[Lecture] Gaussian Mixture Model (GMM)

Jae Yun JUN KIM*

August 20, 2019

Reference: Andrew Ng's lecture notes.

1 Motivational example: Density function estimation

Imagine that there is an aircraft engine company. Let us suppose that the following dataset tell some important properties of engines manufactured in this company.

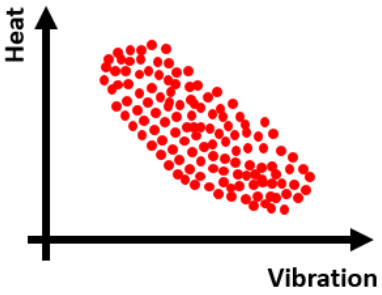
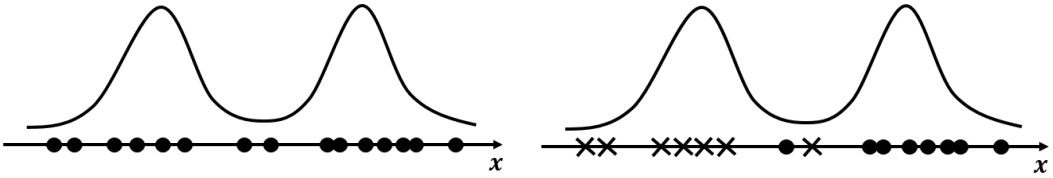


Figure 1: Motivation

What we want is to estimate the joint distribution of the amount of vibration and of heat produced because, for instance, we want to detect outliers (i.e., those engines that are away from the distribution of points illustrated in Figure 1).

Hence, for a given dataset, we would like to build a model $P(x)$. When the value of the model $P(x)$ of an engine (with certain heat and vibration properties) is small, then this engine can be considered as anomalous.



(a) Unsupervised learning problem (b) Supervised learning problem

Figure 2: Comparison between supervised and unsupervised learning

Consider the unlabeled dataset in Figure 2(a) and assume that this dataset is generated from the sum of two Gaussian distributions.

Just to be clear, a picture that we have in mind is that we are visioning that there are two separate Gaussians that generate this unlabeled dataset in Figure 2(a). But, if only we knew

*ECE Paris Graduate School of Engineering, 37 quai de Grenelle 75015 Paris, France; jae-yun.jun-kim@ece.fr

where the Gaussians were, then we can trace their distributions as Figure 2(b) and sum them up to get the overall distribution.

But, in fact, we do not have access to these labels, and, therefore, we do not know from which of the two Gaussians the data came from. So, we would like to come up with an algorithm to fit this Gaussian Mixture model, even I do not know which of the two Gaussians the data came from.

2 Intution

Assume that, in the model, there is a latent (hidden/unobserved) random variable z. Assume also that $x^{(i)}$, $z^{(i)}$ have a joint distribution

$$P\left(x^{(i)}, z^{(i)}\right) = P\left(x^{(i)}|z^{(i)}\right) P\left(z^{(i)}\right). \tag{1}$$

Assume that $z^{(i)} \sim \text{Multinomial}(\phi)$ (if 2-Gaussians, then $z^{(i)} \sim \text{Bernoulli}(\phi)$) with

$$\phi_j \geq 0, \qquad \sum_{j=1}^{J} \phi_j = 1. \tag{2}$$

Assume also that the distribution of $x^{(i)}$ for given $z^{(i)} = j$ is

$$\left(x^{(i)}|z^{(i)} = j\right) \sim \mathcal{N}(\mu_j, \Sigma_j). \tag{3}$$

Recall that these equations are very similar to those of the *Gaussian Discriminant Analysis(GDA)* [1]. The difference is that for the GDA we used $y^{(i)}$ (instead of $z^{(i)}$, where $y^{(i)}$ are known for all i, while $z^{(i)}$ are unknown. More specifically, to relate the GMM to the GDA, if we knew $z^{(i)}$'s (which is not the case), then we could compute the log-likelihood of its parameters as

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{I} \log P\left(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma\right), \tag{4}$$

and could find the optimal parameters as

$$\begin{aligned}
\phi_j &= \frac{1}{I} \sum_{i=1}^{I} \mathbb{I}\{z^{(i)} = j\}, \\
\mu_j &= \frac{\sum_{i=1}^{I} \mathbb{I}\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{I} \mathbb{I}\{z^{(i)} = j\}}, \\
\Sigma_j &= \frac{1}{I} \sum_{i=1}^{I} \left(x^{(i)} - \mu_j\right) \left(x^{(i)} - \mu_j\right)^T.
\end{aligned} \tag{5}$$

Hence, if we knew $z^{(i)}$, then you could use the *maximum likelihood estimation (MLE)* of parameters. But, in reality, we do not know the values of $z^{(i)}$ (since x is unlabeled dataset).

[1] Recall that the GDA is a generative model in the framework of the supervised learning, whereas the GMM is a generative model in the framework of the unsupervised learning.

Therefore, we write down a specific bootstrap procedure, with the idea to guess the value of $z^{(i)}$ using a model. Afterwards, we use the estimated $z^{(i)}$ to fit the parameters of the model. Then, we iterate these two steps until we have a better guess of $z^{(i)}$, while optimizing the model parameters.

This bootstrap procedure is known as the **Expectation-Maximization (EM)** algorithm.

3 Expectation-Maximization (EM) algorithm

The **Expectation-Maximization (EM)** algorithm consists of repeating the following two procedures until convergence:

- **E-step** (guess values of $z^{(i)}$):

$$\begin{aligned} w_j^{(i)} &= P\left(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma\right) \\ &= \frac{P\left(x^{(i)} | z^{(i)} = j\right) P\left(z^{(i)} = j\right)}{\sum_{l=1}^{k} P\left(x^{(i)} | z^{(i)} = l\right) P\left(z^{(i)} = l\right)} \\ &= \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right)\phi_j}{\sum_{l=1}^{k} \frac{1}{(2\pi)^{n/2}|\Sigma_l|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1}(x^{(i)} - \mu_l)\right)\phi_l}. \end{aligned} \tag{6}$$

- **M-step** (update the estimated parameters):

$$\begin{aligned} \phi_j &= \frac{1}{I}\sum_{i=1}^{I} w_j^{(i)}, \\ \mu_j &= \frac{\sum_{i=1}^{I} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{I} w_j^{(i)}}, \\ \Sigma_j &= \frac{\sum_{i=1}^{I} w_j^{(i)} \left(x^{(i)} - \mu_j\right)\left(x^{(i)} - \mu_j\right)^T}{\sum_{i=1}^{I} w_j^{(i)}}. \end{aligned} \tag{7}$$

Hence, we clearly see two differences with respect to the GDA algorithm:

- The GDA algorithm uses indicator functions to define the optimal parameters, while the EM algorithm uses "soft" weights instead.
- In the GDA algorithm, a same covariance matrix is used for all possible labels, while in the EM algorithm, a covariance matrix is defined per label j.

What we just saw is a specific case of the EM algorithm for the case of the GMM. Let us now look at a general view of the EM algorithm. Afterwards, we will justify how the above parameters are found using this general version of the EM algorithm.

4 General version of the EM algorithm

4.1 Jensen's inequality

Let f be a convex function (i.e., $f''(x) \geq 0$). Let x be a random variable. Then,

$$f\left(\mathbb{E}(x)\right) \leq \mathbb{E}\left(f(x)\right), \tag{8}$$

where $\mathbb{E}(\cdot)$ represents the expectation of a random variable.
Further, if $f''(x) > 0$ (i.e., f is strictly convex), then

$$\mathbb{E}(f(x)) = f(\mathbb{E}(x)) \Leftrightarrow x = \mathbb{E}(x) \quad \text{with probability 1.} \tag{9}$$

If $f''(x) \leq 0$ (i.e., f is concave), then

$$f\left(\mathbb{E}(x)\right) \geq \mathbb{E}\left(f(x)\right). \tag{10}$$

If $f''(x) < 0$ (i.e., f is strictly concave), then

$$\mathbb{E}(f(x)) = f(\mathbb{E}(x)) \Leftrightarrow x = \mathbb{E}(x) \quad \text{with probability 1.} \tag{11}$$

4.2 Intuition

Suppose that we have some model $P(x, z; \theta)$ and that we observe only x. We would like to maximize the log-likelihood of the parameters as

$$\ell(\theta) = \sum_{i=1}^{I} \log P\left(x^{(i)}; \theta\right) = \sum_{i=1}^{I} \log \sum_{z^{(i)}} P\left(x^{(i)}, z^{(i)}; \theta\right). \tag{12}$$

The EM algorithm can be viewed as performing the MLE of the above $\ell(\theta)$, which is complicated by the fact that $z^{(i)}$'s are unobserved. This issue can be overcome through the EM algorithm as follows:

- Initialize θ to $\theta^{(0)}$.
- Repeat until convergence
  - Construct a lower bound for this log-likelihood function for $\theta^{(t)}$.
  - Find the maximum value of the lower bound and the corresponding $\theta^{(t+1)}$.

We can mathematically express the above steps as follows. First, our optimization problem can be defined as

$$\max_{\theta} \ell(\theta), \tag{13}$$

where

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{I} \log P\left(x^{(i)}; \theta\right) \\
&= \sum_{i=1}^{I} \log \sum_{z^{(i)}} P\left(x^{(i)}, z^{(i)}; \theta\right) \\
&= \sum_{i=1}^{I} \log \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \\
&= \sum_{i=1}^{I} \log \mathbb{E}_{z^{(i)} \sim Q_i} \left(\frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \right),
\end{aligned} \tag{14}$$

where, in turn,

$$Q_i\left(z^{(i)}\right) \geq 0 \quad \text{and} \quad \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) = 1. \tag{15}$$

Recall that $\log(x)$ is a concave function, and using the Jensen's inequality for concave functions we obtain,

$$\log \mathbb{E}(x) \geq \mathbb{E}(\log(x)). \tag{16}$$

Hence,

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{I} \log \mathbb{E}_{z^{(i)} \sim Q_i} \left(\frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \right) \\ &\geq \sum_{i=1}^{I} \mathbb{E}_{z^{(i)} \sim Q_i} \left(\log \left(\frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \right) \right) \\ &= \sum_{i=1}^{I} \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log \left(\frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \right). \end{aligned} \tag{17}$$

Then, by maximizing the lower bound we can also maximize $\ell(\theta)$. Now, the only aspect that is missing is that $\ell(\theta)$ and its lower bound meets at each $\theta^{(t)}$. For this, let us recall again the Jensen's inequality introduced in Section 4.1. There, we saw that $x = \mathbb{E}(x) \Leftrightarrow f(\mathbb{E}(x)) = \mathbb{E}(f(x))$. Hence, we need to choose $Q_i\left(z^{(i)}\right)$ such that

$$\sum_{i=1}^{I} \log \mathbb{E}_{z^{(i)} \sim Q_i} \left(\frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \right) = \sum_{i=1}^{I} \mathbb{E}_{z^{(i)} \sim Q_i} \left(\log \left(\frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \right) \right). \tag{18}$$

This equality will hold when

$$\frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} = \text{constant for all values of } z^{(i)}. \tag{19}$$

Hence, $Q_i\left(z^{(i)}\right)$ should be chosen such that

$$Q_i\left(z^{(i)}\right) \propto P\left(x^{(i)}, z^{(i)}; \theta\right). \tag{20}$$

On the other hand, because $Q_i\left(z^{(i)}\right)$ is a probability distribution, it must satisfy

$$\sum_{z^{(i)}} Q_i\left(z^{(i)}\right) = 1. \tag{21}$$

Hence, choose

$$\begin{aligned} Q_i\left(z^{(i)}\right) &= \frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{\sum_{z^{(i)}} P\left(x^{(i)}, z^{(i)}; \theta\right)} \\ &= \frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{P\left(x^{(i)}; \theta\right)} \\ &= P\left(z^{(i)} | x^{(i)}; \theta\right) \end{aligned} \tag{22}$$

So, we can summarize the general version of the EM algorithm as

- **E-step**: Set

$$Q_i\left(z^{(i)}\right) = P\left(z^{(i)} | x^{(i)}; \theta\right). \tag{23}$$

- **M-step**: Set

$$\theta = \arg\max_{\theta} \sum_{i=1}^{I} \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log\left(\frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)}\right). \tag{24}$$

5 Gaussian Mixture Model (GMM)

Recall that the EM algorithm tells

$$Q_i\left(z^{(i)} = j\right) = P\left(z^{(i)} = j | x^{(i)}; \phi_j, \mu_j, \Sigma_j\right). \tag{25}$$

Then, using the Bayes' rule,

$$Q_i\left(z^{(i)} = j\right) = \frac{P\left(x^{(i)} | z^{(i)} = j\right) P\left(z^{(i)} = j\right)}{\sum_{l=1}^{k} P\left(x^{(i)} | z^{(i)} = l\right) P\left(z^{(i)} = l\right)}. \tag{26}$$

In fact,

- **E-step**:

$$w_j^{(i)} = Q_i\left(z^{(i)}\right). \tag{27}$$

- **M-step**:

$$\max_{\phi,\mu,\Sigma} \sum_{i=1}^{I} \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log\left(\frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)}\right), \tag{28}$$

  which is equivalent to

$$\max_{\phi,\mu,\Sigma} \sum_{i=1}^{I} \sum_{j=1}^{k} w_j^{(i)} \log\left(\frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_j\right) \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right)\right) \phi_j}{w_j^{(i)}}\right). \tag{29}$$

Name the lower bound of the log-likelihood function as

$$\ell\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{I} \sum_{j=1}^{k} w_j^{(i)} \log\left(\frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_j\right) \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right)\right) \phi_j}{w_j^{(i)}}\right). \tag{30}$$

Now, we maximize the above function with respect to ϕ, μ, Σ as follows:

$$\nabla_{\mu} \ell\ell(\phi, \mu, \Sigma) \overset{\text{set}}{=} 0 \Rightarrow \mu_j = \frac{\sum_{i=1}^{I} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{I} w_j^{(i)}}. \tag{31}$$

$$\nabla_{\Sigma} \ell\ell(\phi, \mu, \Sigma) \overset{\text{set}}{=} 0 \Rightarrow \Sigma_j = \frac{\sum_{i=1}^{I} w_j^{(i)} \left(x^{(i)} - \mu_j\right) \left(x^{(i)} - \mu_j\right)^T}{\sum_{i=1}^{I} w_j^{(i)}}. \tag{32}$$

Finally, in order to compute the optimal ϕ, we need to recall that it is a probability distribution (in particular, a multinomial distribution), hence we need to impose the constraint that

$$\sum_{j=1}^{k} \phi_j = 1. \tag{33}$$

Hence, in this case, we have a constrained optimization problem, and, to solve it, we need to construct the corresponding Lagrangian function as

$$\mathcal{L} = \ell\ell(\phi, \mu, \Sigma) + \beta \left(\sum_{j=1}^{k} \phi_j - 1 \right). \tag{34}$$

Then,

$$\nabla_\phi \mathcal{L} \overset{\text{set}}{=} 0 \Rightarrow \phi_j = \frac{1}{I} \sum_{i=1}^{I} w_j^{(i)}. \tag{35}$$

6 Mixture of Naive Bayes

The **mixture of Naive Bayes** allow one to model unlabeled discrete-valued input feature vectors using the EM algorithm. This algorithm can be used for instance for solving a text-clustering problem.

Suppose that a training set is given $\{x^{(1)}, \cdots, x^{(I)}\}$ where $x^{(i)} = \{0,1\}^N$. That is,

$$x_j^{(i)} = \mathbb{I}\{\text{word j appears in document i}\}. \tag{36}$$

In addition, suppose that

$$z^{(i)} \in \{0,1\} \quad (2 \text{ clusters}). \tag{37}$$

Hence,

$$z^{(i)} \sim \textbf{Bernoulli}(\phi). \tag{38}$$

Then, if we make the *Naive Bayes' assumption* (i.e., x_j^i's are conditionally independent given $z^{(i)}$), then we have

$$P\left(x^{(i)}|z^{(i)}\right) = \prod_{i=1}^{I} P\left(x_j^{(i)}|z^{(i)}\right). \tag{39}$$

Hence, let us define the following parameters

$$\begin{aligned} P\left(x_j^{(i)} = 1|z^{(i)} = 0\right) &= \phi_{j|z^{(i)}=0}, \\ P\left(x_j^{(i)} = 1|z^{(i)} = 1\right) &= \phi_{j|z^{(i)}=1}, \\ P\left(z^{(i)} = 1\right) &= \phi_z. \end{aligned} \tag{40}$$

Now, if we consider the joint distribution over $x^{(i)}$ and $z^{(i)}$, and if we work on the derivation of the EM algorithm for the MLE, we then find:

- **E-step**:

$$w^{(i)} = P(\left(z^{(i)}|x^{(i)}; \phi_{j|z}, \phi_z\right). \tag{41}$$

- **M-step**:

$$
\begin{aligned}
\phi_z &= \frac{1}{I}\sum_{i=1}^{I} w^{(i)}, \\
\phi_{j|z=1} &= \frac{\sum_{i=1}^{I} w^{(i)} \mathbb{I}\left\{x_j^{(i)} = 1\right\}}{\sum_{i=1}^{I} w_j^{(i)}}, \\
\phi_{j|z=0} &= \frac{\sum_{i=1}^{I} \left(1 - w^{(i)}\right) \mathbb{I}\left\{x_j^{(i)} = 1\right\}}{\sum_{i=1}^{I} w_j^{(i)}}.
\end{aligned}
\tag{42}
$$