Clustering and Fitting

Harvi Gandhi

Introduction

Clustering is the process of creating the groups in which objects within a group be like one another and different from the objects in the other group. The quality of clustering is depending on the similarity and dissimilarity of the objects. There are different types of clustering techniques. And finding the curve that minimizes a point's vertical (y-axis) deviation from the curve is what is commonly meant by "fitting."

Clustering

In essence, it is a kind of unsupervised learning technique. It is typically used as a method to identify the groups, generative qualities, and significant structures that are inherent in a set of instances.

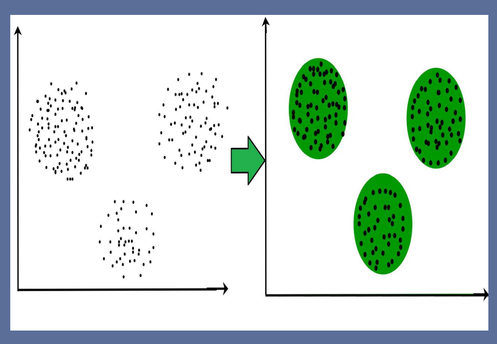


Fig : 1.1 :Example of Clustering

The objective of clustering is to divide the population or set of data points into a number of groups so that the data points within each group are more similar to one another and different from the data points within the other groups.

Fitting

Find the model that best captures the data given a dataset made up of a collection of points is called as 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.

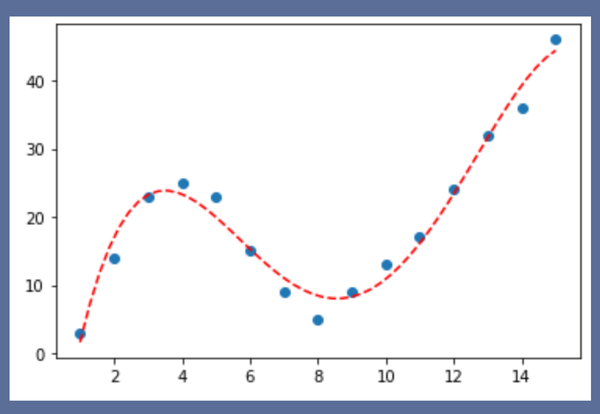
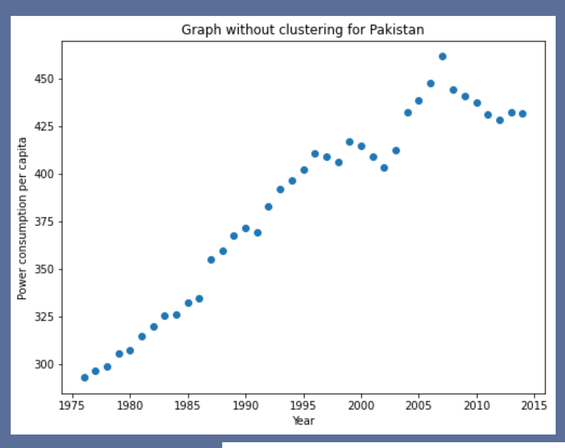


Fig : 1.2 :Example of curve fit

Finding the values for a dataset through which a specific set of explanatory variables may accurately describe another variable is the aim of curve-fitting.

Methodology

To explain clustering, I have chosen the power consumption data set of Pakistan and India for the time span of 25 years from world data bank.



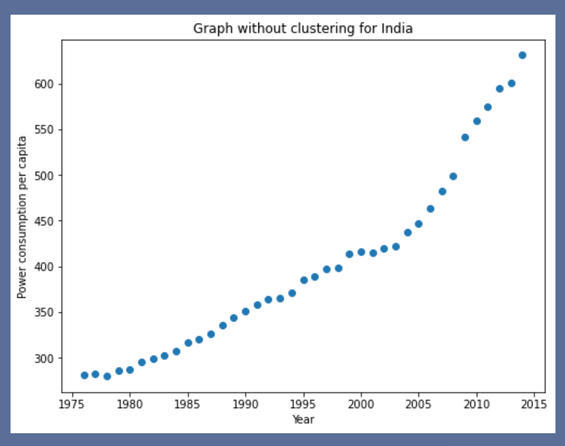
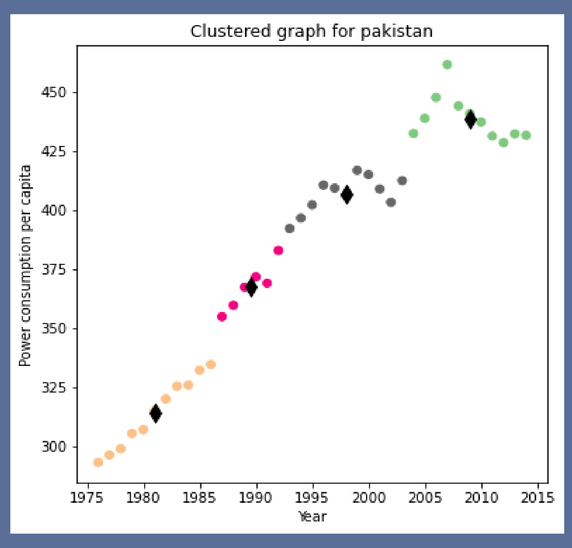


Fig : 1.3 : Scatter graph for Pakistan and India

These above graph is a scatter graph of a power consumption per capita for Pakistan and India between 1975 to 2015.

After plotting the scatter plot, we can easily plot clustering based graph. To plot cluster graph, there are different types of technique. To explain this combination, we are using k-means clustering. From using pandas library, we can plot the graph.. Here, we have chosen 4 clusters.



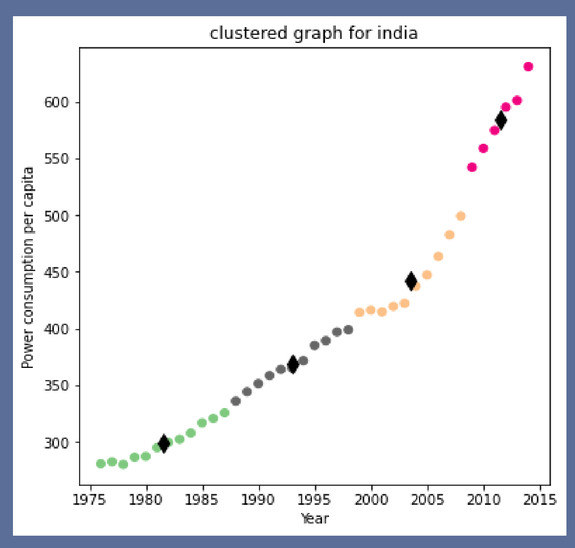
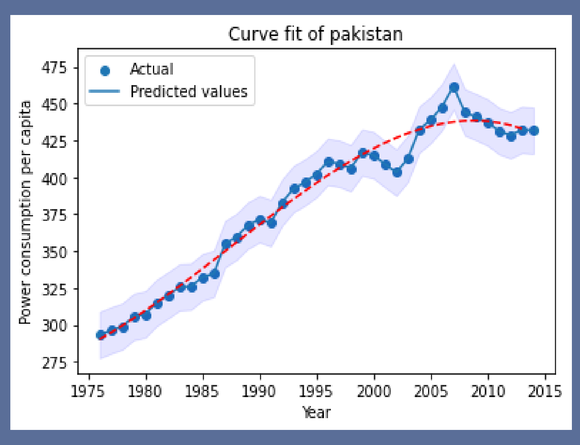


Fig : 1.4 : Clustered graph for Pakistan and India

According to centroids of the cluster, the data are placed into the different clusters. The black diamond label in the graph is the centroids of the clusters.

Now, we have to find the best fit for the data which could be the line fit or polynomial fit or sin fit to find the accuracy of data. From observing the graph, its visible that polynomial would be great fit to choose.



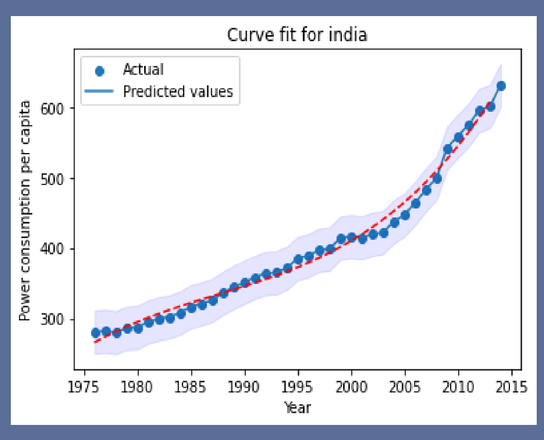


Fig : 1.5 : Curve fit graph

The above graph shows the best fit for particular graph in which blue dots are the actual value, red dotted line is the best predicted value and blue opaque is **confidence range**.

Conclusion

To conclude this, clustering is the method to group similar kind of data from the data set which is useful for in the data science to find the patterns or to predict the future about particular field and fitting will show us that how accurate the data is. From the above combination of tow countries, we can predict the future that the power consumption will increase in the future.

GitHub Link:
https://github.com/Harvi05/ADSPoster/blob/main/clusteringFitting.py