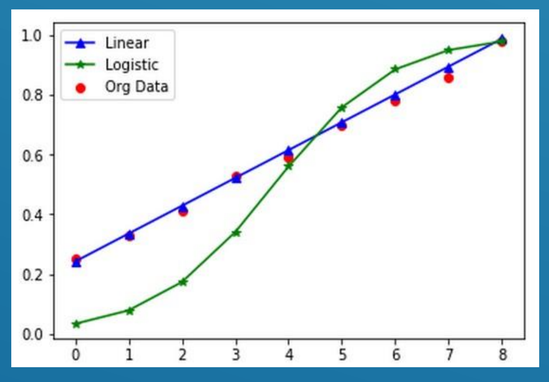
Clustering

The purpose of clustering is to divide a population or set of data points into groups so that the data points within each group are similar to and different from the data points within other groups. The quality of clustering is depending on the similarity and dissimilarity of the objects. There are different types of clustering techniques.

Fig: 1.1 : Clustering

Fitting

We frequently have a dataset with data that generally follow a path, but because each data point has a standard deviation, they are dispersed along the line of best fit. Finding the curve that minimizes the vertical (y-axis) deviation of a point from the curve is what is commonly called "fitting". There are different types of fitting such as overfitting and Under fitting.



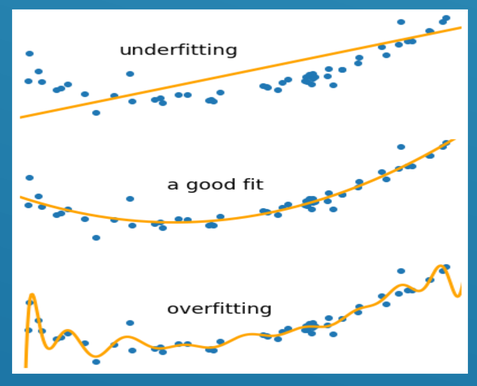


Fig: 1.2 : Fitting

Methodology

To illustrate the clustering, we selected a 30-year Sri Lanka Co2 Emission dataset from a global database. Then I drew the graph below from Python's pandas library.

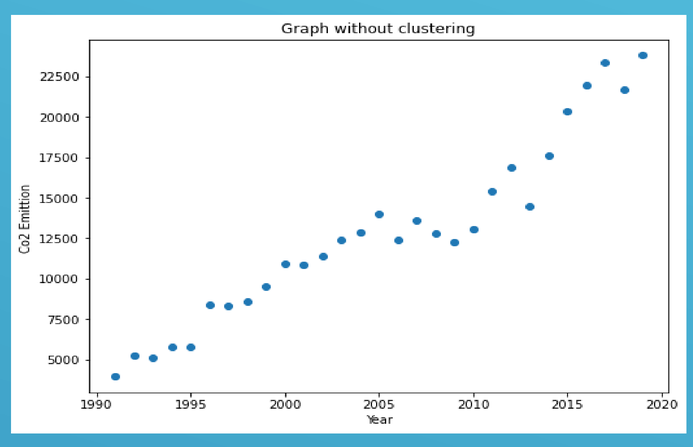


Fig: 1.3 : Scatter graph

This Figure is a Co2 Emition-Year Graph of the Sri Lanka for the time span of 30 Years. In which there is total Co2 Emittion is plotted with respect to the year

Once you've drawn your scatterplot, you can easily draw cluster-based charts. To illustrate this combination, we use k-means clustering. You can draw graphs from the pandas library. Here we have selected four clusters. The data are arranged in different clusters according to cluster focal points. The black diamonds in the graphic are cluster focal points.

Now we need to find the best fit to the data for line fit, polynomial fit, or sine fit. Observing the graph shows that the line is a good choice.

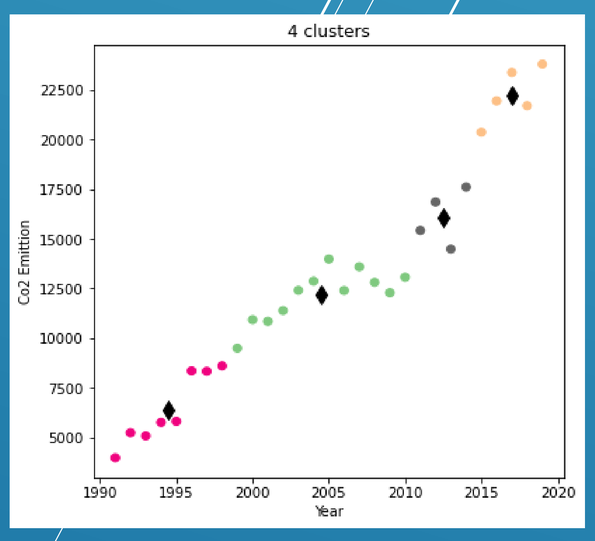


Fig: 1.4 : Clustered graph

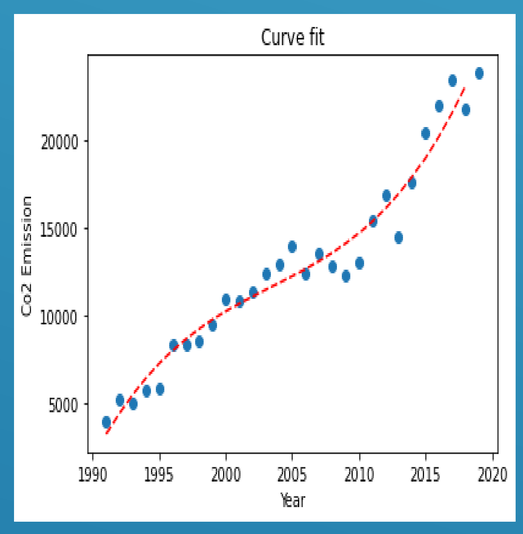


Fig: 1.5 : line fit graph

Conclusion

In summary, clustering is a method of grouping similar types of data from a dataset, which helps data science find patterns and predict the future of a particular area. Fitting shows how accurate the data are. To conclude this, Co2 Emission is increasing continuously.

Github Link: https://github.com/is2295/ADS-Assignment-3-Clustering-and-fitting-