GMM : Gaussian Mixture Model

Kenneth EZUKWOKE
Department of Mathematics and industrial Engineering
Ecole Nationale Superieure des Mines de Saint-Etienne (ENSM-SE)
{ifeanyi.ezukwoke}@emse.fr

Abstract
This paper presents a brief introduction to Gaussian Mixture Model (GMM) together with its mathematical formulation - a probabilistic (soft) clustering algorithm for unsupervised modeling.

Introduction

Machine learning is a subset of artificial intelligence and ad is divided into supervised, unsupervised and semi-supervised learning. Clustering is categorized under unsupervised learning and is used to extract structure from an unstructured data set with no prior knowledge of the underlying distribution of the data set. Clustering can further be splitted into different categories including hard clustering (KMeans, spectral clustering), density based clustering (Local outlier factor, DBSCAN, HDBSCAN, OPTICS) and soft clustering (soft KMeans, Guassian Mixture Model).

Formal definiition

Guassian Mixture Model is a linear superposition of Guassian components used for soft clustering of data containing mixtures of Guassians. As a probabilistic framework this implies that it has flexibility of assigning assignment to clusters as opposed to hard assignments used by KMeans. This means that a data point can be generated by any distribution in the mixtures of observed Guassians with some probability. Furthermore, a distribution has some *responsibility* for generating a particular observation (usually from a latent unobserved distribution).
[1]

Notation and Derivations

We employ the following notations to describe the Guassian mixture model

- X : input space or the training datat set which follows a multivariate normal distribution $\mathcal{N}$
- The dimension of the matrix X is $\mathbb{R}^{n\times m}$, where n is the instances and m is the dimension or variable span.

1. code available on github

— $Z \in \mathbb{R}^{n \times k}$ is the latent variable which follows a categorical distribution with latent points $z_n \in \{0, 1\}$ where $\sum_k z_{k,n} = 1$
— π is the weight or mixing probability which defines the size of a Guassian
— K represent the numbers of Guassians present in X we intend to model.
— An observation in X is represented as $\{(x_n)\}_{n=i}^{N}$ where $x_n \subset X$.
— A row vector can be represented as $[x_n, \ldots, x_N]^T$ and the column vector is represented as $[x_n, \ldots, x_m]$.
— μ and Σ represent the mean and covariance matrix respectively. The mean defines the center of a Guassian while covariance defines the dimension of the ellipsoid.
— θ represents the Guassian parameter pairs $\{\pi, \mu, \Sigma\}$

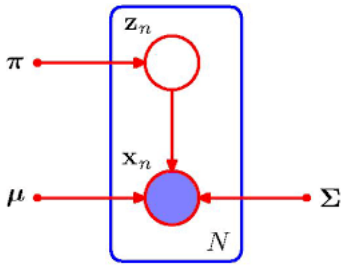

FIGURE 1 – Bayesian network for GMM represented using plate notation

Using the plate notation we can make the following deductions about the Bayesian network

$$\begin{aligned} \text{Marginal distribution :} \quad & p(z_n) \sim Categorical(\pi) \\ \text{Conditional distribution :} \quad & p(x_n|z_n = k) \sim \mathcal{N}(x_n|\mu_k, \Sigma_k) \\ \text{Mixture/joint distribution :} \quad & p(x_n, z_n) \sim p(z_n)p(x_n|z_n) \end{aligned}$$

If $k = 1$, the marginal distrubution $p(z_n = 1) = \pi_k$ where π_k satisfies the constraints $0 \leq \pi_k \leq 1$ and $\sum_k^K \pi_k = 1$; and for K numbers of components,

$$p(z_n) = \Pi_{k=1}^{K} \pi_k^{z_{k,n}} \tag{1}$$

as z uses one of the K. Hence, z_n can only be a binary vector with values $[0, \ldots, 1, \ldots, 0]^T \in \mathbb{R}^k$. The implication of this is that the overall probability of observing a point that comes from form a k Guassian is equivalent to mixing the coefficients of the Guassian. We also know that the mixture distribution can be written as a marginal distribution over z

$$p(x_n) = \sum_{k=1}^{K} p(x_n, z_n = k) \tag{2}$$

$$= \sum_{k=1}^{K} p(z_n = k)p(x_n|z_n = k) \tag{3}$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \tag{4}$$

This suffices that for every observation x_n, their is a corresponding z_n latent varaible (Figure 1). Hence we can chose the joint distribution $p(x_n, z_n)$ over marginal distribution $p(x)$ for maximum likelihood estilmation of the parameters θ. This implies that we evaluate the expectation of the complete data likelihood $p(x_n, z_n|\theta)$ with respect to posterior $p(z_n|x_n, \theta)$. Since z_n does not depend on μ_k, Σ_k, we can rewrite the posterior probability as $p(z_n = k|x_n)$.
Using Bayes rule, we can evalaute $p(z_n = k|x_n)$ as follows

$$p(z_n = k|x_n) = \frac{p(x_n|z_n = k)p(z_n = k)}{p(x_n)} \tag{5}$$

$$p(z_n = k|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} = r_{k,n} \tag{6}$$

$r_{k,n}$ is the responsibility on the observation x_n by the k-th component. Concretely, we require the condition $p(x_n|z_n)$ in order to evaluate $p(x_n, z_n)$. However, we know that the conditional probability $p(x_n|z_n = k, \theta)$ follows a normal distribution $\mathcal{N}(x_n|\mu_k, \Sigma_k)$ for one component and $p(x_n|z_n, \theta)$ is therefore $\Pi_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{k,n}}$ for K components Guassian. Also note that the complete data likelihood for K components is expressed as :

$$p(x_n, z_n|\theta) = p(z_n|\theta)p(x_n|z_n, \theta) \tag{7}$$

$$= \prod_{k=1}^{K} \pi_k^{z_{k,n}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{k,n}} \tag{8}$$

where,

$$\mathcal{N}(x_n|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{m/2}|\Sigma_k|^{1/2}} \exp(-\frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k)) \tag{9}$$

Eqution 9 is a multivariate normal distribution with μ_k and Σ_k parameters.

Expectation Maximation (EM) Algorithm

Expectation–maximization (EM) algorithm is an iterative method to find maximum likelihood estimates of parameters of a Guassian model. Its is an iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the $\{\theta\}$, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step.

The main steps of an EM algorithm is the maximum likelihood estimate (MLE) of the unknown parameters determined by maximizing the marginal likelihood of the observed data. The algorithm can be summarized as follows :

Step 1 (Initialization) Initiatize parameters $\{\theta\}$; the result of KMeans algorithm is sometimes used for θ.

Step 2 (Expectation) estimate the responsibilities $r_{k,n}$ using the initial or current values of the parameters. Recall the responsiblility is given by

$$r_{k,n} = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \tag{10}$$

Step 3 (Maximization) find the parameters that maximizes Q. This essentially means re-estimating the parameters of Q using current responsibility $r_{n,k}$
We find current estimates of the parameter $\{\theta\}$ by evaluating the expected value of the log-likelihood of θ with respect to the conditional distribution of Z given X

$$\mathcal{Q}(\theta|\theta^t) = \mathop{\mathbb{E}}_{z|x,\theta^t}[\ln \prod_{n=1}^{N} p(x_n, z_n|\theta^t)] = \mathop{\mathbb{E}}_{z|x,\theta^t}[\sum_{n=1}^{N} \ln p(x_n, z_n|\theta^t)] \tag{11}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_n = k|x_n, \theta^t) \ln p(x_n, z_n|\theta^t) \tag{12}$$

Substituting the Equations (6) and (8) into (11) we have that,

$$\mathcal{Q}(\theta|\theta^t) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{k,n} \ln[\pi_k^{z_{k,n}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{k,n}}] \tag{13}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{k,n} \cdot z_{k,n} \ln[\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)] \tag{14}$$

Since $z_{k,n}$ has an binary expectation of 1 given x_n, that is, the summation of all latent varibles over a component should be 1 from the previous constraint we earlier stated ($\sum_k^K \pi_k = 1$). Hence we can easily set $z_{k,n} = 1$ and maximizing the log-likelihood with respect to π_k by introducing a lagrangian multiplier

$$\mathcal{Q}(\theta|\theta^t, \lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{k,n} \ln[\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)] + \lambda(1 - \sum_{k=1}^{K} \pi_k) \tag{15}$$

$$\mathcal{Q}(\theta|\{\pi_k, \mu_k, \Sigma_k\}^t, \lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{k,n}[\ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)] + \lambda(1 - \sum_{k=1}^{K} \pi_k) \tag{16}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{k,n}[\ln \pi_k - \frac{m}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2}(x_n - \mu_k)^T |\Sigma_k|^{-1}(x_n - \mu_k)] + \lambda(1 - \sum_{k=1}^{K} \pi_k) \tag{17}$$

We find the parameters by differentiating $w.r.t$ the individual parameters

$$\frac{\partial[\mathcal{Q}(\theta|\{\pi_k, \mu_k, \Sigma_k\}^t, \lambda)]}{\partial \mu_k} = -\sum_{n=1}^{N} r_{k,n} \Sigma_k^{-1}(x_n - \mu_k) = 0 \tag{18}$$

$$\mu_k = \frac{\sum_{n=1}^{N} r_{k,n} x_n}{\sum_{n=1}^{N} r_{k,n}} \tag{19}$$

If we set $N_k = \sum_{n=1}^{N} r_{k,n}$ the number of points assigned to the cluster, the mean takes the form of

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{k,n} x_n \tag{20}$$

We can then conclude that the mean of the *k-th* component is the weighted mean (*weighted by the responsibilities or posterior probability*) of all the points in the data set. Similarly,

$$\frac{\partial[\mathcal{Q}(\theta|\{\pi_k,\mu_k,\Sigma_k\}^t,\lambda)]}{\partial\Sigma_k} = \sum_{n=1}^{N} r_{k,n}\frac{1}{2}(\Sigma_k - (x_n-\mu_k)(x_n-\mu_k)^T) = 0 \tag{21}$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{k,n}(x_n-\mu_k)(x_n-\mu_k)^T \tag{22}$$

$$\frac{\partial[\mathcal{Q}(\theta|\{\pi_k,\mu_k,\Sigma_k\}^t,\lambda)]}{\partial\pi_k} = \sum_{n=1}^{N} r_{k,n}\frac{1}{\pi_k} - \lambda = 0 \tag{23}$$

$$\pi_k = \frac{\sum_{n=1}^{N} r_{k,n}}{\lambda} \tag{24}$$

Since the $\sum_{k=1}^{K}\pi_k = 1$, λ is equal to N for this condition to be satisfied. Hence,

$$\pi_k = \frac{\sum_{n=1}^{N} r_{k,n}}{N} = \frac{N_k}{N} \tag{25}$$

Given the new estimate of θ^t, we can then evaluate subsequent parameters of the log-likelihood of the complete data, until the algorithm converges.

$$\theta^{t+1} = \underset{\theta^t}{argmax}\,\mathcal{Q}(\theta,\theta^t) \tag{26}$$

Step 4 (Test of Convergence) We check if the algorithm converges at a local maxima when

$$Q(\theta^{t+1}|\theta^t) - Q(\theta^t|\theta^{t-1}) \leq \epsilon \tag{27}$$

Where ϵ is the threshold value for which we expect stability or convergence.