GAUSSIAN MIXTURE MODEL EXPERIMENT REPORT

GANESHREDDY ANNAPAREDDY

50442295

ganeshre@gmail.com

1. GAUSSIAN MIXTURE MODEL

A probabilistic model called a "Gaussian mixing model" posits that all of the data points were produced by combining a limited number of Gaussian distributions with unknown parameters. Mixture models may be seen as a generalization of k-means clustering to include details on the covariance structure of the data as well as the locations of the latent Gaussian centres. The expectation-maximization (EM) approach for fitting a mixture of Gaussian models is implemented by the GaussianMixture object. Additionally, it can compute the Bayesian Information Criterion to determine how many clusters there are in the data and create confidence ellipsoids for multivariate models. A Gaussian Mixture Model may be learned from train data using the GaussianMixture.fit technique. Using the GaussianMixture.predict technique and test data, it may assign each sample the Gaussian that it most likely belongs to. For number of components selection we use SILHOUETTE SCORE.

SILHOUETTE SCORE :

The silhouette value gauges an object's cohesiveness with its own cluster in comparison to other clusters (separation). A high number on the silhouette implies that the object is well matched to its own cluster and poorly matched to nearby clusters. The silhouette has a range of 1 to +1.

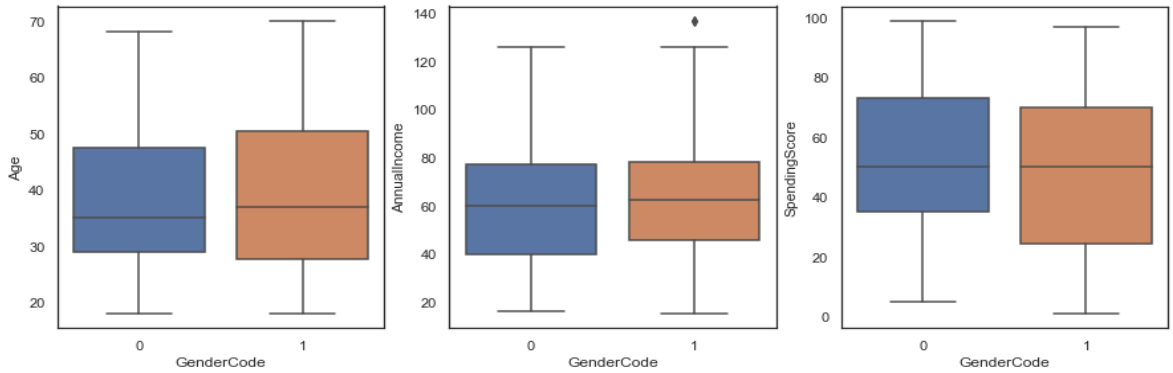
This model can be understood cearly when used a dataset for representation.

DATASET

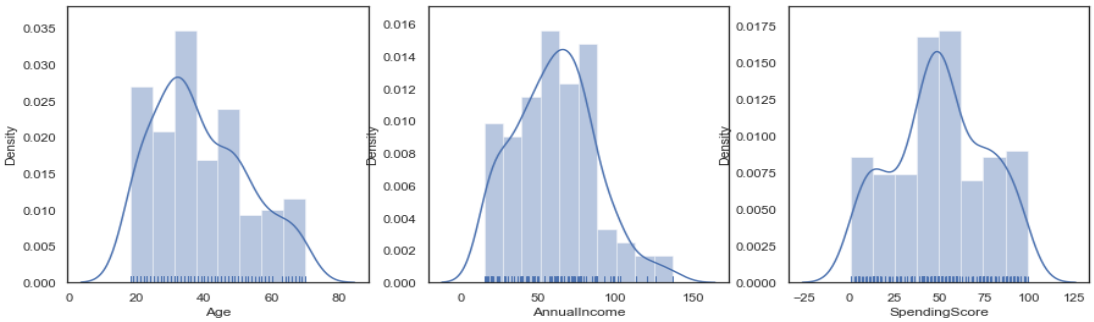
I used the Mall_customer dataset(https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python) to get the Gaussian mixture model experiment and see the evaluation result. Python was used for this task on the Jupiter notebook. I have used NumPy, matplotlib, seaborn, pandas , sk-learns and mpl_toolkits.mplot3d (for 3D visualization of data) libraries for the experiment. The dataset consists of

COLUMN	DESCRIPTION
CustomerID	It gives us the information on how many customers are present
Gender	It gives us the information on how many genders are there
Age	It gives us the information on what the age is of the customer
Annual Income (k$)	It gives us the information on how much income the customer has
Spending Score (1-100)	It gives us the information on how much a customer is spending

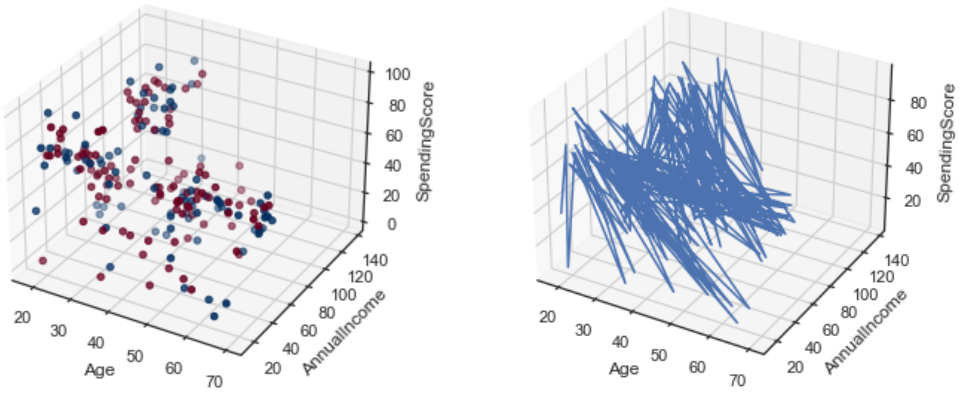
I used box plots to see the annual income and spending scores for both male and female customers.



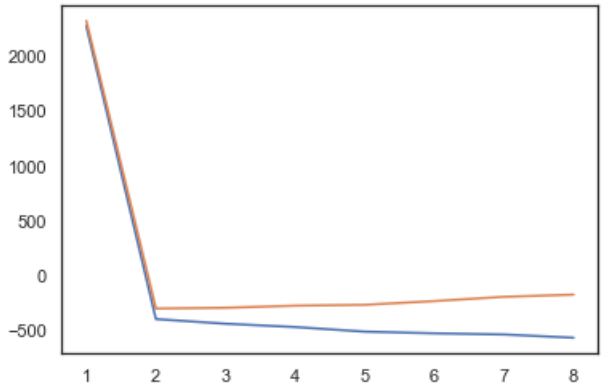
I used the distribution plots to look at the density next.



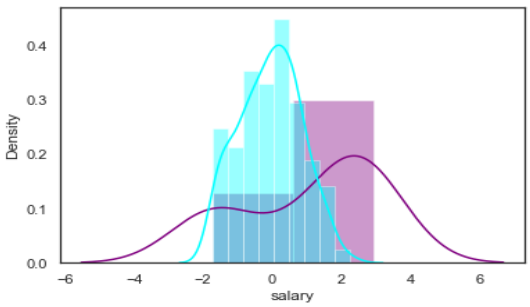
I have used 3D plots to visualize all 3 components which are Age, Annual Income and Spending score to get a better look at the dataset and understand.

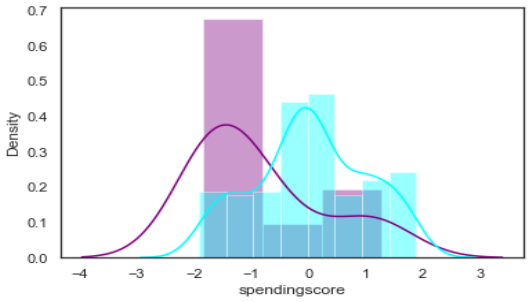


Based on Scree plot using Gaussian Mixture Algorithm, we could finalize 2 cluster groups.

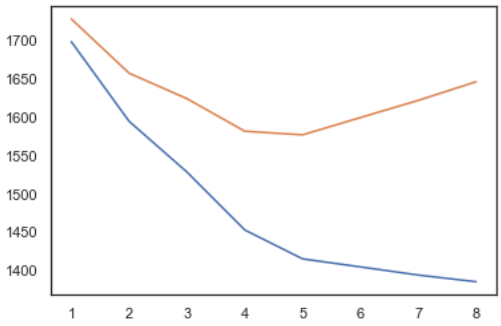


To look at the density of the data I used “distplot” library and got two plots for the annual income and spending score.

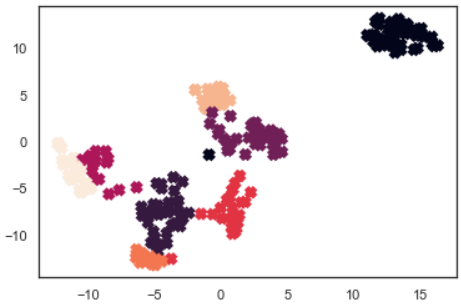




With gender column I got two clusters which can be seen from below figure.



Without the gender column I got 5 more clusters which can be seen from below figure.



PROS :

- It is the fastest algorithm for learning mixture models
- This algorithm will not bias the means towards zero or the cluster sizes to have particular structures that may or may not apply because it optimizes just the likelihood.

CONS :

- It becomes difficult to estimate the covariance matrices when there aren't enough points in each combination, and unless the covariances are artificially regularized, the process is known to diverge and produce solutions with infinite likelihood.

CONCLUSION:

Gaussian Mixture Model (GMM) Clustering handles ellipsoidal distributions, and makes 'soft' assignments to clusters, but is much slower than k-means for large datasets.