Performance of Gaussian Mixture Model on Woman’s Fertility Dataset

Shrikant patro,

School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127

Email:shrikantjagannath.2018@vitstudent.ac.in

Abstract

In real world situation k-means is no suitable due to its practical challenge faced due to non-probabilistic nature of k-means and its use of simple distance from cluster center to assign cluster membership leads to poor performance for many real world application. Gaussian mixture model can be viewed as extension of the k-means algorithm suitable for real life problem. This paper will discuss in detail working of the Gaussian mixture model (GMM). The application of GMM on Woman Fertility dataset from UCI Repository.

Introduction

K-Means lack flexibility in cluster center shape probabilistic cluster assignment – mean that for many datasets, it may not perform as expectation. In order to measure weaknesses by generalizing k-means model. We can measure the uncertainty in cluster assignment by comparing the distance of each point to all cluster centers, rather than focusing at only one cluster center. The cluster boundaries can be made elliptical rather than circular. So, as to account non-circular clusters. It turned out these are two essential component of a different type of clustering model, Gaussian mixture model (GMM). A GMM attempt to find a mixture of multidimensional Gaussian probability distribution that best model any input dataset. GMM can be used in same manner as k-means is used in clustering. Because GMM contains a probabilistic model under the hood, it is also possible to find probabilistic cluster assignment. In scikit-learn, this is done by using the predict_proba method. This return a matrix of size [n_sample, n_cluster] which measure the probability that any point belong to cluster. We can visualize this uncertainty by making the size of each point proportional to the certainty of its prediction; precisely the point at the boundaries between clusters that reflect this uncertainty of cluster assignment.

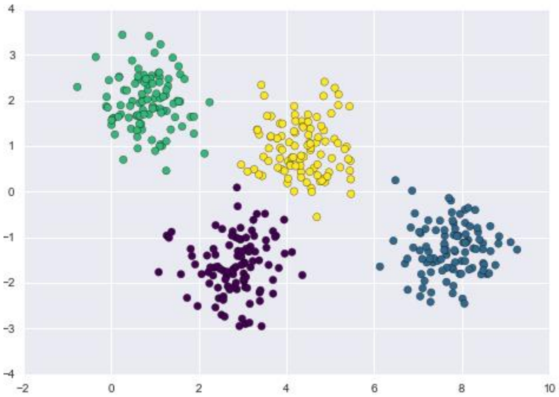

Fig1. Sample dataset

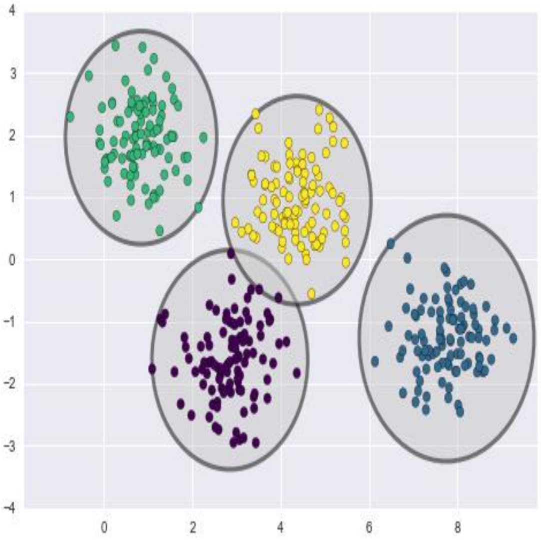

Fig 2. K-means algorithm performance

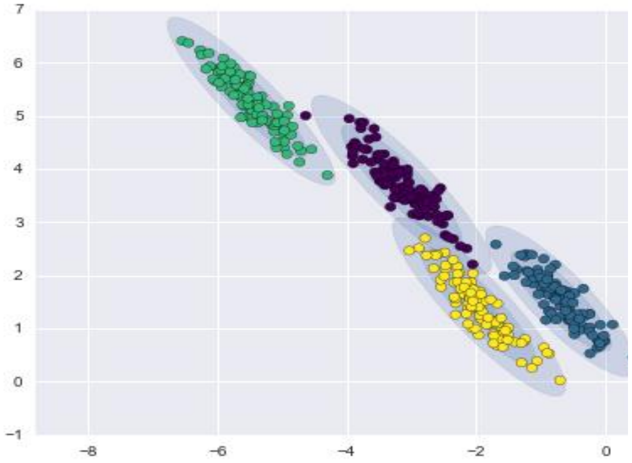

Fig 3. Gaussian mixture model performance

Methodology

Gaussian mixture model is very much similar to k-means. It uses as expectation maximization approach:

1. Choose starting gausses from the location and shape.
2. Repeat until converge:
    a. E-Step: for each point, find weights encoding the probability of membership in each cluster.
    b. 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 is that each cluster is associated not with a hard-edged sphere, but with a smooth Gaussian model. Just as in the *k*-means expectation–maximization approach, this algorithm can sometimes miss the globally optimal solution, and thus in practice multiple random initializations are used.

Choosing the convergence type; We will see that the convergence_type option was set differently within each. This hyperparameter controls the degrees of freedom in the shape of each cluster; it is essential to set this carefully for any given problem. The default is convariance _type=”diag”, which means that the size of the cluster along each dimension can be set independently, with the resulting ellipse constrained to align with the axes. A slightly simpler and faster model is covariance_type=”spherical”, which constrains the shape of the cluster such that all dimensions are equal. The resulting clustering will have similar characteristics to that of *k*-means, though it is not entirely equivalent. A more complicated and computationally expensive model (especially as the number of dimensions grows) is to use covariance_type=full, which allows each cluster to be modeled as an ellipse with arbitrary orientation.

Experiment

To implement Gaussian mixture model(GMM) as Expectation maximization technique. The import statement must include GMM package from the sklearn. mixture. Specify the component using the n_component is set to ‘n’. The call the fit method from the constructed model. This fit method result into formation of ‘n’ GMM among the data point. We can also specify the covariance_type and random state.

Conclusion

Using the Gaussian mixture model the performance of the Expectation maximization can be improved. It overcome the drawback of the k-means algorithm and makes EM technique more suitable for the real life problem.