Tutorial on Generalized Expectation Maximization

Javier R. Movellan

1 Preliminaries

The goal of this primer is to introduce the EM (expectation maximization) algorithm and some of its modern generalizations, including variational approximations.

Notational conventions Unless otherwise stated, capital letters are used for random variables, small letters for specific values taken by random variables, and Greek letters for model parameters. We adhere to a Bayesian framework and treat model parameters as random variables with known prior. From this point of view Maximum-Likelihood methods can be interpreted as using weak priors. The probability space in which random variables are defined is left implicit and assumed to be endowed with the conditions needed to support the derivations being presented. We present the results using discrete random variables. Conversion to continuous variables simply requires changing probability mass functions into probability density fucntions and sums into integrals. When the context makes it clear, we identify probability functions by their arguments, and drop commas between arguments: e.g., $p(xy)$ is shorthand for the joint probability mass or joint probability density that the random variable X takes the specific value x and the random variable Y takes the value y.

Let O, H be random vectors representing observable data and hidden states. Let Λ represent model parameters controlling the distribution of O, H. We treat Λ as a random variable with known prior. We have two problems of interest:

- For a fixed sample o from O find values of Λ with large posterior
- For a fixed sample o from O find values of H with large posterior

Both problems are formally identical so we will focus on the first one. Note

$$p(\lambda \mid o) = \frac{1}{p(o)} p(o, \lambda) \tag{1}$$

Thus

$$\operatorname*{argmax}_{\lambda} p(\lambda \mid o) = \operatorname*{argmax}_{\lambda} \log p(o, \lambda) \tag{2}$$

Let $q = \{q_\theta(\cdot \mid o) : \theta \in \Re^p\}$ be a family of distributions of H parameterized by θ. We call q a variational family, and θ the variational parameters of that family. Note

$$\log p(o, \lambda) = \sum_h q_\theta(h \mid o) \log p(o, \lambda) \tag{3}$$

$$= \sum_h q_\theta(h \mid o) \log \frac{p(oh\lambda)}{q_\theta(h \mid o)} \frac{q_\theta(h \mid o)}{p(h \mid o, \lambda)} \tag{4}$$

$$= \mathcal{F}(\theta, \lambda) + K(\theta, \lambda) \tag{5}$$

where

$$\mathcal{F}(\theta, \lambda) \stackrel{\text{def}}{=} \sum_h q_\theta(h \mid o) \log \frac{p(o, h, \lambda)}{q_\theta(h \mid o)} \tag{6}$$

$$K(\theta, \lambda) \stackrel{\text{def}}{=} \sum_h q_\theta(h \mid o) \log \frac{q_\theta(h \mid o)}{p(h \mid o, \lambda)} \tag{7}$$

Note $K(\theta, \lambda)$ is the KL divergence between the distribution $q_\theta(\cdot \mid o)$ and $p(\cdot \mid o\ \theta)$. Since KL divergences are non-negative, it follows that $\mathcal{F}(\theta, \lambda)$ is a lower bound on $\log(o, \lambda)$,i.e.,

$$\log p(o, \lambda) \geq \mathcal{F}(\theta, \lambda) \tag{8}$$

This equation becomes an equality for values of θ for which $K(\theta, \lambda) = 0$, i.e., values of θ such that $q_\theta(h \mid o) = p(h \mid o\, \lambda)$ for all h.

2 The Generalized EM algorithm

We obtain a sequence, $(\lambda^{(1)}, \theta^{(1)}), (\lambda^{(2)}, \theta^{(2)}) \cdots$ by iteration over two steps:

- **E Step:**
$$\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{F}(\theta, \lambda^{(k)}) \tag{9}$$
Note since
$$\mathcal{F}(\theta, \lambda) = \log p(o, \lambda) + K(\theta, \lambda) \tag{10}$$
and since $\log p(o, \lambda)$ is a constant with respect to θ, this step amounts to minimizing $K(\theta, \lambda^{(k)})$ with respect to θ, i.e., choose a member of the variational family q which is as close as possible to the current p.
- **M Step:**
$$\lambda^{(k+1)} = \underset{\lambda}{\operatorname{argmax}}\, \mathcal{F}(\theta^{k+1}, \lambda) \tag{11}$$

Successive application of EM maximize the lower bound $\mathcal{F}$ on $\log p(o, \lambda)$, i.e,

$$\mathcal{F}(\theta^{(k+1)}, \lambda^{(k)}) \geq \mathcal{F}(\theta^{(k)}, \lambda^{(k)}) \tag{12}$$

and

$$\mathcal{F}(\theta^{(k+1)}, \lambda^{(k+1)}) \geq \mathcal{F}(\theta^{(k+1)}, \lambda^{(k)}) \tag{13}$$

2.1 Interpretation

- Optimizing $\mathcal{F}(\theta, \lambda)$ with respect to λ is equivalent to optimizing
$$\sum_h q_\theta(h \mid o) \log p(o, h, \lambda) \tag{14}$$
and since the $\log$ function is concave from below then
$$\sum_h q_\theta(h \mid o) \log p(o, h, \lambda) \leq \log \sum_h p(o, h, \lambda) = \log p(o, \lambda) \tag{15}$$
- Successive applications of EM increase a lower bound $\mathcal{F}$ on $\log p(o, \lambda)$.
- This lower bound consists of the sum of two terms: a data driven term $\log p(o, \lambda)$ that measures how well the distribution $p(\cdot\lambda)$ fits the observable data, and the term $KL(\theta, \lambda)$ that penalizes deviations from the variational family q:
$$\mathcal{F}(\theta, \lambda) = \log p(o, \lambda) - K(\theta, \lambda) \tag{16}$$
Thus we can think of the Genearlized EM algorithm as solving a penalized maximum likelihood problem.
- Note
$$\log p(o, \lambda^{(k+1)}) - \log p(o, \lambda^{(k)}) \geq K(\theta^{(k+1)}, \lambda^{(k+1)}) - K(\theta^{(k+1)}, \lambda^{(k)}) \tag{17}$$
Note $q_{\theta^{(k+1)}}$ was chosen to be closest to $p(\cdot|o, \lambda^{(k)})$. Thus it is not unreasonable (but also not guaranteed) to expect that it may not be as close to $p(\cdot|o, \lambda^{(k)})$. In other words, it is not unreasonalbe (but also not guaranteed) to expect that
$$K(\theta^{(k+1)}, \lambda^{(k+1)}) - K(\theta^{(k+1)}, \lambda^{(k)}) \geq 0 \tag{18}$$
and thus
$$\log p(o, \lambda^{(k+1)}) \geq \log p(o, \lambda^{(k)}) \tag{19}$$

- An important special case occurs when the family $\{q_\theta(\cdot \mid o)\}$ equals the family $\{p(\cdot \mid o, \lambda)\}$. In this case $\theta^{(k+1)} = \lambda^{(k)}$ and we can guarantee that

$$\log p(o, \lambda^{(k+1)}) - \log p(o, \lambda^{(k)}) \geq K(\lambda^{(k)}, \lambda^{(k+1)}) \geq 0 \tag{20}$$

  Moreover in this case to maximize $\mathcal{F}$ with respect to $\lambda^{(k+1)}$ we just need to maximize

$$Q(\lambda^{(k)}, \lambda^{(k+1)}) \stackrel{\text{def}}{=} \sum_h p(h \mid o, \lambda^{(k)}) \log p(\lambda^{(k+1)}) p(o, h \mid \lambda^{(k+1)}) \tag{21}$$

  and if we use an uninformative priors, then we just need to

$$Q(\lambda^{(k)}, \lambda^{(k+1)}) \stackrel{\text{def}}{=} \sum_h p(h \mid o, \lambda^{(k)}) \log p(o, h \mid \lambda^{(k+1)}) \tag{22}$$

  which is the objective function maximized by the standard EM algorithm.

- In the same vein, note that $\mathcal{F}(\theta, \lambda)$ is a free energy, i.e., the expected energy of states plus the entropy of the distribution under which the expected value is computed. In this case the energy of a state h is $-\log p(o, h, \lambda)$. Thus if there are no further constraints, the optimal distribution $q(\cdot \mid o, \theta)$ is Boltzmann

$$p(h \mid o, \theta) \propto \exp(\log p(o, h, \lambda)) = p(o, h, \lambda)) \tag{23}$$

$$p(h \mid o, \theta) = \frac{p(o, h, \lambda))}{\sum_h p(o, h, \lambda)} = p(h \mid o, \lambda) \tag{24}$$

- Consider the case in which we are given a set of iid observations $o = (o_1, \cdot o_n)$. If we directly optimize $\log p(o \mid \lambda)$ with respect to λ we get

$$\nabla_\lambda \log p(o \mid \lambda) = \sum_{i=1}^n \nabla_\lambda \log p(o_i \mid \lambda) = \sum_{i=1}^n \frac{1}{p(o_i \mid \lambda)} \sum_h \nabla_\lambda p(o_i, h \mid \lambda) \tag{25}$$

$$= \sum_{i=1}^n \frac{1}{p(o_i \mid \lambda)} \sum_h p(o_i, h \mid \lambda) \nabla_\lambda \log p(o_i, h \mid \lambda) \tag{26}$$

$$= \sum_{i=1}^n \sum_h p(h \mid o_i, \lambda) \nabla_\lambda \log p(o_i, h \mid \lambda) = 0 \tag{27}$$

  In contrast, when using the EM method we have

$$\nabla_\lambda Q(\lambda, \tilde{\lambda}) = \sum_{i=1}^n \sum_h p(h \mid o_i, \tilde{\lambda}) \nabla_\lambda \log p(o_i, h \mid \lambda) = 0 \tag{28}$$

  where $\tilde{\lambda}$ and thus $p(h \mid o_i, \tilde{\lambda})$ is no longer a function of λ.

3 Example 1

Consider a simple Gaussian mixture model and a vector of independent observations $o = (o_1, \ldots, o_n)^T$ from that model

$$\log p(o \mid \lambda) = \sum_{i=1}^n \log p(o_i \mid \lambda) \tag{29}$$

where

$$p(o_i \mid \lambda) = (1-\pi)p(o_i \mid H=0,\lambda) + \pi p(o_i \mid H=1,\lambda) = (1-\pi)g(o_i,0) + \pi(o_i,\lambda) \tag{30}$$

$$g(o_i,\lambda) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(o_i-\lambda)^2} \tag{31}$$

where the prior mixture term π is fixed. Taking derivatives with respect to λ we get

$$\frac{\partial \log p(o \mid \lambda)}{\partial \lambda} = \sum_i \frac{1}{p(o_i \mid \lambda)} \pi p(o_i \mid H=1,\lambda)(\lambda - o_i) \tag{32}$$

$$= \sum_i p(H=1 \mid o_i,\lambda)\ (\lambda - o_i) = 0 \tag{33}$$

which is a non-linear equation difficult to solve. However EM asks us to optimize

$$\sum_i p(H=1 \mid o_i,\tilde{\lambda}) \log p(o_i, H=1 \mid \lambda) = \tag{34}$$

Taking derivatives we get

$$\sum_i p(H=1 \mid o_i,\tilde{\lambda})\ (\lambda - o_i) = 0 \tag{35}$$

which is easily solved

$$\lambda = \frac{\sum_i o_i\ p(H=1 \mid o_i,\tilde{\lambda})}{\sum_i p(H=1 \mid o_i,\tilde{\lambda})} \tag{36}$$

4 History

- The first version of this document was written by Javier R. Movellan in January 2005, as part of the Kolmogorov project.