An Introduction to Expectation-Maximization

Dahua Lin

Abstract

This notes reviews the basics about the Expectation-Maximization (EM) algorithm, a popular approach to perform model estimation of the generative model with latent variables. We first describe the E-steps and M-steps, and then use finite mixture model as an example to illustrate this procedure in practice. Finally, we discuss its intrinsic relations with an optimization problem, which reveals the nature of E-M.

1 The Expectation and Maximization

Consider a generative model with parameter θ. The model generates a set of data D, which comprises two parts: (1) *the observed data* X, and (2) *the latent data* Z, which is observed. With this model, the complete likelihood of D is given by

$$p(D|\theta) = p(X, Z|\theta), \tag{1}$$

and the marginal likelihood of the observed data X is given by

$$p(X|\theta) = \sum_z p(X, z|\theta). \tag{2}$$

Here, we sum over all possible assignments z to the variable Z. If Z is continuous, then we can change the sum to the integral over the domain of Z. The maximum likelihood estimation (MLE) of θ given X is to find the parameter $\theta \in \Theta$ that maximizes the marginal likelihood, as

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, p(X|\theta) = \underset{\theta\in\Theta}{\operatorname{argmax}} \log p(X|\theta). \tag{3}$$

Here, Θ is the parameter domain, *i.e.* the set of all valid parameters. In practice, it is usually easier to work with the log-likelihood instead of the likelihood itself. For many problems, including all the examples that we shall see later, the size of the domain of Z grows exponentially as the problem scale increases, making it computationally intractable to exactly evaluate (or even optimize) the marginal likelihood as above. The expectation maximization (E-M) algorithm was developed to address this issue, which provides an iterative approach to perform MLE.

The E-M algorithm, as described below, alternates between *E-steps* and *M-steps* until convergence.

1. Initialize $\hat{\theta}^{(0)}$. One can simply set $\theta^{(0)}$ to some random value in Θ, or employ some problem-specific heuristics to derive an initial guess.

2. For $t = 1, 2, \ldots$, repeat:

   a. **E-step.** Compute the posterior distribution of Z given X and $\theta^{(t-1)}$, as

   $$q^{(t)}(Z) = p(Z|X; \hat{\theta}^{(t-1)}). \tag{4}$$

   Here, we use $q^{(t)}$ to denote the posterior distribution of Z obtained at the t-th iteration.

   b. **M-step.** Solve the optimal $\hat{\theta}^{(t)}$ by maximizing the expectation of the complete log-likelihood with respect to $q^{(t)}$, as

   $$\hat{\theta}^{(t)} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, E_{q^{(t)}}\left(\log p(X, Z|\theta)\right) = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_z q^{(t)}(z) \log p(X, z|\theta). \tag{5}$$

   The computation of this sum can often be greatly simplified by taking advantage of the independence between variables, as we shall see in next section.

3. The iterations stop when some convergence criterion is met. For example, it can be stopped when the difference between $\theta^{(t)}$ and $\theta^{(t-1)}$ is below some threshold.

2 E-M in Practice: Estimation of Finite Mixture Models

One of the most important application of E-M is the maximum likelihood estimation of finite mixture models. Generally, a *finite mixture model* is composed of K *component models*, whose parameters are respective $\theta_1, \ldots, \theta_K$, and a prior distribution over these components, denoted by π. The probability density function (pdf) of this finite mixture model is given by

$$p(x|M) = \sum_{k=1}^{K} \pi_k p_c(x|\theta_k). \tag{6}$$

Here, we use $M = (\pi, \theta_1, \ldots, \theta_K)$ to denote all parameters of the mixture model. p_c is the pdf of a component model, which has a parameter θ_k. When each component model is a Gaussian model, then this finite mixture model is called a *Gaussian mixture model.*

Generation of a sample x from a finite mixture model as described above can be done in two steps: (1) draw an indicator variable z from the prior distribution π as $z \sim \pi = (\pi_1, \ldots, \pi_K)$. Here, z can take an integer value in $\{1, \ldots, K\}$. This value indicates which particular component is going to be used to generate x. (2) Given z, draw x from the z-th component model. Thus, we can consider this as a process that generates two values: z and x.

Given a set of samples $X = \{x^1, \ldots, x^n\}$, we can treat the associated set of indicators $Z = \{z^1, \ldots, z^n\}$ as latent variables. To estimate the model parameters, we can use E-M algorithm. To derive the E-steps and the M-steps, we go through three steps as follows.

1. **Write down the complete log-likelihood.** The complete log-likelihood is the log of the joint pdf of both the observed and latent variables, given the model parameters. In a finite mixture model, it is given by

$$\log p(X, Z|M) = \log \prod_{i=1}^{n} p(x^i|z^i; M) p(z^i|M) \tag{7}$$

Here, $p(x^i|z^i; M) = p_c(x^i|\theta_{z^i})$ and $p(z^i|M) = \pi_{z^i}$. Therefore, substitution of these into the formula above results in

$$\log p(X, Z|M) = \sum_{i=1}^{n} \left(\log \pi_{z^i} + \log p_c(x^i|\theta_{z^i}) \right). \tag{8}$$

This form is not easier to work with when deriving M-steps. The general trick is to introduce an indicator function δ_k, which is defined by

$$\delta_k(z) = \begin{cases} 1 & (z = k) \\ 0 & (z \neq k). \end{cases} \tag{9}$$

With this notation, we can further rewrite the complete log-likelihood into

$$\log p(X, Z|M) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_k(z^i) \left(\log \pi_k + \log p_c(x^i|\theta_k) \right). \tag{10}$$

2. **Derive the E-steps.** Given the model parameters M and x_i, the indicator z_i is independent from other samples, and as a result, we have

$$p(Z|X; M) = \prod_{i=1}^{n} p(z^i|x^i; M). \tag{11}$$

Following the Bayes rule, we can get

$$p(z^i = k|x^i; M) = \frac{\pi_k p_c(x^i|\theta_k)}{\sum_{l=1}^{K} \pi_l p_c(x^i|\theta_l)}. \tag{12}$$

The model parameters M are updated at each iteration t. To simplify notation, we use $M^{(t)}$ to denote the model parameters obtained at iteration t, and $q^{(t)}$ to denote the posterior distribution obtained based on $M^{(t-1)}$. Then, we have

$$q^{(t)}(Z) = \prod_{i=1}^{n} q_i^{(t)}(z^i), \quad \text{with } q_i^{(t)}(z^i) = p(z^i|x^i; M^{(t-1)}). \tag{13}$$

3. **Derive the M-steps.** With respect to $q^{(t)}$, the expectation of the complete log-likelihood is given by

$$\mathrm{E}_{q^{(t)}}\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_k(z^i)\left(\log\pi_k+\log p_c(x^i|\theta_k)\right)\right)=\sum_{i=1}^{n}\sum_{k=1}^{K}\mathrm{E}_{q^{(t)}}\left(\delta_k(z^i)(\log\pi_k+\log p_c(x^i|\theta_k))\right). \tag{14}$$

   Here, each term has

$$\mathrm{E}_{q^{(t)}}\left(\delta_k(z^i)(\log\pi_k+\log p_c(x^i|\theta_k))\right)=\mathrm{E}_{q^{(t)}}\left(\delta_k(z^i)\right)\cdot(\log\pi_k+\log p_c(x^i|\theta_k)), \tag{15}$$

   and since the value of $\delta_k(z^i)$ can only be either 0 or 1, we have

$$\mathrm{E}_{q^{(t)}}\left(\delta_k(z^i)\right)=\mathrm{Pr}_{q^{(t)}}(\delta_k(z^i)=1)=q^{(t)}(k). \tag{16}$$

   Consequently, the expected complete log-likelihood can be written into

$$\sum_{i=1}^{n}\sum_{k=1}^{K}q_i^{(t)}(k)\left(\log\pi_k+\log p_c(x^i|\theta_k)\right). \tag{17}$$

   Hence, the optimal value of θ_k at time t, denoted by $\hat{\theta}_k^{(t)}$, can thus be obtained by

$$\hat{\theta}_k^{(t)}=\underset{\theta}{\mathrm{argmax}}\sum_{i=1}^{n}q_i^{(t)}(k)\log p_c(x^i|\theta_k). \tag{18}$$

   This has a similar form as the ordinary MLE problem, except that the term for each sample is modulated by a coefficient $q_i^{(t)}$. In addition, we have to estimate the value of π_k, which can be done by solving the following problem

$$\hat{\pi}^{(t)}=\underset{\pi}{\mathrm{argmax}}\sum_{i=1}^{n}q_i^{(t)}(k)\log\pi_k,\qquad \text{s.t.}\sum_{k=1}^{K}\pi_k=1,\ \pi_k\geq 0,\ \forall k=1,\dots,K. \tag{19}$$

   It is not difficult to derive (with Lagrange multipliers) that the optimal solution to this problem is given by

$$\hat{\pi}^{(t)}=\frac{1}{n}\sum_{i=1}^{n}q_i^{(t)}(k). \tag{20}$$

Summary: the E-step is to calculate the posterior probabilities of z^i, as

$$q_i^{(t)}(z^i)=p(z^i|x^i;M^{(t-1)})=\frac{\pi_k p_c(x^i|\theta_k)}{\sum_{l=1}^{K}\pi_l p_c(x^i|\theta_l)}. \tag{21}$$

The M-step is to optimize the model parameters, as

$$\hat{\theta}_k^{(t)}=\underset{\theta}{\mathrm{argmax}}\sum_{i=1}^{n}q_i^{(t)}(k)\log p_c(x^i|\theta_k), \tag{22}$$

and

$$\hat{\pi}_k^{(t)}=\frac{1}{n}\sum_{i=1}^{n}q_i^{(t)}(k). \tag{23}$$

In general, Eq.(21) and Eq.(23) are the same for all finite mixture models, while Eq.(22) depends on the particular form of the component model. Here are some examples:

1. **Univariate gaussian distribution.** Here, $\theta_k=(\mu_k,\sigma_k^2)$, and the pdf is given by

$$p_c(x|\theta_k)=\frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right). \tag{24}$$

   The optimal solution in the M-step is given by

$$\hat{\mu}_k^{(t)}=\frac{1}{\pi_k^{(t)}}\sum_{i=1}^{n}q_i^{(t)}(k)x^i,\quad\text{and}\quad\left(\hat{\sigma}_k^{(t)}\right)^2=\frac{1}{\pi_k^{(t)}}\sum_{i=1}^{n}q_i^{(t)}(x^i-\hat{\mu}_k^{(t)})^2. \tag{25}$$

   Here, $\pi_k^{(t)}=\sum_{i=1}^{n}q_i^{(t)}(k)$.

2. **Multivariate gaussian distribution over R^d.** Here, $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and the pdf is given by

$$p_c(\mathbf{x}|\theta_k) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right). \tag{26}$$

   The optimal solution in the M-step is given by

$$\hat{\boldsymbol{\mu}}_k^{(t)} = \frac{1}{\pi_k^{(t)}} \sum_{i=1}^n q_i^{(t)}(k)\mathbf{x}^i, \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_k^{(t)} = \frac{1}{\pi_k^{(t)}} \sum_{i=1}^n q_i^{(t)}(\mathbf{x}^i - \hat{\boldsymbol{\mu}}_k^{(t)})(\mathbf{x}^i - \hat{\boldsymbol{\mu}}_k^{(t)})^T. \tag{27}$$

3. **Simple topic model.** Suppose each component model is a topic characterized by a word distribution θ_k, with $\theta_k(w)$ being the probability of generating the word w from the topic, and each sample x is a bag of words, with $x(w)$ being the number of times the word w appears in the bag. Then

$$p_c(x|\theta_k) = \prod_{w \in \mathcal{W}} \theta_k(w)^{x(w)}. \tag{28}$$

   Here, $\mathcal{W}$ is the vocabulary, *i.e.* the set of all possible words. Under this model, the optimal solution in M-step is given by

$$\hat{\theta}_k(w) = \frac{1}{N} \sum_{i=1}^n q_i^{(t)} x^i(w). \tag{29}$$

   Here, N is the total number of words in all observed bags.

3 An Optimization View: Relating E-M and MLE

In this section, we show that the E-M algorithm is actually maximizing the following objective function via coordinate ascend.

$$Q(\theta; q) = \mathrm{E}_q\left(\log p(X, Z|\theta)\right) + H(q) = \mathrm{E}_q\left(\log p(X, Z|\theta)\right) - \mathrm{E}_q(\log q(Z)). \tag{30}$$

Here, q is a distribution over Z, and θ is the model parameter. $H(q)$ is called the *entropy* of the distribution q, which is defined to be

$$H(q) \triangleq -\mathrm{E}_q(\log q(Z)) = -\sum_z q(z)\log q(z). \tag{31}$$

The complete log likelihood here can be decomposed into

$$\log p(X, Z|\theta) = \log p(X|\theta) + \log p(Z|X;\theta). \tag{32}$$

We substitute Eq.(32) into Eq.(30), leading to

$$Q(\theta; q) = \mathrm{E}_q(\log p(X|\theta)) + \mathrm{E}_q(\log p(Z|X;\theta)) - \mathrm{E}_q(\log q(Z)) = \log p(X|\theta) - KL(q(Z)||p(Z|X;\theta)). \tag{33}$$

Therefore, with θ fixed, we have

$$\hat{q} = \underset{q}{\operatorname{argmax}}\, Q(\theta; q) = \underset{q}{\operatorname{argmax}}\, \mathrm{E}_q(\log p(Z|X;\theta)) - \mathrm{E}_q(\log q(Z)). \tag{34}$$

Here,

$$\mathrm{E}_q(\log p(Z|X;\theta)) - \mathrm{E}_q(\log q(Z)) = \mathrm{E}_q\left(\log p(Z|X;\theta) - \log q(Z)\right) = -KL\left(q(Z)||p(Z|X;\theta)\right). \tag{35}$$

Therefore, maximizing $Q(\theta, q)$ with θ fixed is equivalent to minimizing the Kullback-Leibler divergence between $q(Z)$ and the posterior distribution $p(Z|X;\theta)$, which attains the minimum (zero), when $q(Z) = p(Z|X;\theta)$. In other words, the optimal solution q to this problem is given by the posterior distribution $p(Z|X;\theta)$. Recall that the E-step is to compute this posterior distribution. In this sense, the E-step can be considered as the step to maximize $Q(\theta; q)$ with respect to q, with θ fixed. From Eq.(30), we can see that with q fixed, the optimal θ is given by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta; q) = \underset{\theta}{\operatorname{argmax}}\, \mathrm{E}_q\left(\log p(X, Z|\theta)\right). \tag{36}$$

This is to maximize the expectation of the complete likelihood, which is exactly what the M-step is doing. From the analysis above, we can see that the E-M is actually a coordinate ascend procedure to maximize $Q(\theta; q)$. In particular, the E-step is to optimize q with θ fixed, while the M-step is to optimize θ with q fixed. Consequently, the value of $Q(\theta; q)$ will keep increasing during this process, until it reaches a local optima.

To gain further insight, we revisit Eq.(33), and see that $Q(\theta; q)$ equals the marginal log-likelihood of X with respect to θ minus the KL divergence between $q(Z)$ and $p(Z|X; \theta)$. Since KL-divergence is always non-negative, $Q(\theta; q)$ is a lower bound of θ, and the E-M is to maximize this lower bound. Furthermore, it is worth noting that at each E-step, when we obtain the optimal q by minimizing the KL divergence to zero, we close the gap between the lower bound and the true marginal log-likelihood with respect to $\theta^{(t-1)}$. To make it explicit,

$$Q(\hat{q}^{(t)}, \hat{\theta}^{(t-1)}) = \log p(X|\theta^{(t-1)}). \tag{37}$$

Since $Q(\hat{q}^{(t)}, \hat{\theta}^{(t-1)}) \leq Q(\hat{q}^{(t)}, \hat{\theta}^{(t)}) \leq Q(\hat{q}^{(t+1)}, \hat{\theta}^{(t)})$, we have

$$\log p(X|\theta^{(t-1)}) \leq \log p(X|\theta^{(t)}). \tag{38}$$

This shows that the actual value of the marginal log-likelihood of X increases as t increases, implying that the E-M algorithm is actually maximizing the marginal log-likelihood of X, *i.e.* $\log p(X|\theta)$, the goal of MLE.