Gaussian Mixture Model

An overview and Its applications

Hasti Hojabr
ID: 97216040
June 25, 2022

Abstract

Gaussian mixture models is a probabilistic model and a popular clustering approach. It can be solved by expectation-maximization algorithm. We will implement this method on Iris Data-set and Online Retail Data-set and evaluated it using internal and external evaluation methods.

Introduction

A Gaussian mixture model (GMM) is a probabilistic model which involves the mixture or superposition of multiple and finite Gaussian distributions. A data-set has its own probability distribution, by using GMM we get finite number of Gaussian distributions for this data-set. Each Gaussian distributions will give us a cluster. The summation of these distributions will indeed give us the original distributions. For K Gaussian distributions $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k\sum_k)$ we have

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k\sum_k) \tag{1}$$

$$0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1 \tag{2}$$

where π_k are mixture weights. [1]

GMM can provide us the probabilities of the data point belonging to each of the possible clusters so we can say it's a more general form of k-mean clustering. The model parameter we need to find are

$$\boldsymbol{\theta} = \{\pi_k, \sum_k, \boldsymbol{\mu}_k : k = 0, ..., K\} \tag{3}$$

We can use Expectation Maximization to find these parameters and make cluster of our data. So GMM can used to soft cluster our data in a unsupervised learning manner.

Algorithm

The Expectation Maximization (EM) algorithm is an iterative method that produces maximum-likelihood (ML) estimates of parameters when there is a many-to-one mapping from an underlying distribution to the distribution governing the observation. [2]

Expectation Maximization algorithm consists of two steps, E-step and M-step. E-step Evaluate the responsibilities r_{nk} (posterior probability of data point n belonging to mixture component k) and M-step uses the updated responsibilities to re estimate the parameters given in Eq.3. EM algorithm for finding parameters of GMM is as follows:

1. Initialize $\pi_k, \sum_k, \mu_k$.
2. Evaluate responsibilities r_{nk} for every data point.

$$r_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k \sum_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}|\mu_k \sum_k)} \tag{4}$$

3. Re estimate parameters using the new r_{nk} from E-step.

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mu_n \tag{5}$$

$$\sum_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \mu_k)(\boldsymbol{x}_n - \mu_k)^T \tag{6}$$

$$\pi_k = \frac{N_k}{N} \tag{7}$$

4. Repeat the last two process untill the parameter estimate has converged. [3] [1]

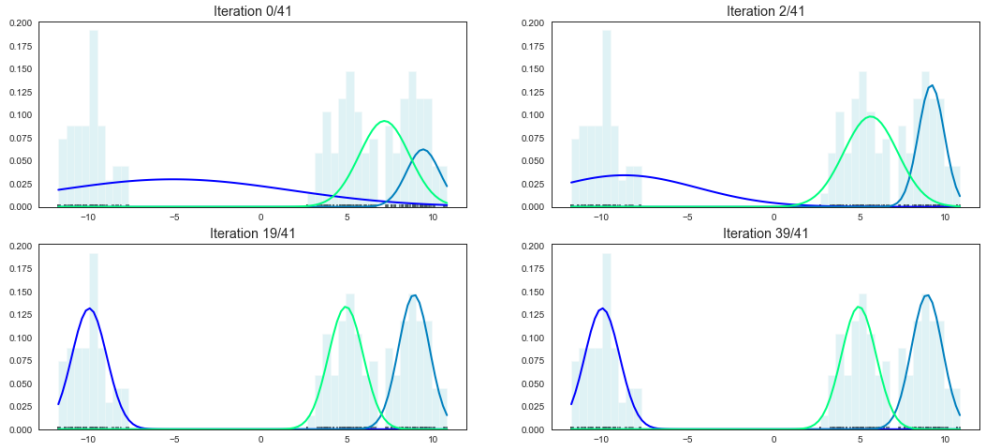

Figure 1: One dimensional EM algorithm results for a hypothetical data set.

The entire iterative process repeats until the algorithm converges, giving a maximum likelihood estimate.

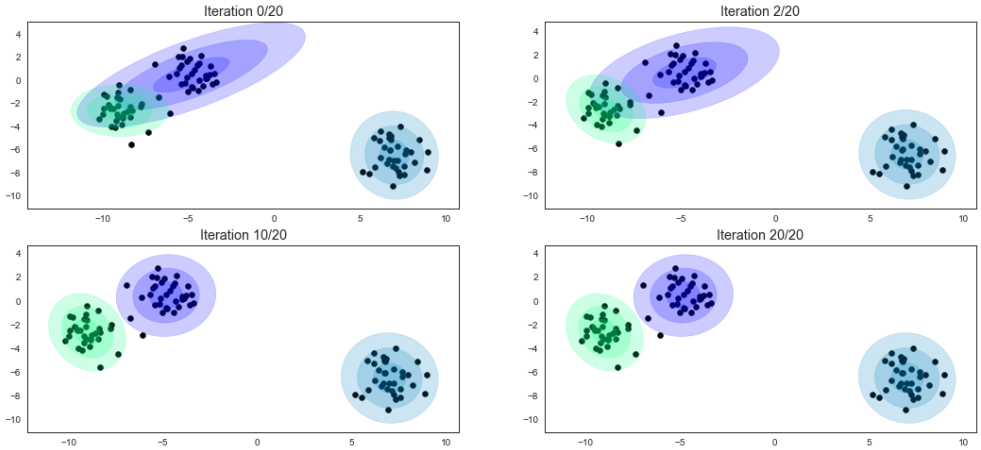

Figure 2: Two dimensional EM algorithm results for a hypothetical data set using make$_b$*lobslibrary*.

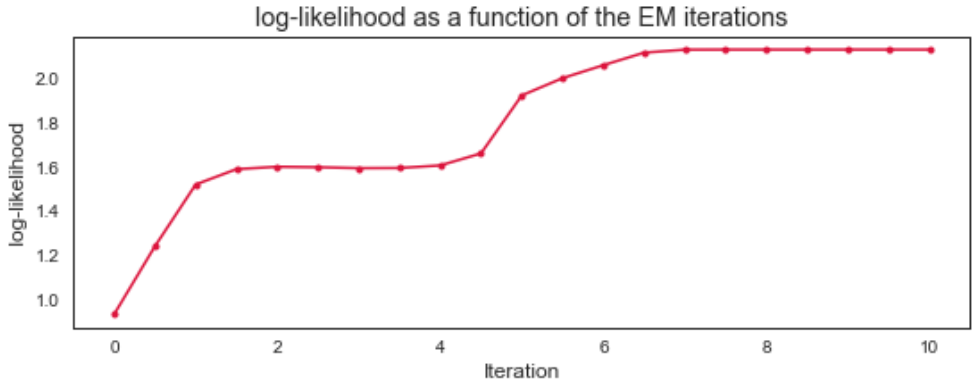

Figure 3: Log-likelihood as a function of the EM iterations for the hypothetical data set using make$_b$*lobslibrary*.

Visualization

We visualize clusters in two dimension using Ellipses. We know that covariance matrix can be diagonalized.

$$cov = \begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{pmatrix} \tag{8}$$

We can diagonlized cov using numpy library and command *numpy.linalg.svd*. As its given in numpy official website, When the given matrix a is a 2D array, and full matrices=False, then it can be factorized as u @ np.diag(s) @ vh = (u * s) @ vh, where u and the Hermitian transpose of vh are 2D arrays with orthonormal columns and s is a 1D array of a's singular values.

Evaluation

In a clustering problem we have to understand how well the data is grouped into different clusters by the algorithm and how many clusters we need. In

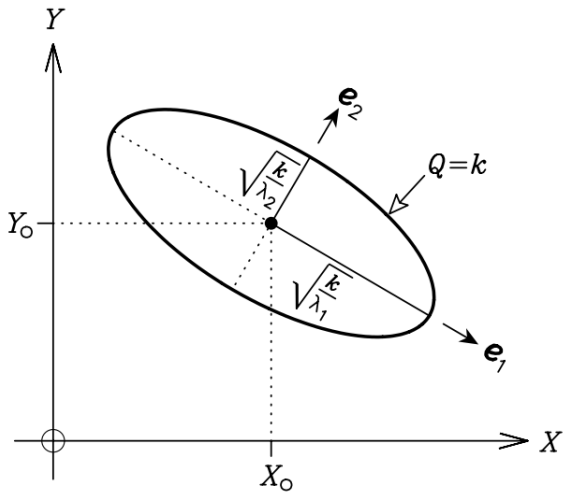

Figure 4: The contour in the X–Y parameter space along which Q = k, a constant. It is an ellipse, centred at (μ_x , μ_y), the characteristics of which are determined by the eigenvalues λ and eigenvectors e of the second-derivative matrix.[4]

internal evaluation, clustering results are evaluated based on data used for the clustering but in external evaluation results are evaluated based on external benchmarks and class labels no present in clustering itself.

Silhouette score

Determining the optimal number of clusters for a data set is an important problem. The silhouette score also tells us how good the clustering algorithm performs. The Silhouette Coefficient is for a single sample is then given as

$$s = \frac{b - a}{max(a, b)} \tag{9}$$

For finding the right K for GNN, we loop through 1 to n clusters and caluclate Silhouette coefficient for each data. [5]

Bayesian information criterion

Another way of choosing the right number of clusters is by optimizing the Bayesian information criterion (BIC).

$$BIC(C_k) = \sum_{i=1}^{k}\{-\frac{1}{2}n_i log|\sum_i| - Nk(d + \frac{1}{2}d(d+1))\} \tag{10}$$

where n_i is the number of samples in cluster c_i, C_k is a clustering with k c_i clusters. We must choose the clustering which maximizes the BIG criterion Eq.6. [6]

BIC is the standard method of model selection for GMM, however it assumes that the true generating model is among the tested models. In general, that is not the case. So you may want to use another method such as AIC or

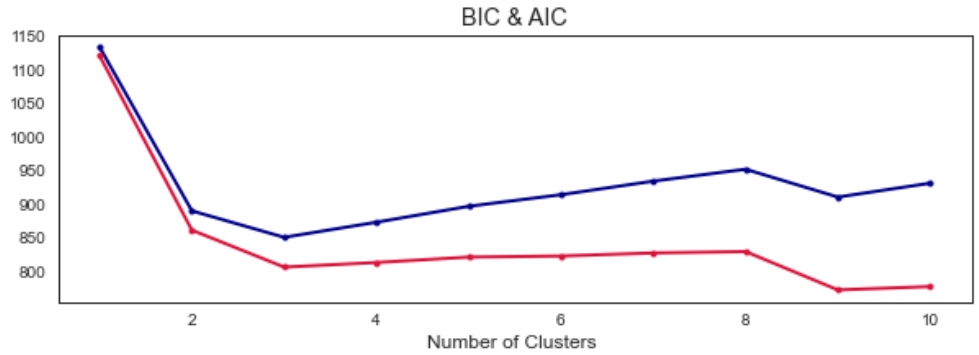

Figure 5: BIC for the hypothetical make$_b lobs' dataset$.

silhouette coefficient to compare. All these methods can be appropriate and they usually agree. If they don't agree, you may also want to consider infinite GMM, which determines the model complexity for you.

Confusion matrix

Confusion matrix is a specific table layout that allows visualization of the performance of an algorithm.

$$M = \begin{bmatrix} \text{True Negatives} & \text{False Positives} \\ \text{False Negatives} & \text{True Positives} \end{bmatrix} \tag{11}$$

Applications

GMM has many applications. It can generate data, soft clustering, resolve point set registration problems in image processing and many more. We have compered sklearn's GMM and our own EM algorithm in clustering. We implement this method on two data-sets, *The Iris flower data set* and *Online Retail data set*.

The Iris flower data set

It is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper [7] The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.

The data set consists of 150 samples and 4 targets, It consists from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. [8]

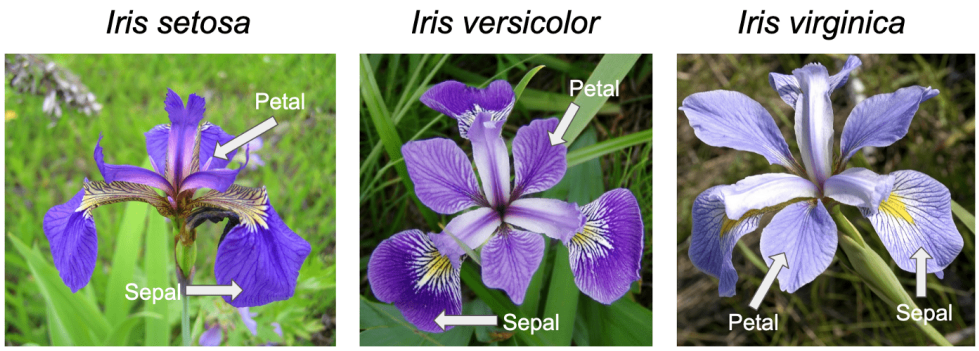

Figure 6: The Iris flower data set

The data-set can be used using sklearn library. We ignore the target column and cluster data features using GMM. Then we can compare GMM's prediction and the actual target and do an external evaluation of our model.

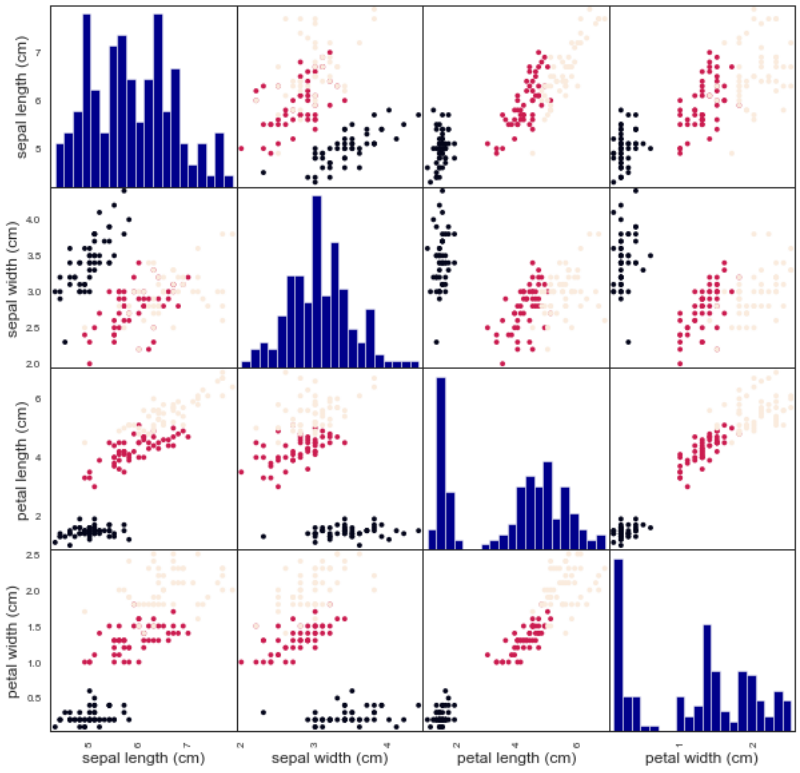

Figure 7: The Iris flower data set scatter matrix.

First using BIC and AIC we find the right number of clusters. It is 3 and it corresponds to our target classes correctly.

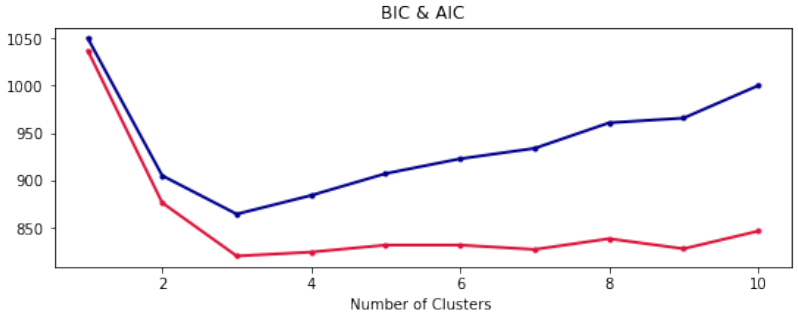

Figure 8: BIC and AIC on the features of the Iris data set.

Now using the right number of clusters we can cluster out data. One of the benefits of GMM is that it clusters out data in ellipses unlike the k-means which clusters in circles. Ellipses have more flexibility than circles and it increase model's accuracy.

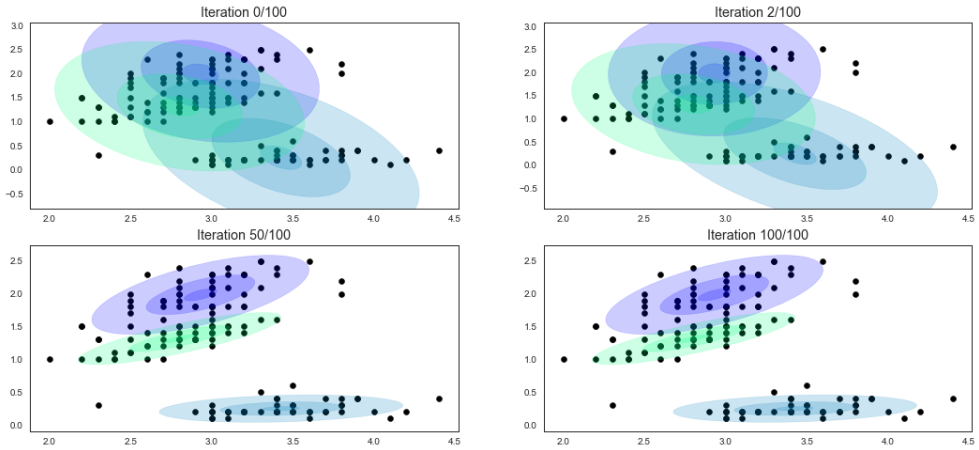

Figure 9: Clustering the Iris data set using EM algorithm.

We can see that predicted clusters matches target classes accurately.

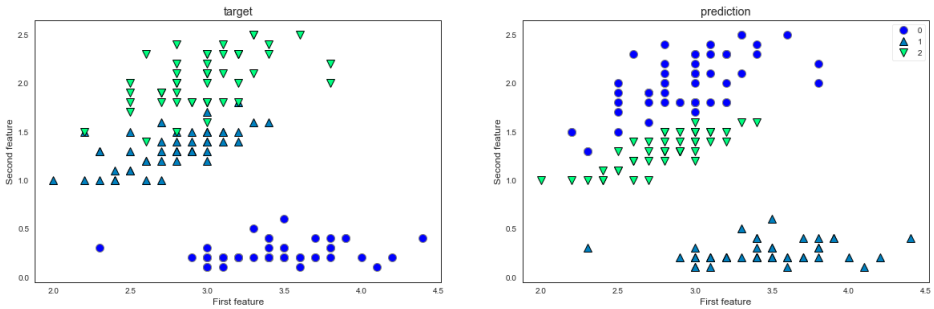

Figure 10: Target and prediction classes.

Online Retail

Its is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. [9]

To maintain and extend business, it is important to know the customers and group them. The goal is to maximize the impact of customized plans focused on targeted customers.

RFM is an effective customer segmentation technique. Recency (R) value is the number of days a customer takes between two purchases. Frequency (F) is defined as the number of purchases a customer makes in a specific period. Monetary (M) is defined as the amount of money spent by the customer during a certain period.

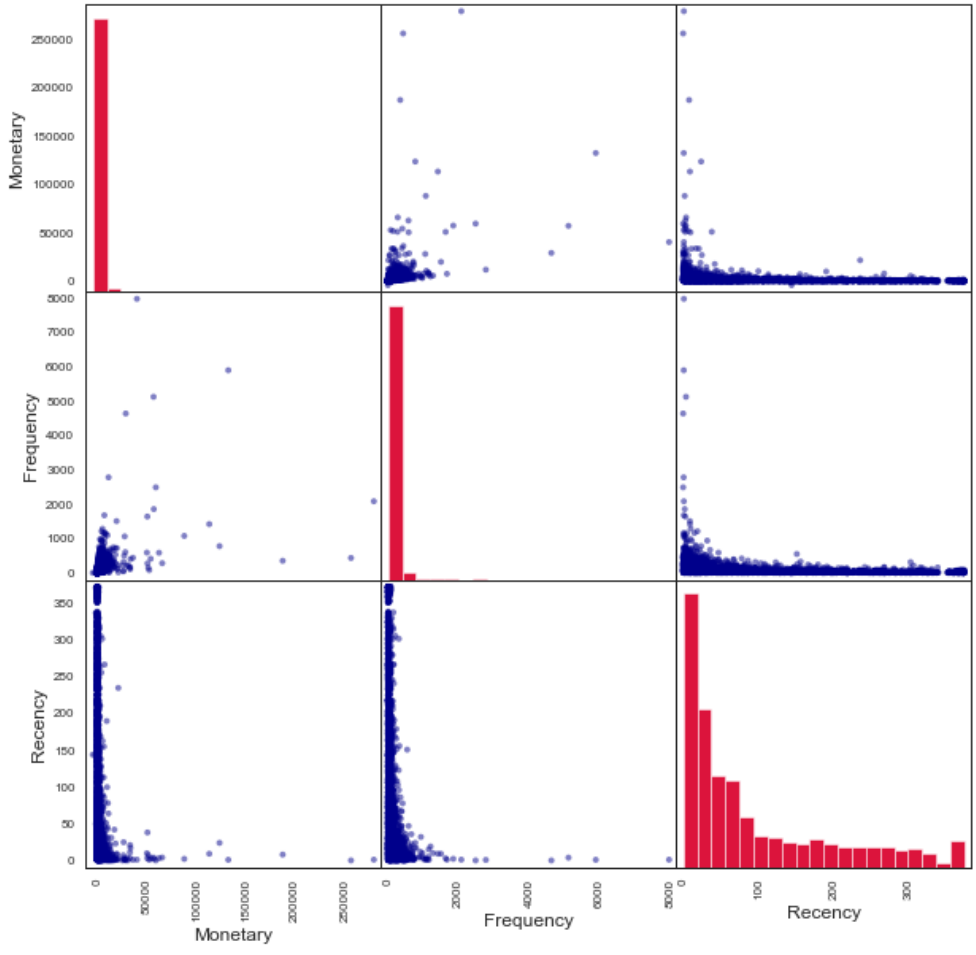

Figure 11: RFM scatter matrix

We try to find the right number of clusters using BIC. As we can see it does not have a minimum. Its means the model will work better if we would have more clusters. But if we increase the number of clusters it will make our model overfit. We can see after 3 clusters the curve becomes fairly smooth, so

we could assume we have 3 to 4 clusters.

We can Silhouette Scores though its not recommended for Mixture models. But in our problem it will give us more certainty about our assumption of 3 to 4 cluster.

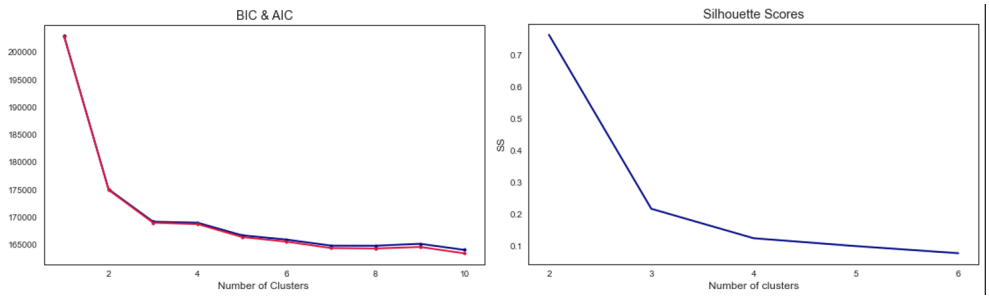

Figure 12: left) BIC and AIC. right) Silhouette Scores.

Using Sklearn we can cluster our data for K=3 and 4. We have removed outliners for better accuracy. We can see that two of the clusters in case k=4 are fairly close to another. So when we are labeling them we might consider them as one.

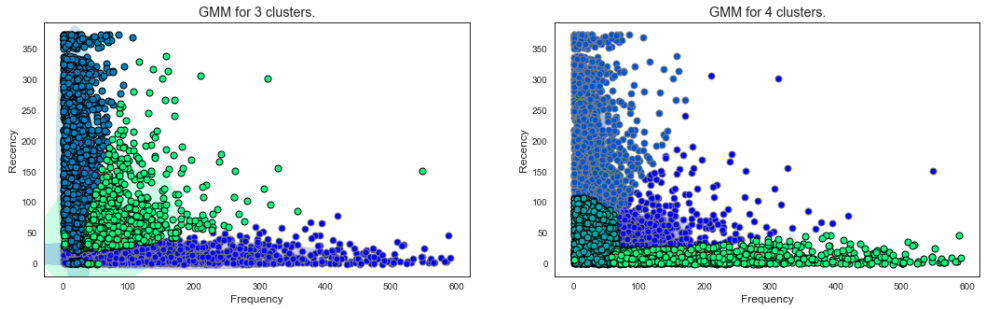

Figure 13: F-R : Clustering Online retail data set using sklearn for left) 3 clusters and right) 4 clusters.

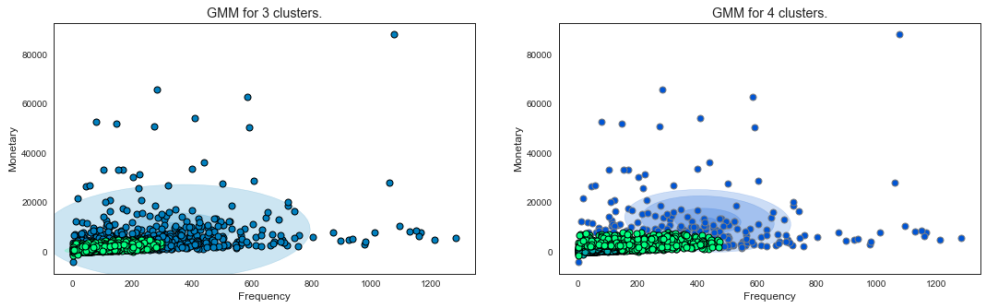

Figure 14: F-M : Clustering Online retail data set using sklearn for left) 3 clusters and right) 4 clusters.

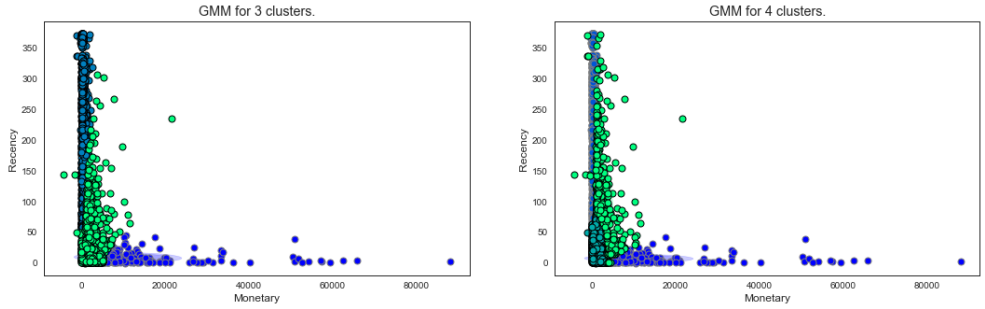

Figure 15: M-R : Clustering Online retail data set using sklearn for left) 3 clusters and right) 4 clusters.

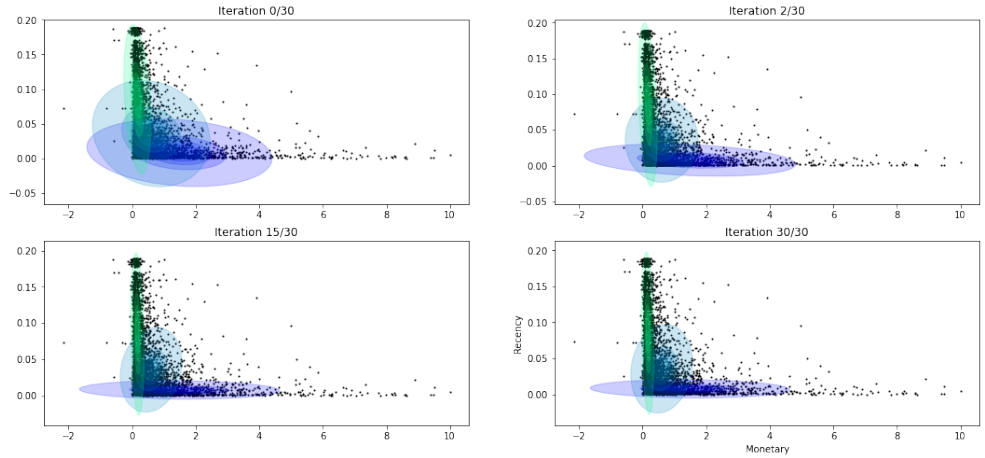

Figure 16: M-F : Clustering Online retail data set using EM for 3 clusters.

We can see EM's results correspond to Sklearn's results. The clusters are the same.

So we have three types of customers. Type one spends a lot and have shop frequently. Type 2 are recent shoppers with low monetary, and type 3 are lost or losing customers which have a high M and F but a low R. So we can target the customers and make a better advertisement plan. If we had more features we can make better clustering and be more precise with covariance and boundaries of ellipse.

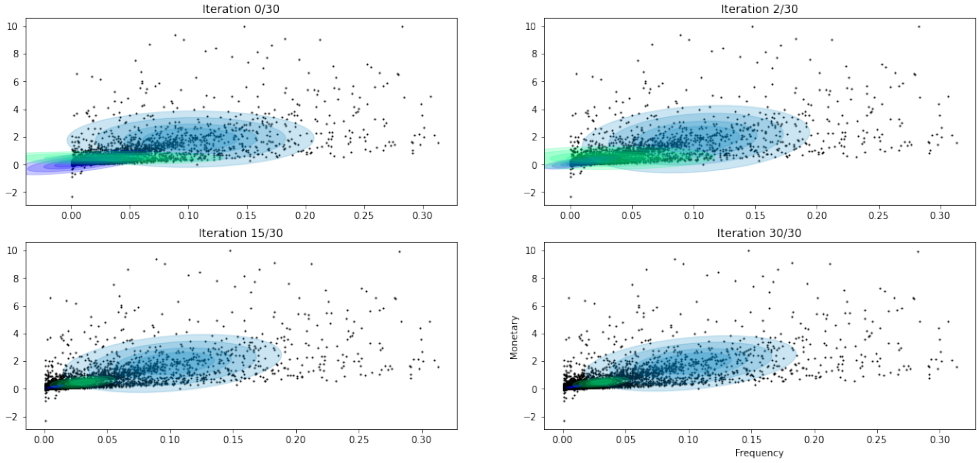

Figure 17: F-M : Clustering Online retail data set using EM for 3 clusters.

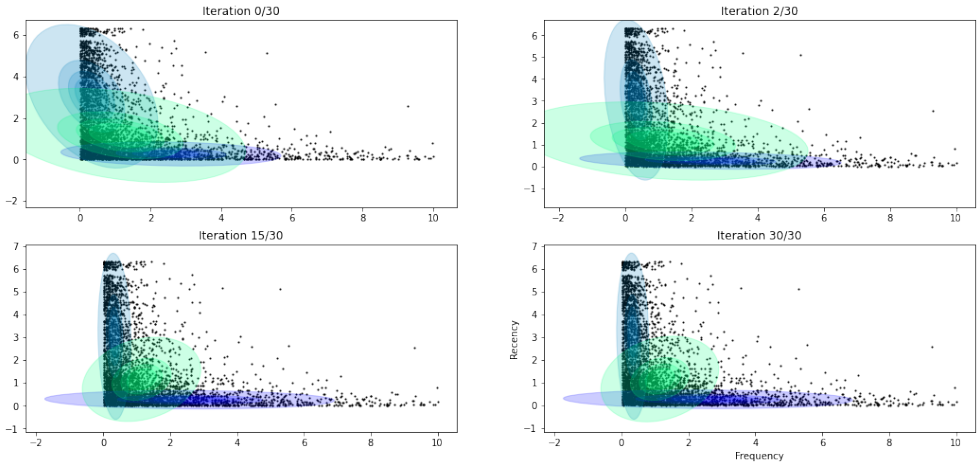

Figure 18: F-R : Clustering Online retail data set using EM for 3 clusters.

References

1. Deisenroth Marc Peter , Faisal A. Aldo, Ong C.S. (2019) *Mathematics For Machine Learning*. Cambridge University Press.

2. Moon, T. K. (1996). The expectation-maximization algorithm. IEEE Signal Processing Magazine, 13(6), 47–60. doi:10.1109/79.543975 doi:10.1109/79.543975

3. Guorong Xuan, Wei Zhang, Peiqi Chai. (n.d.). *EM algorithms of Gaussian mixture model and hidden Markov model*. Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205). doi:10.1109/icip.2001.958974

4. Sivia D.S, Skilling J. (2006) *Data analysis A Bayesian Tutorial*.Oxford University Press, USA.

5. Shahapure, K. R., Nicholas, C. (2020). *Cluster Quality Analysis Using Silhouette Score*. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). doi:10.1109/dsaa49011.2020.000

6. Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association, 66(336), 846. doi:10.2307/2284239

7. FISHER, R. A. (1936). THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. Annals of Eugenics, 7(2), 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

8. Fisher, R. A. (1936). *The Iris flower data set*. UCI Machine Learning Repository: Iris Data Set

9. Daqing Chen (2015). *The Online Retail data set*. UCI Machine Learning Repository: Online Retail Data Set