EM-algorithm and Clustering: a Tutorial

Jacob Su Wang SHREKWANG@UTEXAS.EDU
Department of Statistics and Data Science, the University of Texas Austin

Abstract

The EM-algorithm (Dempster et al., 1997) applies widely in unsupervised learning, in particular clustering models, e.g. K-means (Kanungo et al., 2002) and Bernoulli Mixture models (Juan & Vidal, 2004). Many, however, have treated the algorithm as a pure blackbox in application. In this guide, I derive the EM in detail and demonstrate its uses in clustering models, for the convenience of a deeper understanding of the algorithm for machine learning practitioners. The readers are assumed to be familiar with basic clustering techniques and general knowledge in probability and statistics.

1. K-Means: an Intro Example

The simple formulation of the parameter estimation of the K-means is familiar to many machine learning researchers, while few know that it falls under the family of EM-algorithms as a special variant. In this section I outline the K-means in this general framework.

The general idea of the K-means goes as follows: We start by randomly initializing a set of cluster centroids, then assign the observations to their nearest centroids. On completing the cluster assignment, we re-estimate the centroid based on it. We iterate through the procedure, until the centroids stop shifting in re-estimation.

Now we analyze the algorithm formally. Let $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_K\}$ be the set of centroid variables, where $\mu_k \in \mathbb{R}^d$, and let $\boldsymbol{x} = \{x_1, \dots, x_N\}$ be the set of observations. The parameters we are interested in estimating are $\boldsymbol{\mu}$. Now let α_{nk} be an cluster-assignment indicator such that

$$\alpha_{nk} = \begin{cases} 1, & \text{if } x_n \text{ is assigned to cluster } k \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In an ideal clustering, we expect the within-cluster distance between observations and centroids to be as small as possible, i.e. cohesive clusters. Therefore, taking the distance to be Euclidean, we formulate an objective function as follows:

$$\mathcal{L} = \sum_{k=1}^{K} \sum_{n=1}^{N} \alpha_{nk} \parallel x_n - \mu_k \parallel^2 \tag{2}$$

The optimization objective, thus, is

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu}}{\operatorname{argmin}}\ \mathcal{L} \tag{3}$$

From calculus, we know that $\boldsymbol{\mu}$ can be found by setting $\nabla_{\boldsymbol{\mu}} \mathcal{L} \triangleq 0$, then solve for $\boldsymbol{\mu}$. However, due to the incompleteness of the data, i.e. the cluster assignments of $\boldsymbol{x}$ are unknown, we cannot perform the estimation directly. This leads us to the following iterative routine:

1. Initialize $\boldsymbol{\mu}$
2. Assign $\{x_n\}_{n=1}^N$ by
$$\alpha_{nk} = \begin{cases} 1, & \text{if } k = \underset{k}{\operatorname{argmin}} \parallel x_n - \mu_k \parallel^2 \\ 0, & \text{otherwise} \end{cases}$$
3. Set $\nabla_{\boldsymbol{\mu}} \mathcal{L} \triangleq 0$, then solve for $\boldsymbol{\mu}$

The procedure above is a special case of the EM-algorithm: step 2 (the E-step) assigns the observations to their respective most-likely expected clusters, and step 3 (the M-step) minimizes $\mathcal{L}$ with respect to the centroid parameters. The example also illustrates that EM does not necessarily translates into "Expectation *Maximization*", as here we have a case of the opposite. One may understand the M-step in general as the adjustment of the objective parameters to meet the optimization goal.

2. Deriving EM-Algorithm

We start by stating the general formulation of the EM-algorithm, then proceed to a complete derivation[1], and

[1] The work extends from Jeff Miller's Machine Learning course (Miller, 2011).

finally close with a simple proof of the crucial theorem which states that the EM is guaranteed to make improvement at every step.

Let $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ be the observation set, and X, Z be the data variables and latent variables[2] respectively. Suppose X, Z are distributed by some density p parameterized by $\theta \in \Theta$. Our optimization objective, e.g. MLE, is then

$$\theta_{MLE} = \underset{\theta}{\text{argmax}}\ p_\theta(x) = \underset{\theta}{\text{argmax}} \sum_z p_\theta(x, z) \quad (4)$$

As oftentimes p_θ can be complex (e.g. multimodal), the optimization cannot be solved analytically. We thus resort to the EM-algorithm, which is formulated as follows[3]

- Initialization: θ_0
- E-step: For $t = 0, 1, \ldots$

$$Q(\theta, \theta_t) = \mathbb{E}_{\theta_t}\left[\log p_\theta(X, Z) | X = x\right] \quad (5)$$

- M-step:

$$\theta_{t+1} = \underset{\theta}{\text{argmax}}\ Q(\theta, \theta_t) \quad (6)$$

To understand Eq. 5, we look at a natural sequence of derivations that leads to it. In particular, we consider the case where the density $p_\theta(x, z)$ is in the *exponential family*[4], i.e. it can be written in the form

$$p_\theta(x, z) = \frac{1}{c(\theta)} \cdot \exp\left\{g(\theta)^\intercal s(X, Z)\right\} \cdot h(X, Z) \quad (7)$$

where $g(\theta)$ are the *natural parameters*[5], $s(X, Z)$ are the *sufficient statistics*, $h(X, Z)$ the scaling constant, and $c(\theta)$ the normalizing constant. To estimate the MLE parameters for $p_\theta(x, z)$, first note that we know that $\underset{\theta}{\text{argmax}}\ p_\theta(x) = \underset{\theta}{\text{argmax}} \log p_\theta(x)$, where $p_\theta(x) = \sum_z p_\theta(x, z)$. Another crucial property[6] of the exponential family is $\nabla_\theta \log c(\theta) = \mathbb{E}_\theta\left[s(X, Z)\right]$, which we exploits extensively later. Setting our optimization objective as

$$\theta^* = \underset{\theta}{\text{argmax}} \log p_\theta(x)$$

and Eq. 7, we calculate the gradient $\nabla_{\theta_i} \log p_\theta(x)$ as follows

$$\begin{aligned}
&\nabla_\theta \log p_\theta(x) \\
&= \frac{1}{p_\theta(x)} \nabla_\theta p_\theta(x) \\
&= \frac{1}{p_\theta(x)} \sum_z \nabla_\theta p_\theta(x, z) \\
&= \frac{1}{p_\theta(x)} \sum_z \left(s(X, Z) - \nabla_\theta \log c(\theta)\right) \\
&\quad \times e^{\theta^\intercal s(X,Z) - \log\ c(\theta)} \cdot h(X, Z) \\
&= \frac{1}{p_\theta(x)} \sum_z \left(s(X, Z) - \mathbb{E}_\theta\left[s(X, Z)\right]\right) \cdot p_\theta(x, z) \\
&= \left(\sum_z \left(s(X, Z) \cdot p_\theta(z|x)\right) - \mathbb{E}_\theta\left[s(X, Z)\right] \sum_z p_\theta(z|x)\right) \\
&\quad \times \sum_z p_\theta(z|x) \\
&= \mathbb{E}_\theta\left[s(X, Z) | X = x\right] - \mathbb{E}_\theta\left[s(X, Z)\right]
\end{aligned}$$

Now, set $\nabla_\theta \log p_\theta(x) \triangleq 0$, we have

$$\mathbb{E}_{\theta_t}\left[s(X, Z) | X = x\right] = \mathbb{E}_{\theta_{t+1}}\left[s(X, Z)\right] \quad (8)$$

where we initialize the parameter on the LHS to be some θ_0, solve for θ_1 on the RHS, then plug the result in the LHS again, and iterate until convergence of θ.

To see the connection between Eq. 8 and Eq. 5

$$\begin{aligned}
&p_\theta(x, z) = e^{\theta^\intercal s(X,Z) - \log\ c(\theta)} h(X, Z) \\
&\Rightarrow \log p_\theta(x, z) = \theta^\intercal s(X, Z) - \log c(\theta) + \log h(X, Z) \\
&\Rightarrow \mathbb{E}_{\theta_0}\left[\log p_\theta(X, Z) | X = x\right] = \\
&\qquad \theta^\intercal \mathbb{E}_{\theta_0}\left[s(X, Z) | X = x\right] - \log c(\theta) + \log h(X, Z)
\end{aligned}$$

from which we have the gradient

$$\begin{aligned}
&\nabla_\theta \mathbb{E}_{\theta_0}\left[\log p_\theta(X, Z) | X = x\right] = \\
&\qquad \mathbb{E}_{\theta_0}\left[s(X, Z) | X = x\right] - \mathbb{E}_\theta\left[s(X, Z)\right]
\end{aligned}$$

Setting $\nabla_\theta \mathbb{E}_{\theta_0}\left[\log p_\theta(X, Z) | X = x\right] = 0$

$$\mathbb{E}_{\theta_0}\left[s(X, Z) | X = x\right] = \mathbb{E}_\theta\left[s(X, Z)\right] \quad (9)$$

Eq. 9 is equivalent to Eq. 8, justifying the formulation in Eq. 5 and 6.

Every EM round effects nondescreasing performance gain. The property is stated formally in the following theorem

Theorem 1. *The sequence of EM estimates $\theta_0, \theta_1, \ldots$ satisfies the following property for any observation x*

$$p_{\theta_t}(x) \leq p_{\theta_{t+1}}(x), \quad \forall t = 1, 2, \ldots$$

[2] In K-means, for example, the latent variables are $\{\alpha_{nk}\}$.

[3] We will see soon, in the proof of Theorem 1, that $\underset{\theta}{\text{argmax}}\ p_\theta(x)$ is equivalent to $\underset{\theta}{\text{argmax}}\ Q(\theta, \theta_t)$.

[4] In fact, the EM specializes to the exponential family. For more details on the exponential family, cf. Murphy (2012)

[5] When $g(\theta) = \theta$, we call it the *canonical form*.

[6] Proved in Murphy (2012), section 9.2.

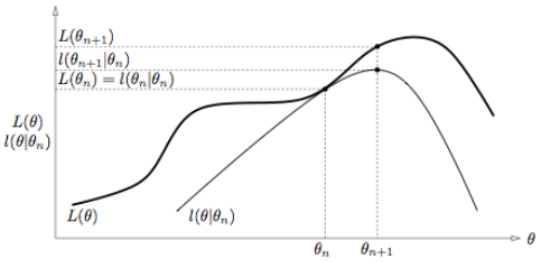

Figure 1. EM nondecreasing

Proof. Let $L(\theta) = \log\, p_\theta(x)$, we know that if $L(\theta)$ is nondecreasing, Theorem 1 follows. We now derive $L(\theta)$ to link it to the EM formula Eq. 5.

$$\begin{aligned}
L(\theta) &= \log\, p_\theta(x) \\
&= \sum_z q(z) \cdot \log \left(p_\theta(x) \cdot p_\theta(z|x) \cdot \frac{q(z)}{p_\theta(z|x)} \cdot \frac{1}{q(z)} \right) \\
&= \sum_z q(z) \log\, p_\theta(x, z) \\
&\quad + \sum_z q(z) \log\, \frac{q(z)}{p_\theta(z|x)} \\
&\quad + \sum_z q(z) \log\, q(z) \\
&= \mathbb{E}_q \left[\log\, p_\theta(X, Z) \right] + D(q \parallel p_\theta) - H(q)
\end{aligned}$$

where D is the *KL divergence*, which is nonnegative, and H is the entropy. Now if we set $q(z) \triangleq p_{\theta_n}(z|x)$ for some $\theta_n \neq \theta$, we may define

$$\begin{aligned}
l(\theta|\theta_n) &= \mathbb{E}_{\theta_n} \left[\log\, p_\theta(X, Z) | X = x \right] + D(p_{\theta_n} \parallel p_\theta) - H(\theta_t) \\
&= Q(\theta, \theta_n) - H(\theta_n) + D(p_{\theta_n} \parallel p_\theta) \\
&\geq Q(\theta, \theta_n) - H(\theta_n)
\end{aligned}$$

Further, if we set $q(z) \triangleq p_\theta(z|x)$ instead, $D(p_\theta \parallel p_\theta) = 0$, and we have

$$l(\theta|\theta) = L(\theta) = Q(\theta, \theta) - H(\theta)$$

Now we are ready to show that $L(\theta)$ is nondecreasing. In the M-step of the EM loop, we maximize Q by $\theta_{n+1} = \underset{\theta}{\text{argmax}}\, Q(\theta, \theta_n)$. As we are maximizing for θ, $H(\theta_n)$ is but a constant. It is therefore clear that $Q(\theta, \theta_{n+1}) \geq Q(\theta, \theta_n) \Rightarrow L(\theta_{n+1}) \geq L(\theta_n)$, i.e. $L(\theta)$ is nondecreasing (Figure 1). □

3. Bernoulli Mixture: a Demo

Bernoulli mixture model has two levels of variables: the mixture components which correspond to clusters, and a separate independent Bernoulli distribution under each cluster. For a good intuition, we draw analogy from the nationality of an individual and the binary status of him whether his height is higher than 6 feet. Take each country z as a mixture component, with any instance individual having a probability over countries $p(z)$ defined for him (i.e. his nationality). Representing an individual i's height status with x_i, conceivably it depends on his home country, i.e. $p(x_i|z)$, a Bernoulli distribution. In addition, roughly it is reasonable to assume that one person's height status in general does not depend on another's.

To describe the model more formally, let Z be the random variable for mixture components which follows a *categorical distribution*, i.e. $Z \sim \text{Cat}(\alpha)$, and X the individual random variable such that $X|Z \sim \text{Bernoulli}(\rho)$. Further let $Z \in \mathbb{R}^K$, the density functions of the variables are then

$$p(z) = \alpha^{\mathbb{1}(z=1)} \cdot \ldots \cdot \alpha^{\mathbb{1}(z=K)} \tag{10}$$

$$p(x) = \rho^x \cdot (1 - \rho)^{1-x} \tag{11}$$

where $0 \leq \alpha \leq 1, \sum_K \alpha = 1$ and $0 \leq \rho \leq 1$. Denoting the parameters of the entire model collectively with $\theta = (\alpha, \theta)$, we have the following densities which we use extensively later

$$p_\theta(x) = \sum_z p(z) p(x|z) = \sum_{k=1}^K \alpha_k \cdot \rho^x (1 - \rho)^{1-x} \tag{12}$$

$$p_\theta(x, z) = \prod_{k=1}^K \alpha_k^{z_k} \cdot \rho_k^{z_k x} (1 - \rho_k)^{z_k (1-x)} \tag{13}$$

$$\begin{aligned}
p_\theta(z|x) &= \frac{p(z) p(x|z)}{p(x)} \\
&= \frac{\alpha_k \cdot \rho_k^x (1 - \rho_k)^{1-x}}{\sum_{l=1}^K \alpha_l \cdot \rho_l^x (1 - \rho_l)^{1-x}} \\
&= \mathbb{E}_\theta \left[Z | X \right]
\end{aligned} \tag{14}$$

Now suppose we have the observations $\boldsymbol{x} = \{x_1, \ldots, x_N\}$, each is paired with a component/cluster assignment variable z, i.e. $\boldsymbol{z} = \{z_1, \ldots, z_N\}$, the objective probability we want to maximize (with respect to θ) will be

$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = \prod_{n=1}^N \prod_{k=1}^K \alpha_k^{z_{nk}} \cdot \rho_k^{z_{nk} x_n} (1 - \rho_k)^{z_{nk}(1-x_n)} \tag{15}$$

Taking exp log, we write the formula in the exponential

form

$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = \exp\left\{\sum_{n=1}^{N}\sum_{k=1}^{K}[z_{nk}\log\,\alpha_k]\right\} \times \exp\left\{\sum_{n=1}^{N}\sum_{k=1}^{K}[x_n z_{nk}\log\,\rho_k]\right\} \times \exp\left\{\sum_{n=1}^{N}\sum_{k=1}^{K}[z_{nk}(1-x_n)\log\,(1-\rho_k)]\right\} \tag{16}$$

Reorganizing the RHS, we end up with two terms

$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = \tag{17}$$

$$\exp\left\{\sum_{k=1}^{K}\left[\left(\sum_{n=1}^{N} z_{nk}\right)(\log\,\alpha_k + \log(1-\rho_k))\right]\right\} \times \exp\left\{\sum_{k=1}^{K}\left[\left(\sum_{n=1}^{N} x_n z_{nk}\right)(\log\,\rho_k - \log(1-\rho_k))\right]\right\} \tag{18}$$

By Eq. 7 and 17, we know the sufficient statistics are

$$\left\{\sum_{n=1}^{N} z_{nk}, \quad \sum_{n=1}^{N} x_n z_{nk}\right\}$$

We treat the two terms in Eq. 17 separately by applying Eq. 9. For the first term, we set

$$\mathbb{E}_{\theta_0}\left[\sum_n Z_{nk}|X=x\right] = \mathbb{E}_\theta\left[\sum_n Z_{nk}\right] \tag{19}$$

where the LHS equals to $\sum_i \mathbb{E}_{\theta_0}[Z_{nk}|X_{x_n}]$, where the expectation is of the conditional density Eq. 14. For notational simplicity, we abbreviate it with r_{nk}. Therefore, we have the LHS as $\sum_n r_{nk}$. For the RHS, by the property of the categorical distribution, we know that $\mathbb{E}_\theta[Z_{nk}] = \alpha_k$, this gives us $\sum_n \alpha_k = N\alpha_k$, as α_k does not depend on n. Finally Eq. 18 becomes $\sum_n r_{nk} = n\alpha_k$. Solving for α_k, we have

$$\alpha_k = \frac{1}{N}\sum_n r_{nk} \tag{20}$$

Now consider the second term in Eq. 17.

$$\mathbb{E}_{\theta_0}\left[\sum_n X_n Z_{nk}|X=x\right] = \mathbb{E}_\theta\left[\sum_n X_n Z_{nk}\right] \tag{21}$$

Exploiting Eq. 14 again, we have the LHS equal to $\sum_n \mathbb{E}_{\theta_0}[Z_{nk}|X=x_n]\,x_n = \sum_n r_{nk}x_n$. For the RHS, we leverage the property of *iterated expectation*, and derive it as follows

$$\begin{aligned}\mathbb{E}_\theta\left[\sum_n X_n Z_{nk}\right] &= \sum_n \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[X_n Z_{nk}|Z_{nk}=1\right]\right] \\ &= \sum_n \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[X_n|Z_{nk}=1\right]\right] \\ &= \sum_n \mathbb{E}_\theta\left[\rho_k\right] \\ &= \sum_n \alpha_k\rho_k = N\alpha_k\rho_k\end{aligned}$$

The derivation draws on the property of the Bernoulli distribution, i.e. $X|Z \sim \text{Bernoulli}(\rho) \Rightarrow \mathbb{E}[X|Z] = \rho$, and again the fact $Z \sim \text{Cat}(\alpha) \Rightarrow \mathbb{E}[Z] = \alpha$. Eq. 20 now becomes $\sum_n r_{nk}x_n = N\alpha_k\rho_k$. Solving for ρ_k, we have

$$\rho_k = \frac{1}{N\alpha_k} r_{nk}x_n \tag{22}$$

For completeness, we outline the EM routine for the Bernoulli mixture model as follows

- INITIALIZATION: $\theta_0 = (\alpha_0, \rho_0)$
- E-STEP: for $t = 0, 1, \ldots$, set

$$\mathbb{E}_{\theta_0}\left[\sum_n Z_{nk}|X=x\right] = \mathbb{E}_\theta\left[\sum_n Z_{nk}\right]$$
$$\mathbb{E}_{\theta_0}\left[\sum_n X_n Z_{nk}|X=x\right] = \mathbb{E}_\theta\left[\sum_n X_n Z_{nk}\right]$$

- M-STEP: solve for $\theta = (\alpha, \rho)$

$$\alpha_k = \frac{1}{N}\sum_n r_{nk}$$
$$\rho_k = \frac{1}{N\alpha_k} r_{nk}x_n$$

We iterate until the change $\Delta(\theta_{t-1}, \theta_t)$ drops below some small number ϵ.

4. Conclusion

We have introduced, in the context of clustering models, the EM-algorithm in complete detail, linking its *prima facie* arbitrary formulation with a natural derivation: EM as an MLE technique. We have also seen the application of EM in the K-means clustering and the Bernoulli mixture model. We note, for the EM to be applicable, the objective density must be in the exponential family.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via EM algorithm. *J. Royal Stat. Soc.*, 39:1–38, 1997.

Juan, Alfons and Vidal, Enrique. Bernoulli mixture-models for binary images. *IEEE*, 2004.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. An efficient K-means clustering algorithm: analysis and implementation. *IEEE*, 24:881–892, 2002.

Miller, Jeff. CS1950: Introduction to machine learning. *Brown University*, 2011.

Murphy, Kevin. *Machine Learning: a probabilistic perspective*. MIT Press, 2012.