**Assignment #2:** Clustering with GMM & Hierarchical Clustering

**Team Name:** Codebusters

**Team Members:**

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Student ID** | **Contributions** |
| Anushri Srinath Aithal | 012506897 | Hierarchical Clustering - Agglomerative Clustering |
| Harini Balakrishnan | 010830755 | Gaussian Mixture Model Clustering |
| Ravi Katta | 012127011 | Gaussian Mixture Model Ellipse &  Stretched-out pattern |
| Sunder Thyagarajan | 011528062 | Hierarchical Clustering - Dendrogram |

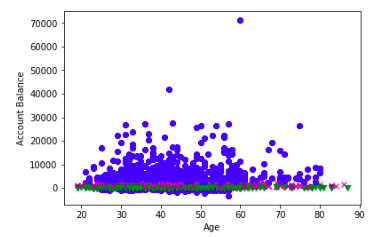
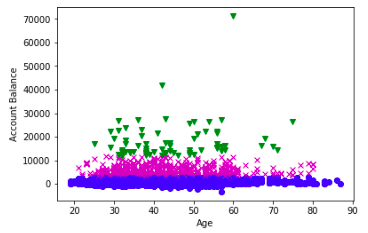
**Dataset:**<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

**Objective:**

The objective of this assignment is to understand the Machine Learning life-cycle using the bank marketing data set and apply Gaussian Mixture Model Algorithm. Compare and contrast with K-means clustering that we achieved with the same data set in Assignment 1. The bank marketing data helps to analyze the age based account balance clustered which helps to get the clustering of all customers who are eligible to take a loan. The goal of clustering thus identifies the customer groups that are eligible for taking loan and the bank can send promotions.

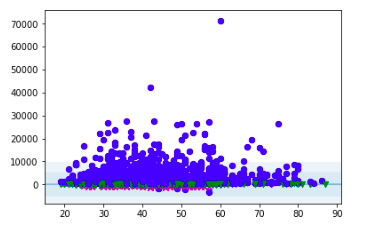
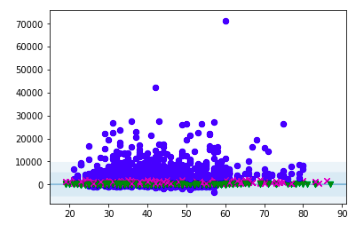
**GMM Clustering:**

Gaussian Mixture Model Algorithm is a powerful tool to estimate the simple clustering better. It is a better algorithm that can be applied to analyze the real-world data. A Gaussian mixture model (GMM) finds a mixture of multidimensional probability distribution that provides best model for any input dataset. By applying GMM on our existing k-means clusters we were able to visualize the changes as shown in fig1.1 and fig 1.2.



**Fig: 1.1  *k*-means Fig: 1.2 Gaussian Mixture Model (GMM)**

It is possible to visualize the uncertainty by making each point promotional to the certainty of its prediction between the cluster’s boundaries. In GMM, the result of each cluster is associated with a hard-edge sphere. In this assignment we have applied GMM on the dataset that has 3 clusters. By creating a function to plot the GMM algorithm with Ellipse patches, we were able to generate ‘spherical’ sphere as shown in the fig 1.3 Choosing the covariance type controls the degree of freedom in the shape of each cluster. We set the covariance\_type=‘spherical’ which constraints the shape of our cluster with all dimensions. There are three clustering types like full, spherical and diag. In order to allow full covariance of the model using GMM approach is to fit the stretched dataset. The cluster will be even more oblong with stretched-out pattern.



**Fig:1.3 GMM with ‘spherical’ covariance Fig: 1.4 GMM**  **stretched-out pattern**

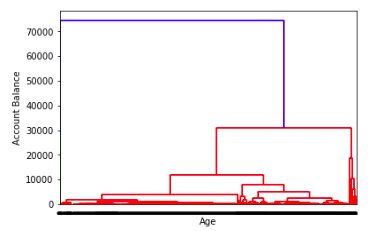
**Compare with K-means:**

The K-means is simple and relatively easy to understand but has poor performance for real-world situations. K-means model defines the center of each cluster with a radius that acts as hard cutoff for training the data set. Another drawback in the k-means algorithm is that it’s not flexible enough to account certain situation. There is no guarantee that this will circulate the global data. Lack of probabilistic cluster for many datasets may not perform well. Even though GMM is similar to the k-means, it maximizes the quality of the clusters to find weights encoding the probability of membership in each cluster. In GMM the result of each cluster is associated with a hard-edge sphere. GMM is used with a notion of "distribution over functions" is more concrete.

In the Bank marketing data set, we observe that the cluster of customers with higher bank balance has a higher density when GMM is applied compared to KMeans. This particular cluster can be targeted for promotions on Home Loans. This helps in identifying different customer segmentation of Bank customers.

**Hierarchical Clustering:**

Hierarchical cluster analysis, is an algorithm used to group up similar objects called clusters. Set of clusters is an endpoint, where each cluster is unique and each clusters have objects that are broadly similar to each other. The algorithm uses distance matrix or raw data which automatically compute a distance matrix in the background. It does two steps repeatedly. (1) Identifies two clusters that are close together, (2) merge the two most similar cluster. The mind output is dendrogram which shows the hierarchical relationship between the clusters. The two most similar part of a cluster (single-linkage) and two least similar bits of a cluster (complete-linkage) are also determined. Fig 1.4 is the hierarchical clustering using dendrogram.

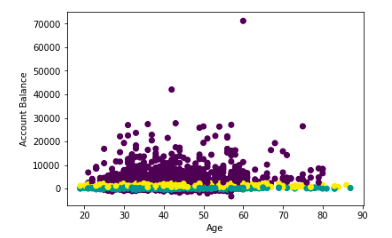


**Fig: 1.4 Hierarchical Clustering**

**Hierarchical Clustering - Agglomerative Clustering:**

Sequentially merging similar clusters using Hierarchical cluster analysis is known as agglomerative hierarchical clustering. This can be defined by grouping the observations initially into one single cluster, and then successively splitting these clusters which is known as divisive hierarchical clustering.

Hierarchical Clustering can be done without specifying the number of clusters present. Provides insights into possible clustering of data. Initially, every point is considered as a different cluster and then a convergence point is arrived at. We have used the Euclidean distance as the affinity and the ward as the linkage. Applying this method on Bank marketing data, we received a clustering which is like the GMM method. The customer segmentation targeted to send out he house loan promotions are the top most cluster in the below image.

****

**Fig: 1.5 Agglomerative Clustering**