Notes on EM Algorithm

Scribe: Jiajia Xie July 23, 2022

In this tutorial, I would like to introduce the well-known Expectation-Maximization (EM) algorithm for learning a mixture of models. The EM algorithm was first introduced by Arthur P. Dempster (Professor of Statistics, Harvard) in 1977 [1]. However, the convergence property of the EM algorithm is established by C.F. Wu [2]. In a modern age dominated by machine learning, the EM algorithm serves as the basic method for maximum likelihood estimation of models with latent variables, and it has generalized versions such as as Bayesian variational inference for learning complex generative models. Latent Dirichlet Allocation is one of these examples [3].

Learning a Mixture of Gaussian

The most illustrative example of implementation of the EM algorithm is, perhaps, learning a mixture of Gaussians. The problem is stated as follows: Given n i.i.d samples $x_i \in \mathbb{R}^d$, a hyper-parameter $\mathcal{K}$ denotes the number of Gaussian distributions, we model the generative process of each data x_i as:

$$p(x_i) = \sum_{k=1}^{\mathcal{K}} \pi_k \mathcal{N}(x_i|\mu_k, \Lambda_k) \tag{1}$$

where the parameters are π_k, the mixture weight of the k-th Gaussian distribution satisfying $\sum_{k=1}^{\mathcal{K}} \pi_k = 1$, and μ_k and Λ_k, the mean and cor-variance matrix of the k-th Gaussian distribution respectively.

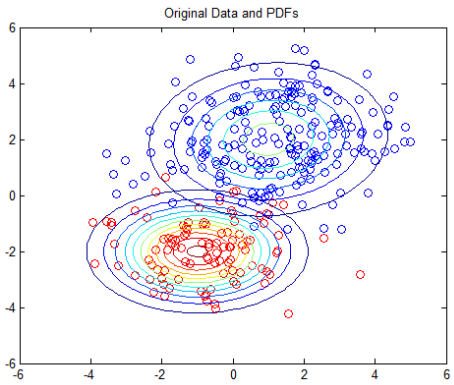

Figure 1: An example of Mixture of 2 Gaussian Distributions

A 2-d example is provided below, see Figure 1. The model reveals the sub-population structure among the data. An example for this 2-d case is: imagine that the y-axis represents height of elephants, and the x-axis represents weight. Someone has collected the data (height, weight) from the population but forgot to record the

biological gender of these samples. Suppose there is an interest to reveal the gender among data, then a good way to recover the information is by learning a 2-d mixture of Gaussian distributions. Therefore, we can inference that the blue dots are likely male because a male has larger height and weight than an female in average. Note that the missing information, gender, is a typical example of incomplete observations, which should be treated as latent variables in statistical modeling. The motivation of EM algorithm is exactly to inference the parameters of these models with latent variables.

1 Maximum Likelihood Estimation

The theory of statistical learning does not prevent us from directly training the model parameters via maximum likelihood estimation. But you should see the difficulty of optimizing the objective function. Consider taking the log of the likelihood function for an sample x_i:

$$\log p(x_i) = \log(\sum_{k=1}^{\mathcal{K}} \pi_k \mathcal{N}(x_i|\mu_k, \Lambda_k)) \tag{2}$$

Maximizing this function with respect to the parameters is computationally challenging because the function are not concave in general (since the summation happens inside a log function). This means the problem should has multiple local maxima. However, if $\mathcal{K} = 1$, then problem is equivalent to estimating the parameters of a single Gaussian distribution, and the likelihood function becomes concave now.

2 EM Algorithm

We should break down the computations of the EM algorithm for training the mixture of Gaussian distribution iteratively to see that the final computation of each step are surprisingly simple. The EM algorithm has two steps for for each iteration: the E step, and the M step.

2.1 The E-Step

First, the E step computes the log expectation of the complete likelihood with respect to the posterior $p(\pi|x_i)$. Let us define what the complete likelihood is. Assume that if you did record the gender of the data, then the observation are not just x_i, but (x_i, z_i) where z_i is a 1-of-K encoding in general. For example, in the 2-d case we have assumed $\mathcal{K} = 2$, let x_i be a data of female, then z_i is a 2-d vector with $z_{i,1} = 1$ and $z_{i,2} = 0$ indicating it belongs to the first (female) Gaussian distribution. For a general $\mathcal{K}$, $z_i = [0, 1, 0...0]$ if x_i belongs to the second Gaussian. Let us also define the

distribution of the encoding $p(z_{i,k} = 1) = \pi_k$. The complete likelihood is defined as:

$$p(x_i, z_i) = p(x_i|z_i)p(z_i) = \prod_{k=1}^{\mathcal{K}} \Big(\mathcal{N}(x_i|\mu_k, \Lambda_k)\pi_k\Big)^{z_{i,k}} \tag{3}$$

Note that since the 1-of-K encoder z_i has only one 1, the complete likelihood is just a single Gaussian distribution with its mixture weight.
The posterior, $p(z_i|x_i)$ is computed based on the Bayes's rule:

$$p(z_{i,k} = 1|x_i) = \frac{p(x_i|z_{i,k} = 1)p(z_{i,k} = 1)}{p(x_i)} = \frac{\mathcal{N}(x_i|\mu_k, \Lambda_k)\pi_k}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j, \Lambda_j)\pi_j} \tag{4}$$

Note that you have to compute the above equation for $k = 1, 2, 3...\mathcal{K}$. Therefore, the expectation of the log-complete likelihood of a single x_i is:

$$\begin{aligned}
\mathbb{E}_{p(z_i|x_i)}\Big(\log p(x_i, z_i)\Big) &= \mathbb{E}_{p(z_i|x_i)}\Big(\sum_{k=1}^{\mathcal{K}} z_{i,k}\Big(\log \mathcal{N}(x_i|\mu_k, \Lambda_k) + \log p(z_{i,k} = 1)\Big)\Big) \\
&= \sum_{j=1}^{\mathcal{K}}\sum_{k=1}^{\mathcal{K}} p(z_{i,j} = 1|x_i) z_{i,k}\Big(\log \mathcal{N}(x_i|\mu_k, \Lambda_k) + \log \pi_k\Big) \\
&= \sum_{k=1}^{\mathcal{K}} p(z_{i,k} = 1|x_i)\Big(\log \mathcal{N}(x_i|\mu_k, \Lambda_k) + \log \pi_k\Big)
\end{aligned} \tag{5}$$

Note that since the EM algorithm is an iterative algorithm, at the E step of t, the parameters attained from the previous iteration, denoted by π^{t-1}, μ^{t-1}, and Λ^{t-1}, should be treated as constants for computing the posterior. For convenience, let $\mathcal{Q}^t$ be the log-complete likelihood of of all given x_i at time t, that is:

$$\mathcal{Q}^t(x, \pi, \mu, \Lambda) = \sum_{i=1}^{n}\sum_{k=1}^{\mathcal{K}} \frac{\mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}}\big(\log \mathcal{N}(x_i|\mu_k, \Lambda_k) + \log \pi_k\big) \tag{6}$$

2.2 The M-Step

The M step of the EM algorithm maximize the log-complete likelihood attained from the E step with respect to all parameters. That is, we need to solve:

$$\begin{aligned}
\max_{\pi_k, \mu, \Lambda} \quad & \mathcal{Q}^t(x, \pi, \mu, \Lambda) \\
\text{s.t.} \quad & \textstyle\sum_{k=1}^{\mathcal{K}} \pi_k = 1
\end{aligned} \tag{7}$$

2.2.1 Solve for Parameters π

You should be able to verify that this objective function is concave in all parameters. To solve for π, let us form the Lagrange multiplier, which is:

$$\mathcal{L}(\pi, \lambda) = \mathcal{Q}^t(x, \pi) + \lambda(1 - \sum_{k=1}^{\mathcal{K}} \pi_k) \tag{8}$$

According to the K.K.T conditions, the stationary condition is $\frac{\partial \mathcal{L}}{\partial \pi_k} = 0$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \pi_k} &= -\lambda + \frac{\partial}{\partial \pi_k}\Big(\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{K}} \frac{\mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}} \big(\log \mathcal{N}(x_i|\mu_k, \Lambda_k) + \log \pi_k \big) \Big) \\
&= -\lambda + \frac{1}{\pi_k}\Big(\sum_{i=1}^{n} \frac{\mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}} \Big) \\
&= 0
\end{aligned}$$

Solve the above equation for π_k and λ, it is not hard to see the following relation:

$$\lambda \pi_k = \sum_{i=1}^{n} \frac{\mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}} \tag{9}$$

Remember that we have other conditions from K.K.T. The primal feasibility informs that $\sum_{k=1}^{\mathcal{K}} \pi_k = 1$. If we take the summation over k, we get:

$$\sum_{k=1}^{\mathcal{K}} \lambda \pi_k = \sum_{k=1}^{\mathcal{K}}\sum_{i=1}^{n} \frac{\mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}} \tag{10}$$

The left hand side is just λ, and the left hand side is n.

$$\lambda = \sum_{i=1}^{n} \frac{\sum_{k=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}} = n \tag{11}$$

Therefore, with $\lambda = n$, and plug it into Equation (9) we attain the close form solution for π_k;

$$\pi_k^t = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}} \tag{12}$$

which is just the average of all posteriors.

2.2.2 Solve for Parameters μ and Λ

Maximize $\mathcal{Q}^t(x, \pi, \mu, \Lambda)$ with respect to μ and Λ is really simple. To see this, remember that we shall treat π as constants while maximizing the objective function with respect to any μ_k and Λ_K. Therefore, the objective function is just equivalent to:

$$\mathcal{Q}^t(x, \mu, \Lambda) = \sum_{i=1}^{n}\sum_{k=1}^{\mathcal{K}} \frac{\mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}} \big(\log \mathcal{N}(x_i|\mu_k, \Lambda_k) + C \big) \tag{13}$$

for some constants C. Note that this is a unconstrained optimization problem. If you take the gradient, for example, with respect to μ_k, it is not hard to see:

$$\nabla_{\mu_k} \mathcal{Q}^t(x, \mu, \Lambda) = \sum_{i=1}^{n} \frac{\mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}} \nabla_{\mu_k} \Big(\log \mathcal{N}(x_i|\mu_k, \Lambda_k) \Big) \tag{14}$$

Now if we let $\mathcal{W}_{i,k} = \frac{\mathcal{N}(x_i|\mu_k^{t-1}, \Lambda_k^{t-1})\pi_k^{t-1}}{\sum_{j=1}^{\mathcal{K}} \mathcal{N}(x_i|\mu_j^{t-1}, \Lambda_j^{t-1})\pi_j^{t-1}}$ Isn't it equivalent to maximize the likelihood of the n samples of a single Gaussian distribution $\mathcal{N}(x_i|\mu_k, \Lambda_k)$ with each sample x_i assigned a weight $\mathcal{W}_{i,k}$? You should already knew that (please see PRML 2.1-2.3) the optimal maximum likelihood estimation for a n-dimensional Gaussian distribution is just the sample mean and the sample covariance matrix. Therefore, if you set the above equation to 0 and solve for μ_k, you should get solution weighted by $\mathcal{W}$:

$$\mu_k^t = \frac{\sum_{i=1}^{n} \mathcal{W}_{i,k} x_i}{\sum_{i=1}^{n} \mathcal{W}_{i,k}} \tag{15}$$

Similarly, for Λ_k, you will get:

$$\Lambda_k^t = \frac{\sum_{i=1}^{n} \mathcal{W}_{i,k} (x_i - u_k^t)(x_i - u_k^t)^\top}{\sum_{i=1}^{n} \mathcal{W}_{i,k}} \tag{16}$$

3 Theory behind EM Algorithm

Why does the EM algorithm converge to a local maxima in general for maximum likelihood estimation of models with latent variables? To see the magic behind the EM algorithm, let us analyze the algorithm in a theoretical framework. First, we introduce a measure to quantify the similarity of one probability distribution against the another called Kullback–Leibler (KL) divergence defined below:

$$KL(q(x)||p(x)) = -\int q(x) \log \left(\frac{p(x)}{q(x)} \right) dx \tag{17}$$

Note that we should use integral for generality, and it can be changed to summation if x is defined on a discrete set. The KL divergence is non-negative and is 0 if and only if $q(x) = p(x)$.

Consider a probabilistic model in which we define all observed variables by X and all latent variables (not observed) by Z. The EM algorithm assumes that $p(X|\theta)$, which is the likelihood parameterized by θ, is difficult to compute, but the complete likelihood $p(X, Z|\theta)$ is significantly easier to compute. Our goal is to maximize $p(X|\theta) = \int p(X, Z|\theta) dZ$.

Next, let us introduce an arbitrary distribution of Z, called $q(Z)$. The magic begins if you look at the KL divergence of $q(Z)$ and the posterior $P(Z|X, \theta)$:

$$KL(q(Z)||P(Z|X,\theta)) = -\int q(Z)\log\left(\frac{P(Z|X,\theta)}{q(Z)}\right)dZ \tag{18}$$

According to the Bayes' rule, $P(Z|X,\theta) = \frac{p(X,Z|\theta)}{p(X|\theta)}$, and we put that into the KL divergence:

$$\begin{aligned}
KL(q(Z)||P(Z|X,\theta)) &= -\int q(Z)\log\left(\frac{\frac{p(X,Z|\theta)}{p(X|\theta)}}{q(Z)}\right)dZ \\
&= -\int q(Z)\left(\log\left(\frac{p(X,Z|\theta)}{q(Z)}\right) - \log\left(p(X|\theta)\right)\right)dZ \\
&= -\int q(Z)\log\left(\frac{p(X,Z|\theta)}{q(Z)}\right)dZ + \int q(Z)\log\left(p(X|\theta)\right)dZ \\
&= -\int q(Z)\log\left(\frac{p(X,Z|\theta)}{q(Z)}\right)dZ + \log\left(p(X|\theta)\right)\int q(Z)dZ \\
&= -\int q(Z)\log\left(\frac{p(X,Z|\theta)}{q(Z)}\right)dZ + \log\left(p(X|\theta)\right)
\end{aligned}$$

Therefore, rearrange the terms we have the following relation:

$$\log\left(p(X|\theta)\right) = \int q(Z)\log\left(\frac{p(X,Z|\theta)}{q(Z)}\right)dZ - KL(q(Z)||P(Z|X,\theta)) \tag{19}$$

where the left hand side is just the log-likelihood, and the first term on the right hand side is also familiar: the expected complete log-likelihood with respect to $q(Z)$, denoted by $\mathcal{G}(q,\theta)$. Note that since $KL(q(Z)||P(Z|X,\theta)) \geq 0$, $\mathcal{G}(q,\theta)$ can be treated as a lower bound for $\log\left(p(X|\theta)\right)$, that is:

$$\mathcal{G}(q,\theta) \leq \log\left(p(X|\theta)\right) \tag{20}$$

where the inequality becomes equality only when $q(Z) = P(Z|X,\theta)$. Therefore, the EM algorithm in general can be treated as coordinate ascent algorithm. To see this, let:

$$\mathcal{F}(q,\theta^{t-1}) = \mathcal{G}(q,\theta) - KL(q(Z)||P(Z|X,\theta)) \tag{21}$$

which is just the right hand side of Equation (19) at the beginning of the t step of the EM algorithm. The E step of the algorithm is:

E-Step

$$q^t = \arg\max_q \mathcal{F}(q,\theta^{t-1}) \tag{22}$$

where $q^t = P(Z|X,\theta)$ since the lower bound become inequality when $KL(q(Z)||P(Z|X,\theta)) = 0$. The term $\mathcal{G}(q^t,\theta)$ is the expected log-complete likelihood with respect to the posterior $P(Z|X,\theta)$, which is exactly what we did. The M step can be viewed as:

M-Step

$$\theta^t = \arg\max_{\theta} \mathcal{F}(q^t, \theta) \tag{23}$$

which is just the M step that maximizes the expected log-complete likelihood in parameters θ. Overall, the EM algorithm first create the tightest lower bound for the log-likelihood, then it maximizes this lower bound by varying the parameters.

4 Practical Solution of Computing High Dimensional Gaussian

Recall that the general d-dimensional Gaussian distribution has the probability density function defined as:

$$p(X|\mu, \Lambda) = \frac{1}{\sqrt{(2\pi)^d |\Lambda|}} \exp(-\frac{1}{2}(X-\mu)\Lambda^{-1}(X-\mu)^\top)$$

which involves the computation of the inverse covariance matrix Λ^{-1} and the determinant $|\Lambda|$. In practice, the covariance can be singular which mean the inverse does not exist. We can always make more unrealistic assumption on the distribution to simplify the computation. One example is to assume that the Gaussian distribution is spherical, which means $\Lambda = s^2 I$ where $s \in \mathbb{R}$. The probability density function is given by:

$$p(X|\mu, s^2 I) = \frac{1}{s^d \sqrt{(2\pi)^d}} \exp(-\frac{(X-\mu)^\top (X-\mu)}{2s^2})$$

4.1 Exercise

Suppose you are given n i.i.d samples $x_i \in \mathbb{R}^d$, formulate the maximum likelihood estimation of a spherical Gaussian distribution and solve for the optimal μ and s^2.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[2] C. J. Wu, "On the convergence properties of the em algorithm," *The Annals of statistics*, pp. 95–103, 1983.

[3] M. Hoffman, F. Bach, and D. Blei, "Online learning for latent dirichlet allocation," *advances in neural information processing systems*, vol. 23, 2010.