# Expectation Maximization

lruczu

## 1 Gaussian Mixture Model

Gaussian mixture model is a clustering model that comes from the general family of mixture models that could be specified as follows:

$$\begin{aligned} f(x) &= \sum_{k=1}^{K} \lambda_k f_k(x), \\ \lambda_i \geq 0, &\sum_{k=1}^{K} \lambda_k = 1 \end{aligned} \tag{1}$$

The above equation can be understood in two ways. Either we can imagine that each point x partially belongs to each of the k distributions and the intensity of membership is reflected by $\lambda$'s. Or as it is probabilistic/generative model, we can imagine that first, we randomly select a mixture, where probability of selecting k-th mixture is equal to $\lambda_k$ and after having selected the mixture we generate a point from this distribution. As for the above equation, very often we have $f_1 = .. \ f_K$ and each distribution is parametrized by some vectors. In the case of Gaussian Mixture Model or GMM for short, we have:

$$f_k = \mathcal{N}(x|\mu_k, \Sigma_k) \tag{2}$$

Typically we want to find parameters of this model. The way it is done in Machine Leaning is often by mean of Maximum Likelihood Maximization process.

$$L(\theta) = \sum_{i=1}^{n} log f(x_i|\theta) \tag{3}$$

It turns out that this quantity cannot be solved in a most common way, i.e. by calculating partial derivatives and setting them to zero. The way to go is to find the lowerbounding function of the likelihood function that would be easier to maximize. To make it possible, for each data point $x_i$, random variable $Z_i$ is introduced. It is multinomial random variable that takes value k if and only if i-th point is generated by k-th mixture.

$$L(\theta) = \sum_{i=1}^{n} log \sum_{k=1}^{K} P(X_i, Z_i = k|\theta) \tag{4}$$

In the above equation $Z_i$ is just marginalized out. Along with $Z_i$, for the ease of notation it is worthwhile introducing distribution of it, namely $q_i$. So we have $q_i(\mathrm{k}) = P(Z_i = k)$. Below, we multiply each summand by 1, and make use of Jensen's inequality.

$$\begin{aligned} L(\theta) = \sum_{i=1}^{n} log \sum_{k=1}^{K} \frac{q_i(k)}{q_i(k)} P(X_i, Z_i = k|\theta) \geq \\ \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k) log \frac{P(x_i, Z_i = k|\theta)}{q_i(k)} = L(\theta, q) \end{aligned} \tag{5}$$

$L(\theta, q)$ is the family of lowerbounding functions. The inequality holds for each value of $\theta$ and for each distributions $q_i$. It can be maximized in two distinct phases.

- $q^{k+1} = \arg\max_q L(\theta^k, q)$ (E-step)
- $\theta^{k+1} = \arg\max_q L(\theta, q^{k+1})$ (M-step)

To get E-step, let's calculate the gap between the likelihood and its lowerbound.

$$\begin{aligned} L(\theta) - L(\theta, q) = \sum_{i=1}^{n} log f(x_i|\theta) - \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k) log \frac{P(x_i, Z_i = k|\theta)}{q_i(k)} = \\ \sum_{i=1}^{n} log f(x_i|\theta) \sum_{k=1}^{K} q_i(k) - \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k) log \frac{P(x_i, Z_i = k|\theta)}{q_i(k)} = \\ \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k) log f(x_i|\theta) - \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k) log \frac{P(x_i, Z_i = k|\theta)}{q_i(k)} = \\ \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k) log \frac{f(x_i|\theta) q_i(k)}{P(x_i, Z_i = k|\theta)} \end{aligned} \tag{6}$$

$P(x_i, Z_i = k|\theta)$ can be rewritten as

$$P(x_i, Z_i = k|\theta) = P(Z_i = k|x_i, \theta) P(x_i|\theta) = P(Z_i = k|x_i, \theta) f(x_i|\theta) \tag{7}$$

After substituting it, we get

$$\begin{aligned} L(\theta) - L(\theta, q) = \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k) log \frac{f(x_i|\theta) q_i(k)}{P(Z_i = k|x_i, \theta) f(x_i|\theta)} = \\ \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k) log \frac{q_i(k)}{P(Z_i = k|x_i, \theta)} = E_{q_i(Z_i)} \frac{q_i(Z_i)}{P(Z_i|x_i, \theta)} \end{aligned} \tag{8}$$

It turns out thet the last quantity if Kullback-Leibler divergence, so the gap is equal to

$$L(\theta) - L(\theta, q) = KL(q_i(Z_i)||P(Z_i|x_i, \theta)) \tag{9}$$

As Kullback-Leibler divergence is always nonnegative and is minimized where two distributions are the same, in E-step each distribution of latent variables should be set to the posterior distribution of these random variables.

$$q_i(Z_i) = P(Z_i|x_i, \theta) \tag{10}$$

The way they are calculated can be easily obtained by applying bayes theorem on the posterior distribution.
To get M-step, we have to reformulate the lowerbound derived above.

$$\begin{gathered} L(\theta, q) = \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k) \log \frac{P(x_i, Z_i = k|\theta)}{q_i(k)} = \\ \sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k)(logP(x_i, Z_i = k|\theta) - logq_i(k)) = \\ \sum_{i=1}^{n} E_{q_i(Z_i)} logP(x_i, Z_i|\theta) + const \end{gathered} \tag{11}$$

In M-step we maximize the expectation of the likelihood function (the complete model), where expectation is taken with respect to updated distributions of latent variables. The last quantity can be further rewritten as

$$\begin{gathered} L(\theta, q) = \sum_{i=1}^{n} E_{q_i(Z_i)} logP(x_i, Z_i|\theta) + const = \\ \sum_{i=1}^{n} E_{q_i(Z_i)} logP(x_i|Z_i, \theta) P(Z_i|\theta) + const \end{gathered} \tag{12}$$