PRACTISING UNSUPERVISED CLUSTERING WITH GAUSSIAN MIXTURE MODELS

Asma Umar
Department of Electrical Engineering, , University of Management and Technology, Lahore, Pakistan.
Email: asma.umar@umt.edu.pk

ABSTRACT- *This paper discusses the unsupervised clustering practice of largely distributed data by utilizing Expectation Maximization (EM) and Maximum Likelihood (ML) in Gaussian Mixture Model. Results in this paper state the fact that it is a better approach to represent a largely distributed data with a number of clusters having various different means but minimum variance possible. Instead of describing it with only a single distribution having a single mean but large variance among the data points.*

Keywords- Clustering, Expected Maximization, Gaussian Mixture Model, Maximum Likelihood

1. INTRODUCTION

Clustering or Cluster Analysis is the task of grouping data in such a way that samples in one group share some properties with the samples from the same group but not with the samples from any other group. Unsupervised Clustering is the area of clustering in which no labels are assigned to the clusters. While proper labels are assigned in case of Supervised Learning. These labels can be known prior to the analysis. Clustering can be distinguished in:

- Hard Clustering: an object belongs to a cluster or not
- Soft Clustering: an object belongs to each object to a certain degree (likelihood of belonging to the cluster)

The most commonly used clustering model is the distribution based clustering. In which clusters can be defined as objects belonging to the same distribution. One benefit of using this model is that it provides complex models with correlations and dependencies among objects along with providing the clusters. One issue is that it puts quite burden on the user for choosing the right model.

Normal Gaussian is one of the most widely and commonly used distributions for data model representation. It is being used in statistics, social sciences and neural sciences for this very reason. Hence the clustering problem actually becomes parameter estimation if we intend to present the complete data set with a set of Gaussian Mixture models.

Gaussian Mixture Model (GMM) is nothing else but actually a parametric estimation of a probability density function which is represented by a weighted sum of a number of individual Gaussian distributions. The parameters of these individual distributions are estimated from given training data by utilizing the Expectation Maximization (EM) and Maximum Likelihood (ML) algorithms. Both of them are iterative algorithms. GMM is a soft clustering algorithm that gives the likelihood of each object with each cluster.

The aim of this paper is simply to utilize GMM for unsupervised clustering of data set. The issues in selecting a proper model are:

1. How to initialize the parameters of each component in GMM?
2. How many components at least?
3. How many iterations of EM for ML?

Bayesian Information criteria (BIC) is used for selection of minimum possible clusters. ML approach is used to end the iterations of the EM algorithm.

The paper is arranged in a manner that section 1 provides an introduction of clustering with GMM. The next section describes the literature review on GMM, EM, ML and BIC & AIC algorithms. EM algorithm for GMM is discussed in the section after literature review. The methodology section discusses the data set used and the working of algorithm. Discussion section states the results and discusses the analysis. At the end the conclusion section concludes the work.

2. LITERATURE REVIEW

In real world we often come across data sets that are quite widely distributed. And so the best approach is to present that data with a weighted sum of a number of groups instead of a single one.

A. Gaussian Mixture Model (GMM)

GMM describes a data set as a weighted sum of a number of probability distributions having different parameters. The aim is to keep variance at a low level and achieving a best model.

As fig. 1 shows, the actual data can best be described as a sum of two normal pdfs:

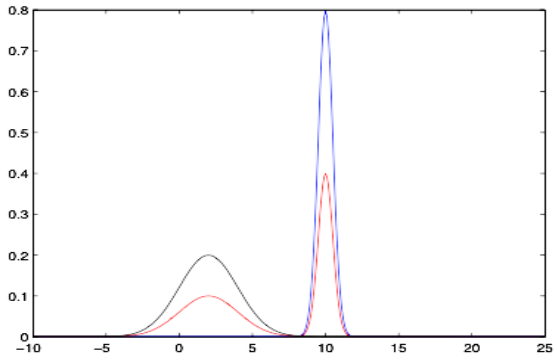

Fig. 1 Data composed of two pdfs

$f_o(x) = N(x; 2, 2)$
$f_1(x) = N(x; 10, .5)$
With equal weights
$\Pi = [.5\ .5]^T$
Π describes the weight of each mixture in the model.

In general, a data consisting of k number of mixtures is described by equation 1 and summarized in equation 3 in terms of normal distributions [1].

$$\boldsymbol{p(x)} = \pi_o f_o(\mathrm{x}) + \pi_1 f_1(\mathrm{x}) + \pi_2 f_2(\mathrm{x}) + \cdots + \pi_k f_k(\mathrm{x}) \quad (1)$$

$$\boldsymbol{p(x)} = \sum_{i=0}^{k} (\pi_i f_i(\mathrm{x})) \quad (2)$$

$$\boldsymbol{p(x)} = \sum_{\boldsymbol{i=0}}^{\boldsymbol{k}} (\pi_i N(x|\mu_i, \sum_i)) \quad (3)$$

Where $\sum_{i=0}^{k}(\pi_i) = 1$ is a necessary condition to be satisfied.

Where each component density is a D variate Gaussian function:

$$N(x|\mu_i, \sum_i) = \frac{1}{2\pi^{D/2}|\sum_i|^{1/2}} \exp\left(\frac{-1}{2}(x - \mu_i)' \sum_i^{-1} (x - \mu_i)\right) \quad (4)$$

Where μ_i and $\sum_i$ are the mean and covariance of ith component respectively. Hence the parameters of each component in GMM can collectively be stated as: $\boldsymbol{\lambda} = \{\pi_i, \mu_i, \sum_i\} \quad i = 1 \ldots k$

B. Expectation-Maximization (EM)

Expectation-Maximization algorithm is a commonly used standard approach for estimating the parameters for GMM [2]. However there are some limitations of this algorithm i.e. some a priori knowledge of parameter initialization, number of components to be incorporated in the model. The results depend on the parameter initialization. Larger the number of components, greater is the log likelihood. But it also creates computational burden and a risk of over fitting. Hence an intelligent choice of k and parameter initialization is required for better results.

C. Maximum Likelihood (ML)

Maximum Likelihood algorithm can be considered as part of Expectation Maximization. EM is an iterative algorithm for estimating parameters for maximizing the likelihood for the data.

D. Bayesian Information Criteria (BIC) & Akaike Information Criteria (AIC)

BIC and AIC serve as model selector for GMM. The base their decision on the log likelihood value of different models and provides the best option model the best fits the data. By choosing the fitted candidate model corresponding to the minimum value of BIC, one is attempting to select the candidate model corresponding to the highest Bayesian posterior probability.

3. EM ALGORITHM FOR GMM

EM Algorithm is an iterative algorithm that starts from some initial parameters λ (e.g. random) and then keeps on updating the parameters until convergence is achieved. It works in two steps:

I. E-Step (Expectation)
II. M-Step (Maximization)

I E-Step

Denote the current parameter values as λ= {π, μ, ∑}. Compute the responsibility terms $T(Z_{nk})$ using the membership weights equation given below, for all data points x_i; 1 ≤ i ≤N and all mixture components 1 ≤k ≤ K. Where N is the total number of data points in the data set.

$$T(Z_{nk}) = \frac{\pi_k N(x|\mu_k, \sum_k)}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} (\pi_j N(x|\mu_j, \sum_j))}$$

Where 1 ≤ i ≤N and 1 ≤k ≤ K.

This equation is computed from direct application of Bayes Rule.

Note that for each data point x_i the membership weights are defined such that $\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} (T(Z_{nk})) = 1$. This yields an N×K matrix of membership weights, where each of the rows sum exactly to 1. $T(Z_{nk})$ tells the probability of ith data point belonging to each of the k clusters.

II M-Step

After calculating the weight/ responsibility vector, use it to estimate the new parameters $\lambda = \{\pi_{new}, \mu_{new}, \sum_{new}\}$.

Let $N_k = \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} (T(Z_{nk}))$ i.e. the sum of the membership weights for the kth component. This is the effective number of data points assigned to component k.

Using this information, the new parameters are learned as:

$$\pi_{k,new} = \frac{N_k}{N}; 1 \le k \le K$$

$$\mu_{k,new} = \frac{\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} (T(Z_{nk}))\, x_n}{N_k}$$

$$\sum_{k,new} = \frac{\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} (T(Z_{nk}))\, (x_n - \mu_k)(x_n - \mu_k)'}{N_k}$$

Note that the dimensions of μ and ∑ are 1×d and d×d respectively. Where d is the number of columns/features in data.

Order of computations in M-Step should be: first of all compute k new π's, then k new μ's and at the end k new ∑'s.

When done with the parameter calculations, the M-Step is complete. The again with the new parameters, the weight vector is recomputed in E-Step followed by new parameter computation using this weight vector, the M-Step is repeated. One pair of E-Step and M-Step is collectively called an iteration.

Parameter Initialization and Convergence issues in EM

EM algorithm can be started one of the two ways:

1. Initialize parameters first and start with the E-Step computations
2. Initialize the weight vector first and directly start with the M-Step first

The initial selection of parameters or weights can be done randomly (e.g. select K random data points as initial means and select the covariance matrix of the whole data set for each of the initial K covariance matrices) or some heuristic method could be utilized for the initial selections (such as first clustering the data by utilizing k-means algorithm and then defining weights based on k-means memberships).

Convergence is achieved by computing the log likelihood by the equation given below:

$$\boldsymbol{ln}(\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{\pi},\mu,\sum) = \sum_{n=0}^{N}\left(\ln(\sum_{k=1}^{K}(\pi_k N(x_n|\mu_k,\sum_k)))\right)$$

After every iteration and stopping when it ceases to change in an effective value i.e. setting a minimum thresh-hold value of change.

4. METHODOLOGY

The data set used for clustering in this paper is death rates and birth rates of 73 countries [3]. Hence the actual data is 73×2.

Fig. 2 shows the original data.

As is clear from fig.2 data visuals, it's not an ideal approach to represent this data with a single distribution with a single mean and a large variance. Instead selecting a model with a number of distributions with different means and smaller variance is a good approach. Hence utilizing GMM.

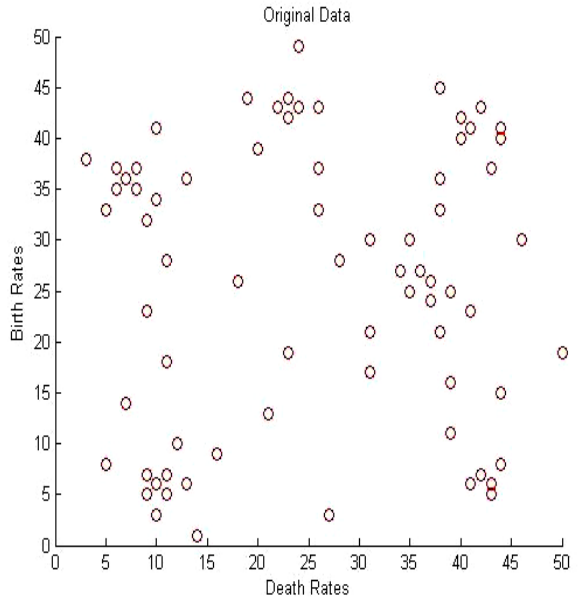

Fig.2 Scattered view of original data

Fig. 3 shows the flow chart of GMM methodology followed in this paper.

For continence and computation simplicity, random parameter initialization is done and algorithm is started from the E-Step first, computing weight vector from randomly initialized parameters. Initialization is done as follows:

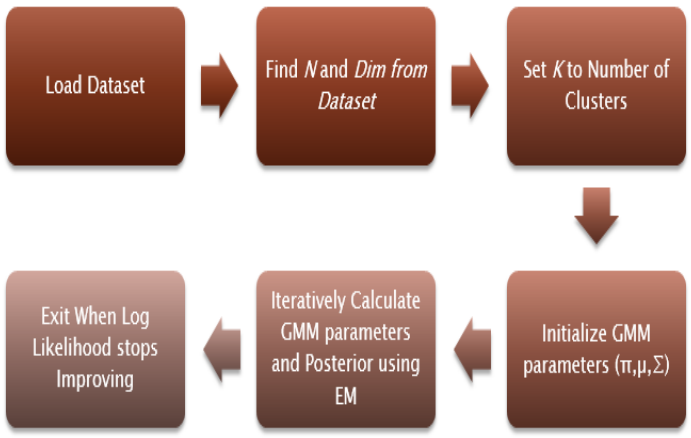

Fig. 3 GMM Implementation Flowchart

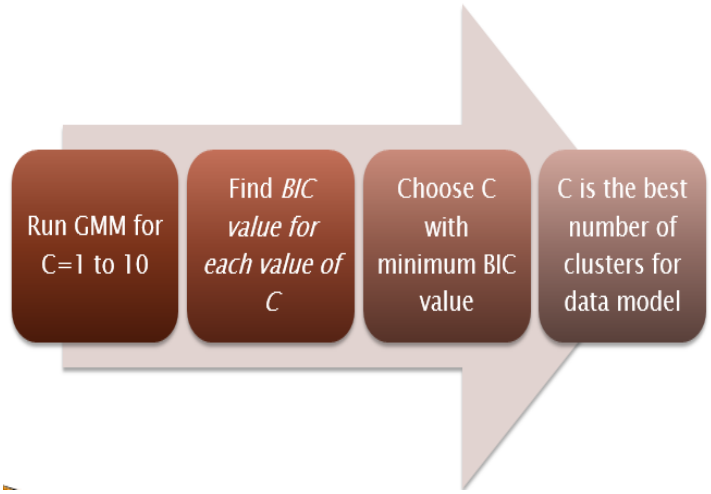

Fig.4 Best model selection through BIC Flowchart

- Mu (μ) set to random 1×d vector
- Sigma (Σ) set to multiple of Identity Matrix of size d×d
- Pi (π) set to 1/K
- Posterior/Tau (τ) set to All Zero.

A maximum number of iterations is set to 100 and number of clusters set to 10 for safe calculations. The log thresh-hold is set to .0001, so as to stop the iterations when the deference between the old and new log values is less than thresh, the algorithm stops.

For best model selection, BIC information is used. GMM-EM algorithm is run for number of clusters from 1 to 10. At the end BIC value of each model is compared and the one with minimum BIC value is choose to be the best option. Fig. 4 shows the flow chart for best model selection for the data.

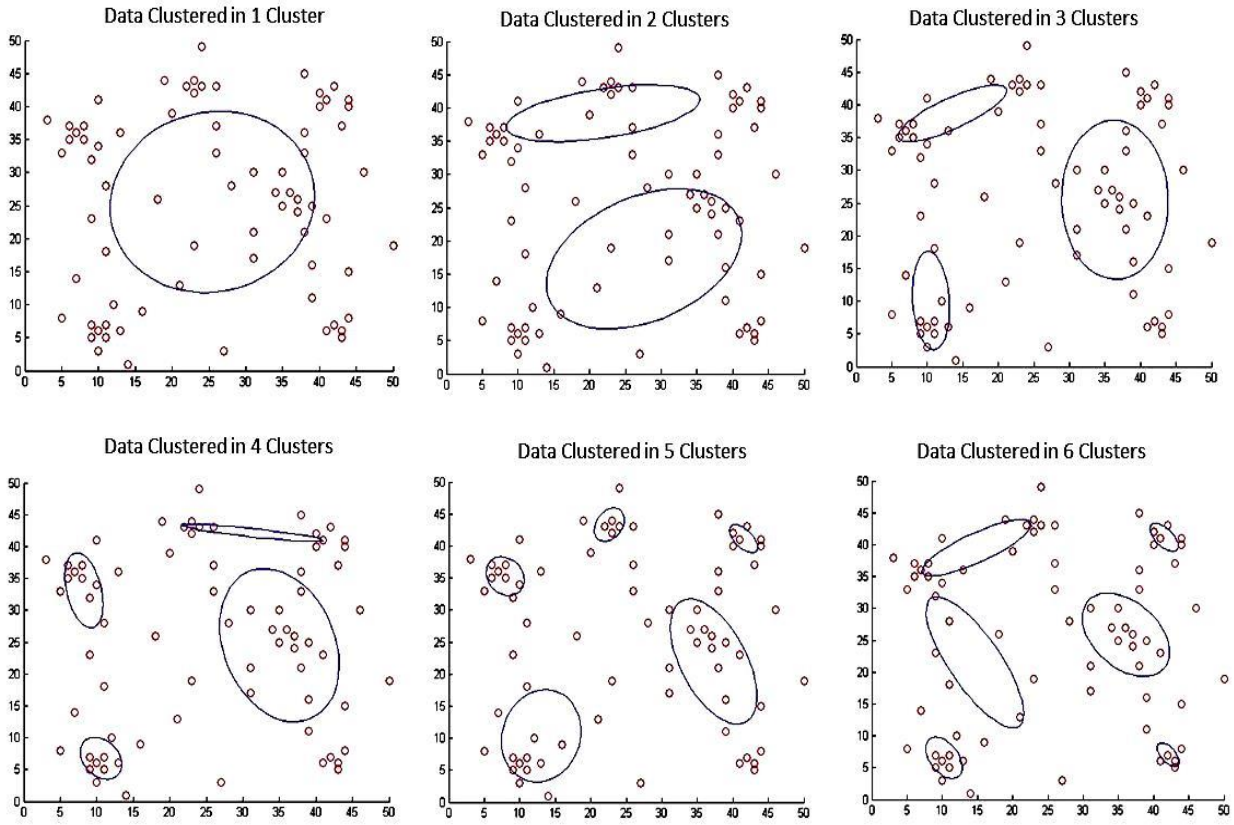

Fig.5 Clustered representation of data for different values of C

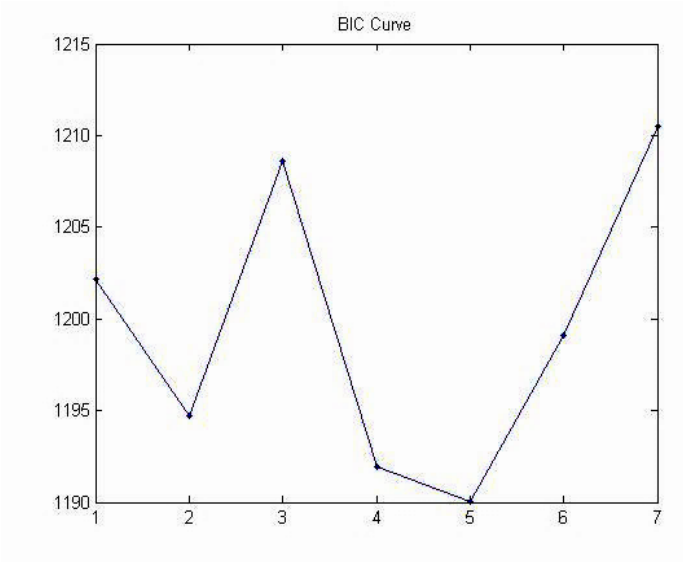

Fig.6 BIC Curve

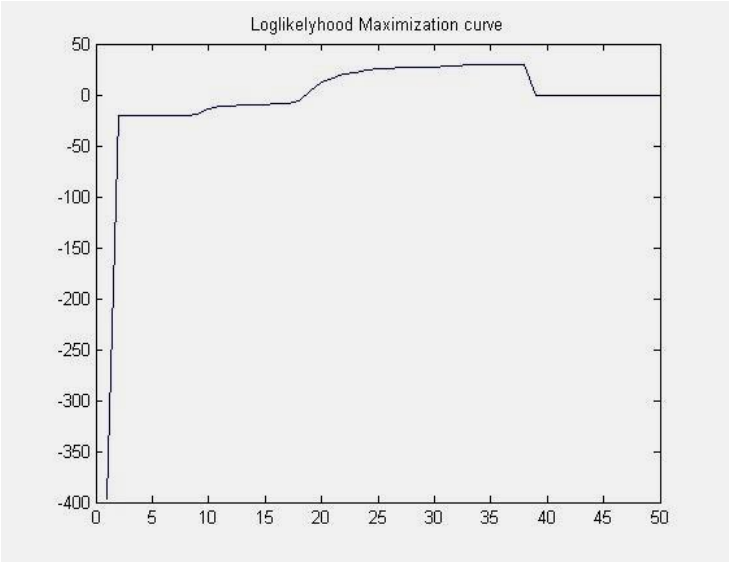

Fig.7 Log Likelihood curve

5. RESULTS & DISCUSSION

Visual cluster represemtation for each value of C is shown in fig.5. fig.6 shows the BIC curve against the corresponding values of C. It is clear from the curve that 5 cluster model is the best option for the used dataset. Which is also clear from fig.5 if we take a look at the clusters. Below 5 clusters, some of the clusters have very large variance in order to include all the data points. While for more than 5 clusters, variance starts to decrease prominently, leading to over-fitting. Hence 5 clusters seems to be the best minimum cluster model.

Fig. 7 shows the log likelihood curve for the data set for 5 cluster model. Log likelihood curve is unstable at the start but then after a number of iterations, it starts to settle. As clear from the curve the algorithm converges at 38 iterations of the EM algorithm.

6. CONCLUSION

The task of clustering is unsupervised learning. For which normal distribution based approach turns out to be a quite good technique. And GMM works quite efficiently for this purpose. BIC criteria and ML convergence criteria are the standard approach for GMM-EM algorithm convergence. One reason, the algorithm took too much iterations for convergens condition satisfaction is the random parameter initialization. Which can be overcome by utilizing K-Means for parameter initialization in EM.

REFERENCES

1. Douglas Reynolds, *Gaussian Mixture Models*, Department of Defense, Air Force Contract FA8721-05-C-0002
2. Hartley, H., *Maximum likelihood estimation from incomplete data*. Biometrics 14, 174–194 (1958)
3. Helmut Spaeth, *Cluster Dissection and Analysis*, Theory, FORTRAN Programs, Examples, Ellis Horwood, page 144, 1985