Clustering

Q1: Beer grouping use case

Company X would like to enter the market with a new beer brand.

Before it decides the kind of beer it will launch, it must understand what kinds of products already exist in the market and what kinds of segments the products address.

To understand the segments, the company collects specifications of few samples of beer brands (as given in the Beer Dataset). Please refer to Beer Grouping dataset on GitHub -

https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment10/Beer_Grouping.xlsx

Assuming there are 3 segments, use K-Means Algorithm to create three clusters using the Beer Dataset. Once the clusters are finalized, interpret the clusters.

Solution:

Please refer to the following Jupyter notebook file for details.

https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment10/AAI_Clust_Assignment_Q1.ipynb

Q2: Distance measures on universities dataset

Please refer to the Universities.xlsx on GitHub: (https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment10/Universities.xlsx).

a) Find out the Euclidean distance between Stanford and MIT.
b) Find out the Manhattan distance between Stanford and MIT.
c) What difference you can convey for Manhattan distance w.r.t. Euclidean distance measure?

Solution:

a) Euclidean distance = 1580.17
b) Manhattan distance = 1618.00
c) Manhattan distance is always greater than Euclidean distance.

We can use Python to easily compute above distances. You can also refer to the excel for simple computation based on the formula for these distance measures (https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment10/AAI_Clust_Assignment_Q2.xlsx).

Q3: Distance measures comparison

What is the Euclidean distance between two points (2,4,6) and (3,8)?
What is the Manhattan distance in the above case?

Solution:

Ideally, we can't compute these distance measures if there are different dimensions. So, in this case, we should not assume and compute.

Q4: Customer Segmentation use case

One of the large smart phone manufacturing giants is analyzing it's customer satisfaction and loyalty data. Here is a sample they have captured from a geography for certain period for analysis purposes. As a Data Scientist, you are tasked to analyze the segmentation pattern and recommend your inferences.

The dataset is here on GitHub - https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment10/Satisfaction_Loyalty.csv

Dataset has two features captured here for analysis consideration - Satisfactory index and Loyalty index of customers.

Use clustering methods and provide your inference.

Solution:

Please refer to the following Jupyter notebook file for details.

https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment10/AAI_Clust_Assignment_Q4.ipynb