University of Illinois Fall 2018

ECE 544NA: Pattern Recognition
Lecture 17: Gaussian Mixture Model (GMM)

Lecturer: Alexander Schwing Scribe: Anay Pattanaik

1 Motivation and Goals

- In the last lecture, we learnt about K-means, but it results in hard cluster assignment to datapoints. Thus, K-means may not work when clusters overlap as shown in Fig. 1 . Gaussian Mixture Model (GMM) alleviates this issue through a probabilistic approach to clustering where each datapoint has "soft/probabilistic" assignment to each cluster.
- We will learn that the optimization of GMM doesn't have a closed form solution. Hence, we will use alternative optimization.
- We will learn that GMM is a generalization of K-means.

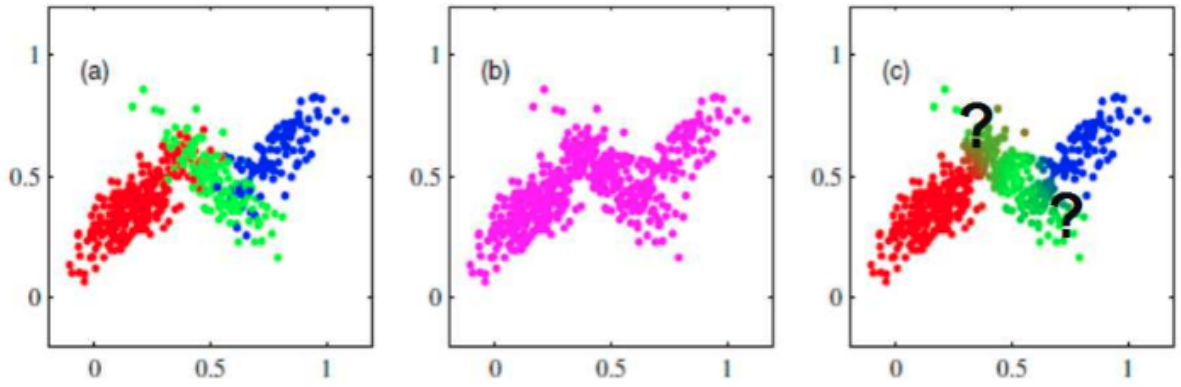

Figure 1: Failure of K-Means when there is overlap amongst clusters. GMM is more suited for this situation as the boundary points should have "soft" assignment. [6]

2 Discriminative and Generative Modelling

2.1 Linear Regression (Discriminative)

Recall that in linear regression (discriminative) model, we are interested in modelling $p(y^i|x^i)$ where y^i are the values corresponding to datapoint x^i. A probabilistic approach that we discussed was by modelling y^i as Gaussian distribution with mean given by $\mathbf{W}^T\phi(x^i)$. Mathematically,

$$p(y^i|x^i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(y^i - \mathbf{W}^T\phi(x^i))^2}{2\sigma^2}\} \tag{1}$$

2.2 Clustering (Generative)

In generative model, we are interested in modelling the distribution of datapoints x^i itself and not necessarily the class labels y^i conditional on datapoint x^i (discriminative method). One might model the data being generated by a Gaussian distribution with mean μ and standard deviation σ. Mathematically,

$$p(x^i|\mu,\sigma) = \mathcal{N}(x^i|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(y^i-\mu)^2}{2\sigma^2}\} \tag{2}$$

3 A Naive Generative Approach to Clustering (Single Gaussian)

As discussed in Subsection 2.2, the simplest data generation mechasim can just be Gaussian. Thus, the learning problem boils down to: *Given a dataset $\mathcal{D} = (x^i)$, how to find parameters ($\theta = (\mu,\sigma)$) of "data generating" Gaussian distribution* . We turn to the standard approach of minimizing the negaive log-likelihood (NLL) of data. This is summarized in the following program:

Program:

$$\begin{aligned}\underset{\mu,\sigma}{\text{minimize}} \quad NLL &:= \underset{\mu,\sigma}{\text{minimize}} - log \prod_{i\in\mathcal{D}} p(x^i|\mu,\sigma) \\ &= \underset{\mu,\sigma}{\text{minimize}} \sum_{i\in\mathcal{D}} \frac{(x^i-\mu)^2}{2\sigma^2} + \frac{N}{2} log(2\pi\sigma^2)\end{aligned} \tag{3}$$

Here, we used the data generation mechanism of Eq. 2. Let us see if we have a closed form solution of above program for clustering using single Gaussian.
Optimality Condition:

- Critical point with respect to μ.

$$\begin{aligned}\frac{\partial NLL}{\partial \mu} &= 0 \\ \frac{\partial \frac{\sum_{i\in\mathcal{D}}(x^i-\mu)^2 + \frac{N}{2}log(2\pi\sigma^2)}{\sigma^2}}{\partial \mu} &= 0 \quad \text{(Using Eq. 3)} \\ \frac{\sum_{i\in\mathcal{D}} 2(x^i-\mu)}{\sigma^2} &= 0 \\ \mu &= \frac{\sum_{i\in\mathcal{D}} x^i}{N}\end{aligned} \tag{4, 5}$$

- Critical point with respect to σ

$$\begin{aligned}\frac{\partial NLL}{\partial \sigma} &= 0 \\ \frac{\partial \frac{\sum_{i\in\mathcal{D}}(x^i-\mu)^2 + \frac{N}{2}log(2\pi\sigma^2)}{\sigma^2}}{\partial \sigma} &= 0 \quad \text{(Using Eq. 3)} \\ \frac{-\sum_{i\in\mathcal{D}}(x^i-\mu)^2}{\sigma^3} + \frac{N}{\sigma} &= 0 \\ \sigma^2 &= \frac{\sum_{i\in\mathcal{D}}(x^i-\mu)^2}{N}\end{aligned} \tag{6}$$

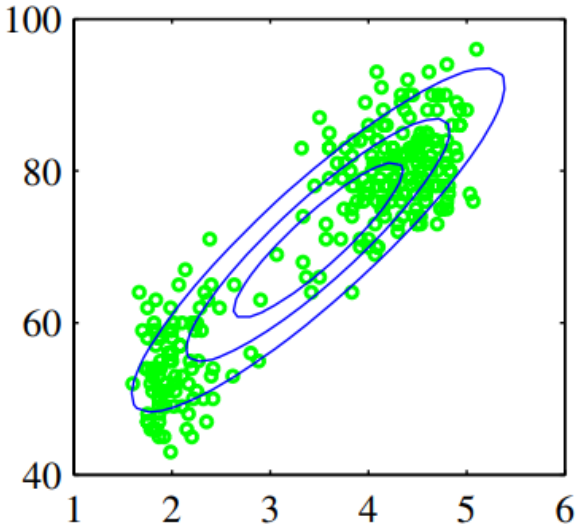

Figure 2: Single Gaussian trying to capture bimodal data distribution
[5]

Great! So, we found closed form solution of the parameters that we set out to learn (μ, σ) in Eq. 5 and 6. Unfortunately, a single Gaussian distribution is not "general enough" as shown in Fig. 2. Can we do better? Yes! An improvement could be to consider the data being generated by a mixture of Gaussian distributions.

4 From Single Gaussian to a Mixture of Gaussian

4.1 Superposition of Gaussians

To fix the limitation of a single Gaussian distribution, we can consider the data to be generated by a linear superpositions of several Gaussian distributions. Mathematically,

$$\begin{aligned} p(x^i|\pi, \mu, \sigma) &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x^i|\mu_k, \sigma_k) \\ \text{such that} \quad \sum_{k=1}^{K} \pi_k &= 1; \pi_k \geq 0 \end{aligned} \tag{7}$$

In other words, each datapoint is "composed" of several Gaussian distributions, each with mean μ_k and variance σ_k^2. Here, π is constrained to be a probability distribution and encodes "weightage" of different clusters. A data generated with degenerate π will basically correspond to data generated by a single cluseter. On the other hand, if π is uniform, this means that each cluster contributes equally to the data generation process.

So, the learning problem boils down to *Given a dataset $\mathcal{D} = \{(x^i)\}$, how to find parameters $(\theta = (\pi, \mu, \sigma))$ of "data generating" mixture of Gaussian distributions.* We turn to our usual approach of minimization of NLL. This can be stated as follows:

Program:

$$\begin{aligned} \underset{\pi,\mu,\sigma}{\text{minimize}} \quad & NLL := -log \prod_{i \in \mathcal{D}} p(x^i|\pi,\mu,\sigma) = -\sum_{i \in \mathcal{D}} log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^i|\mu_k,\sigma_k) \quad \text{(using Eq. 7)} \\ \text{subject to} \quad & \sum_{k=1}^{K} \pi_k = 1 \\ & \pi_k \geq 0 \end{aligned} \tag{8}$$

We can optimize the above objective function with lagrange multiplier corresponding to π_k. However, we won't obtain a closed form solution unlike single Gaussian case. An alternative perspective with introduction of a latent/hidden variable will be illuminating. We consider this in next subsection.

4.2 Alternative Perspective with Latent Variable

In this subsection, we introduce hidden (latent) variable (Z) to the problem. This will result in the same optimization problem but will give elegant solution. Each datapoint is assigned latent variable z. z denotes the cluster assignment for generative process, with each of its component being $z_{ik} \in \{0, 1\}$ that signifies assignment of cluster k to datapoint i. Probabilistically, we can write down

Marginal for z_{ik}:

$$p(z_{ik} = 1) = \pi_k \quad \text{(Prior of cluster assignment)} \tag{9}$$

$$p(\mathbf{z}_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}}; \quad \mathbf{z}_i = [z_{i1}, ..., z_{iK}]^T \tag{10}$$

Now, given the latent variable Z, we can find the conditional distribution of each datapoint.

Conditional for $x^i|\mathbf{z}_i$:

$$p(x^i|\mathbf{z}_i) = p(x^i|\mathbf{z}_{ik} = 1) = \mathcal{N}(x^i|\mu_k,\sigma_k)^{z_{ik}} \tag{11}$$

Given conditional generation of datapoint by each cluster, we can find the total distribution for datapoint x^i by marginalizing over all clusters as follows.

Marginal for x^i

$$\begin{aligned} p(x^i|\pi,\mu,\sigma) &= \sum_{\mathbf{z}_i} p(x^i|\mathbf{z}_i) p(\mathbf{z}_i) \\ &= \sum_{\mathbf{z}_i} \mathcal{N}(x^i|\mu_k,\sigma_k)^{z_{ik}} \prod_{k=1}^{K} \pi_k^{z_{ik}} \quad \text{(using Eqs. 10 and 11)} \\ &= \sum_{k=1}^{K} \mathcal{N}(x^i|\mu_k,\sigma_k)\pi_k \end{aligned} \tag{12}$$

Now, given the marginal distribution for x^i, we can calculate the posterior of assignment to cluster k for datapoint i (also called responsibility) and denote it by r_{ik} as follows.

Posterior of cluster assignment r_{ik}:

$$\begin{aligned} r_{ik} = p(\mathbf{z}_i|x^i) &= p(z_{ik}=1|x^i) \\ &= \frac{p(z_{ik}=1)p(x^i|z_{ik}=1)}{\sum_{\hat{k}=1}^{K} p(z_{i\hat{k}}=1)p(x^i|z_{i\hat{k}}=1)} \quad \text{(Bayes' rule)} \qquad (13) \\ &= \frac{\pi_k \mathcal{N}(x^i|\mu_k,\sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}},\sigma_{\hat{k}})} \quad \text{(Using Eqs. 9, 11, 12)} \qquad (14) \end{aligned}$$

With all of these, we again consider the program that we want to solve in Prog. 8, that is minimization of NLL and we restate it below.

$$\begin{aligned} \underset{\pi,\mu,\sigma}{\text{minimize}} \quad & NLL := -log \prod_{i\in\mathcal{D}} p(x^i|\pi,\mu,\sigma) = -\sum_{i\in\mathcal{D}} log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^i|\mu_k,\sigma_k) \\ \text{subject to} \quad & \sum_{k=1}^{K} \pi_k = 1 \\ & \pi_k \geq 0 \end{aligned}$$

Optimization:
Additionally, let us define the "effective" number of examples assigned to k-th Gaussian cluster as

$$N_k = \sum_{i\in\mathcal{D}} r_{ik} \qquad (15)$$

- Critical point with respect to μ_k

$$\begin{aligned} \frac{\partial NLL}{\partial \mu_k} &= 0 \\ -\frac{\partial \sum_{i\in\mathcal{D}} log \sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}},\sigma_{\hat{k}})}{\partial \mu_k} &= 0 \quad \text{(Using Eq. 8)} \\ \sum_{i\in\mathcal{D}} \frac{\frac{\pi_k \partial \mathcal{N}(x^i|\mu_k,\sigma_k)}{\partial \mu_k}}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}},\sigma_{\hat{k}})} &= 0 \\ \sum_{i\in\mathcal{D}} \frac{\pi_k \mathcal{N}(x^i|\mu_k,\sigma_k) \frac{\partial \frac{(x^i-\mu_k)^2}{2\sigma_k^2}}{\partial \mu_k}}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}},\sigma_{\hat{k}})} &= 0 \\ \sum_{i\in\mathcal{D}} r_{ik}(-\frac{x^i-\mu_k}{\sigma_k^2}) &= 0 \quad \text{(Using Eq. 14)} \\ \mu_k &= \frac{1}{N_k} \sum_{i\in\mathcal{D}} r_{ik} x^i \quad \text{(Using Eq. 15)} \qquad (16) \end{aligned}$$

- Critical point with respect to σ_k

$$\frac{\partial NLL}{\partial \sigma_k} = 0$$

$$-\frac{\partial \sum_{i\in\mathcal{D}} log \sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sigma_{\hat{k}})}{\partial \sigma_k} = 0 \quad \text{(Using Eq. 8)}$$

$$\sum_{i\in\mathcal{D}} \frac{\frac{\pi_k \partial \mathcal{N}(x^i|\mu_k,\sigma_k)}{\partial \sigma_k}}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sigma_{\hat{k}})} = 0$$

$$\sum_{i\in\mathcal{D}} \frac{\pi_k \mathcal{N}(x^i|\mu_k,\sigma_k)(\frac{1}{\sigma_k} - \frac{(x^i-\mu_k)^2}{\sigma_k^2})}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sigma_{\hat{k}})} = 0$$

$$\sum_{i\in\mathcal{D}} r_{ik}(\frac{1}{\sigma_k} - \frac{(x^i-\mu_k)^2}{\sigma_k^2}) = 0 \quad \text{(Using Eq. 14)}$$

$$\sigma_k = \frac{1}{N_k} \sum_{i\in\mathcal{D}} r_{ik}(x^i - \mu_k)^2 \quad \text{(Using Eq. 15)} \tag{17}$$

- Critical point with respect to π_k (add Langrange multiplier to NLL for constraint on π_k)

$$\frac{\partial NLL + \lambda(\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} - 1)}{\partial \pi_k} = 0$$

$$\frac{\partial \sum_{i\in\mathcal{D}} log \sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sigma_{\hat{k}}) + \lambda(\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} - 1)}{\partial \pi_k} = 0 \quad \text{(Using Eq. 8)}$$

$$\sum_{i\in\mathcal{D}} \frac{\mathcal{N}(x^i|\mu_k,\sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sigma_{\hat{k}})} + \lambda = 0 \tag{18}$$

$$\sum_{i\in\mathcal{D}} \frac{\sum_{k=1}^{K} \pi_k \mathcal{N}(x^i|\mu_k,\sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sigma_{\hat{k}})} + \lambda = 0 \quad \text{(Multilpy by } \pi_k \text{ and add)}$$

$$\lambda = -N \tag{19}$$

$$\sum_{i\in\mathcal{D}} \frac{\mathcal{N}(x^i|\mu_k,\sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sigma_{\hat{k}})} - N = 0 \quad \text{(Using Eq. 19 in Eq. 18)}$$

$$\sum_{i\in\mathcal{D}} \frac{\pi_k \mathcal{N}(x^i|\mu_k,\sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sigma_{\hat{k}})} - \pi_k N = 0 \quad \text{(Multiplying by } \pi_k)$$

$$\sum_{i\in\mathcal{D}} r_{ik} - \pi_k N = 0 \quad \text{(Using Eq. 14)}$$

$$\pi_k = \frac{N_k}{N} \tag{20}$$

Note that the update equations of parameters μ, σ, π in Eqs.16, 17, and 20 are not closed form. Thus, we need to resort to an alternating procedure to find the solution. This is summarized in the following algorithm (Expectation Minimization for GMM).

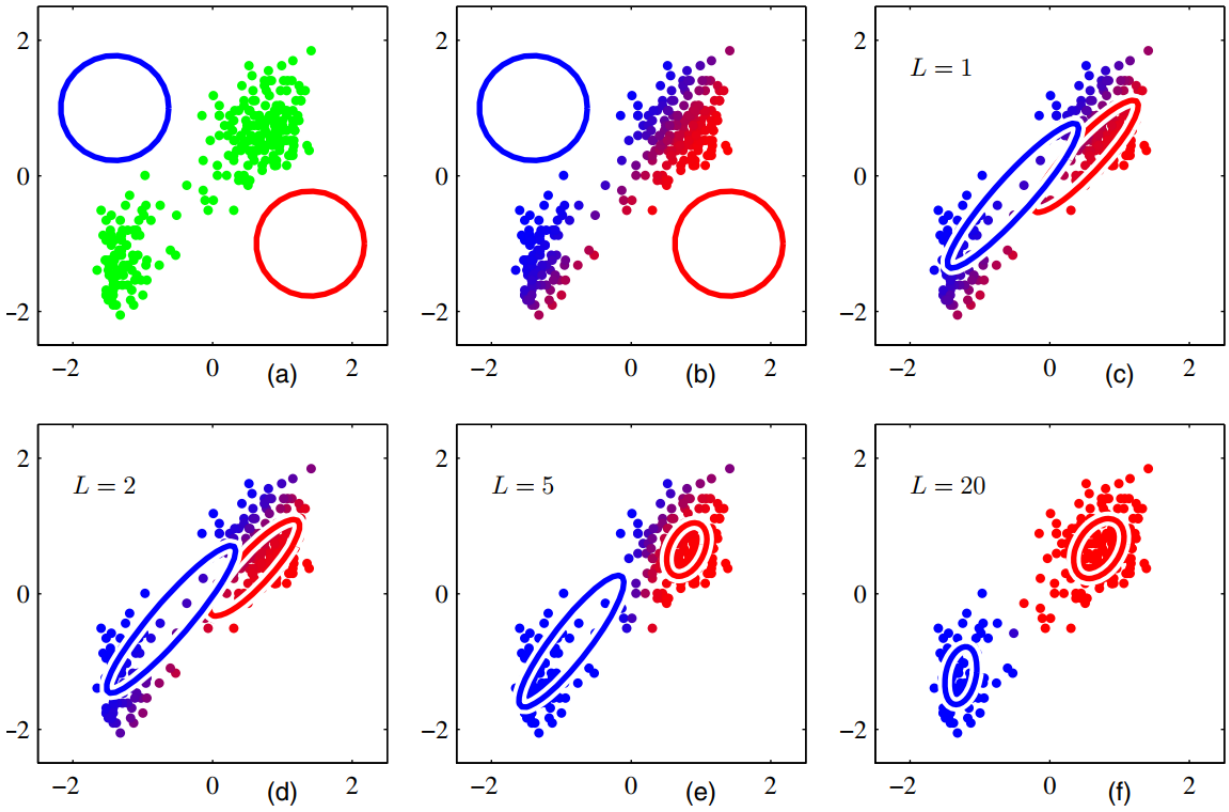

Figure 3: EM algorithm for GMM in action! L represents the iteration number. [1]

EM for GMM

1. Initialize π, μ, σ
2. E-step Update

$$r_{ik} = \frac{\pi_k \mathcal{N}(x^i|\mu_k, \sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sigma_{\hat{k}})} \quad \text{(Eq. 14)}$$

3. M step: Update

$$N_k = \sum_{i \in \mathcal{D}} r_{ik} \quad \text{(Eq. 15)}$$

$$\mu_k = \frac{1}{N_k} \sum_{i \in \mathcal{D}} r_{ik} x^i \quad \text{(Eq. 16)}$$

$$\sigma_k = \frac{1}{N_k} \sum_{i \in \mathcal{D}} r_{ik} (x^i - \mu_k)^2 \quad \text{(Eq. 17)}$$

$$\pi_k = \frac{N_k}{N} \quad \text{(Eq. 20)}$$

We can see how GMM optimizes alternatively to finally converge to the desired cluster in Fig. 3

5 K-Means as a limiting case of GMM

In this section, we will show that K-Means is in fact a special case of GMM when the variance of each cluster centre approaches 0. To deduce this fact mathematically, let us fix the variance of each cluster in GMM, that is, $\sigma_k^2 = \epsilon, \forall k$ (for all clusters). Then the responsibilities r_{ik} can be expressed as

$$\begin{aligned} r_{ik} &= \frac{\pi_k \mathcal{N}(x^i|\mu_k, \sqrt{\epsilon})}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^i|\mu_{\hat{k}}, \sqrt{\epsilon})} \quad \text{(Eq. 14)} \\ &= \frac{\pi_k exp(-\frac{1}{2\epsilon}(x^i - \mu_k)^2)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} exp(-\frac{1}{2\epsilon}(x^i - \mu_{\hat{k}})^2)} \\ \lim_{\epsilon \to 0} r_{ik} &= \begin{cases} \frac{\pi_k exp(-\frac{1}{2\epsilon}(x^i - \mu_k)^2)}{\pi_k exp(-\frac{1}{2\epsilon}(x^i - \mu_k)^2)}; \text{ if } k = \text{argmin}_l \, (x^i - \mu_l)^2 \\ 0; \text{ otherwise} \end{cases} \\ \lim_{\epsilon \to 0} r_{ik} &= \begin{cases} 1; \text{ if } k = \text{argmin}_l \, (x^i - \mu_l)^2 \\ 0; \text{ otherwise} \end{cases} \end{aligned} \tag{21}$$

In Eq. 21, the responsibilities go to 0, except the for the cluster whose distance from x^i is least. Thus, the datapoint x^i is assigned to closest cluster. Similarly, we can observe that the mean assignment of each cluster is

$$\begin{aligned} \mu_k &= \frac{\sum_{i \in \mathcal{D}} r_{ik} x^i}{N_k} \quad \text{(Eq. 16)} \\ &= \frac{\sum_{i \in \mathcal{D}_k} x^i}{N_k} \quad \text{(Using Eq. 21)} \end{aligned} \tag{22}$$

Here, $\mathcal{D}_k$ is the set of datapoints assigned to cluster k. Thus, from Eq. 22, we can see that the update for cluster centre is exactly same as K-Means.
In summary, Eq. 21 and Eq. 22 updates are exactly same as K-Means update.

6 Applications

- GMM has been used in astronomy for pulsar "classification" [3]
- GMM has been used in image segmentation (for bacterial colonies) [4]
- GMM has been used for "classification" of seismic activities [2]

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] H. Kuyuk, E. Yildirim, E. Dogan, and G. Horasan. Application of k-means and gaussian mixture model for classification of seismic activities in istanbul. *Nonlinear Processes in Geophysics*, 19(4):411–419, 2012.

[3] K. Lee, L. Guillemot, Y. Yue, M. Kramer, and D. Champion. Application of the gaussian mixture model in pulsar astronomy-pulsar classification and candidates ranking for the fermi 2fgl catalogue. *Monthly Notices of the Royal Astronomical Society*, 424(4):2832–2840, 2012.

[4] I. S. Maretić and I. Lacković. Application of gaussian mixture models with expectation maximization in bacterial colonies image segmentation for automated counting and identification. In *XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013*, pages 388–391. Springer, 2014.

[5] A. G. Schwing. Pattern recognition. Lecture Notes, ECE 544NA, 2018.

[6] A. Soni. Clustering with gaussian mixture model clustering with gaussian mixture model medium, Dec 2017.