A study on Gausian Mixture Model

Shashank Pathak(15BCE1287)*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: shashank.pathak2015@vit.ac.in*

***Abstract*—Gaussian Mixture Models (GMMs) are among the most statistically mature methods for clustering (though they are also used intensively for density estimation).The optimixzation method used in GMM is Expectation Maximization.We study the performance of the GMM on a height dataset for a class consisting of male and female students.We also see how GMM can be used for othe things like density estimation**

1. Introduction

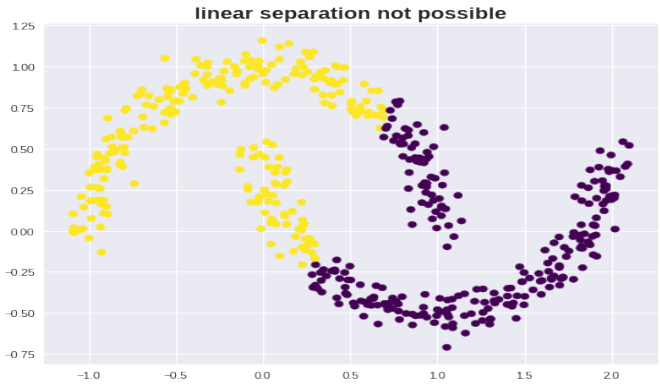
Gaussian mixture models are a probabilistic model for representing normally distributed subpopulations within an overall population. Mixture models in general don't require knowing which subpopulation a data point belongs to, allowing the model to learn the subpopulations automatically. Since subpopulation assignment is not known, this constitutes a form of unsupervised learning.

2. Methodology

For GMMs, we will find the clusters using a technique called Expectation Maximization. This is an iterative technique that feels a lot like the iterative approach used in k-means clustering.

In the Expectation step, we will calculate the probability that each data point belongs to each cluster (using our current estimated mean vectors and covariance matrices). This seems analogous to the cluster assignment step in k-means.

In the Maximization step, well re-calculate the cluster means and covariances based on the probabilities calculated in the expectation step. This seems analogous to the cluster movement step in k-means.



3. Database - Breast Cancer Wisconsin Dataset(Original)

The dataset used is a heights dataset for a class consisting of male and female students

4. Algorithm

In the Expectation step, we calculate the probability that each data point belongs to each cluster.

Well need the equation for the probability density function of a multivariate Gaussian. A multivariate Gaussian (multivariate just means multiple input variables) is more complex because there is the possibility for the different variables to have different variances, and even for there to be correlation between the variables. These properties are captured by the covariance matrix.

$$g_j(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1} (x-\mu_j)}$$

.

5. Experiments

Gaussian distributions give flexibility to the clustering, and the same basic two step E-M algorithm used in K-Means is applied here as well.Randomly initialize location

and shapeAfter this Repeat until converged:there are 2 steps E-step: for each point, find weights encoding the probability of membership in each cluster.and M-step: for each cluster, update its location, normalization, and shape based on all data points, making use of the weights

The result of this process is that we end up with a smooth Gaussian cluster better fitted to the shape of the data, instead of a rigid inflexible circle. Note that because we still are using the E-M algorithm there is no guarantee of a globally optimal result. We can visualize the results of the model.

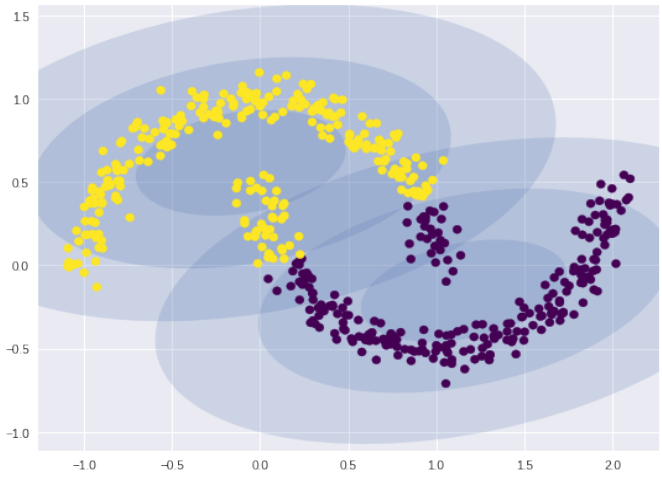


Figure 1.

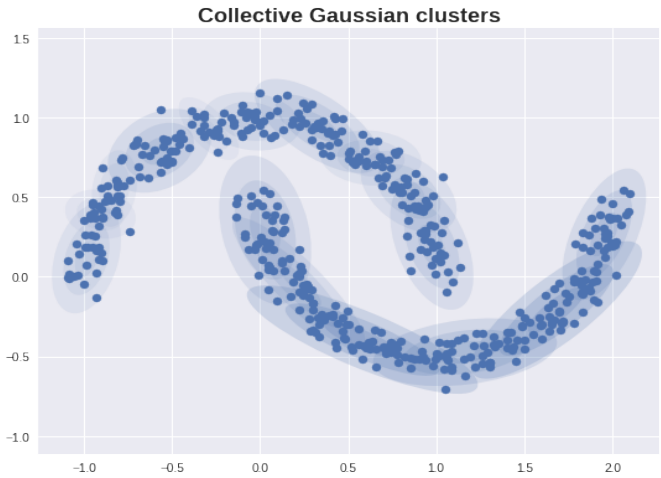


Figure 2.

6. Conclusion

After this experiments we understood how the expectation maximization algorithm works and how Gausian mixture models works.We also realized that how GMM can used as a tool for Density estimator