EECS 545: Machine Learning

Expectation Maximization

Benjamin Bray

Wednesday, March 16, 2016*

Abstract

These notes provide a theoretical treatment of **Expectation-Maximization**, an iterative parameter estimation algorithm used to find local maxima of the likelihood function in the presence of hidden variables. Introductory textbooks [3, 1] typically state the algorithm without explanation and expect students to work blindly through derivations. We find this approach to be unsatisfying, and instead choose to tackle the theory head-on, followed by plenty of examples. Following [4], we view expectation-maximization as coordinate ascent on the **Expectation Lower Bound**. This perspective takes much of the "mystery" out of the algorithm and allows us to easily derive variants like **Hard EM** and **Variational EM**.

Note: I actually wrote these notes for myself when learning about expectation maximization for the first time. Accordingly, they have only been lightly proofread, and some parts are incomplete. Send an email to `benrbray@umich.edu` *if you spot any big mistakes.*

Part 1

Expectation Maximization

Suppose we observe data $\mathcal{X}$ generated from a model p with true parameters θ^* in the presence of hidden variables Z. As usual, we wish to compute the maximum likelihood estimate

$$\hat{\theta}_{ML} = \arg\max_{\theta} \ell(\theta|\mathcal{X}) = \arg\max_{\theta} \log p(\mathcal{X}|\theta) \tag{1}$$

of the parameters given our observed data. In some cases, we also seek to *infer* the values $\mathcal{Z}$ of the hidden variables Z. In the Bayesian spirit, we will treat the parameter θ^* as a realization of some random variable Θ.

The observed data log-likelihood $\ell(\theta|\mathcal{X}) = \log p(\mathcal{X}|\theta)$ of the parameters given the observed data is useful for both inference and parameter estimation, in which we must grapple with uncertainty about the hidden variables. Working directly with this quantity is often difficult in latent variable models because the inner sum cannot be brought out of the

Last modified: Tuesday, March 15, 2016

logarithm when we marginalize over the latent variables:

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

In general, this likelihood is non-convex with many local maxima. In contrast, [3] shows that when $p(x_n, z_n|\theta)$ are exponential family distributions, the likelihood is convex, so learning is much easier. Expectation maximization exploits the fact that learning is easy when we observe all variables. We will alternate between inferring the values of the latent variables and re-estimating the parameters, assuming we have complete data.

1.1 Evidence Lower Bound

Our general approach will be to reason about the hidden variables through a proxy distribution q, which we use to compute a lower-bound on the log-likelihood. This section is devoted to deriving one such bound, called the **Evidence Lower Bound (ELBO)**.

We can expand the data log-likelihood by marginalizing over the hidden variables:

$$\ell(\theta|\mathcal{X}) = \log p(\mathcal{X}|\theta) = \log \sum_z p(\mathcal{X}, z|\theta) \tag{3}$$

Through Jensen's inequality, we obtain the following bound:

$$\begin{aligned}
\ell(\theta|\mathcal{X}) &= \log \sum_z p(\mathcal{X}, z|\theta) && (4)\\
&= \log \sum_z q(z) \frac{p(\mathcal{X}, z|\theta)}{q(z)} && (5)\\
&\geq \sum_z q(z) \log \frac{p(\mathcal{X}, z|\theta)}{q(z)} \equiv \mathcal{L}(q, \theta) && (6)
\end{aligned}$$

The lower bound $\mathcal{L}(q, \theta)$ can be rewritten as follows:

$$\begin{aligned}
\ell(\theta|\mathcal{X}) \geq \mathcal{L}(q, \theta) &= \sum_z q(z) \log \frac{p(\mathcal{X}, z|\theta)}{q(z)} && (7)\\
&= \sum_z q(z) \log p(\mathcal{X}, z|\theta) - \sum_z q(z) \log q(z) && (8)\\
&= E_q[\log p(\mathcal{X}, Z|\theta)] - E_q[\log q(z)] && (9)\\
&= E_q[\log p(\mathcal{X}, Z|\theta)] + H(q) && (10)
\end{aligned}$$

Relationship to Relative Entropy

The first term in the last line above closely resembles the cross entropy between $q(Z)$ and the joint distribution $p(X, Z)$ of the observed and hidden variables. However, the variables X are fixed to our observations $X = \mathcal{X}$ and so $p(\mathcal{X}, Z)$ is an *unnormalized*[1] distribution

[1] In this case, $\int p(\mathcal{X}, z)\, dz \neq 1$.

over Z. It is easy to see that this does not set us back too far; in fact, the lower bound $\mathcal{L}(q, \theta)$ differs from a Kullback-Liebler divergence only by a constant with respect to Z:

$$\begin{aligned} D_{KL}(q||p(Z|\mathcal{X}, \theta)) &= H(q, p(Z|\mathcal{X}, \theta)) - H(q) && (11)\\ &= E_q[-\log p(Z|\mathcal{X}, \theta)] - H(q) && (12)\\ &= E_q[-\log p(\mathcal{X}, Z|\theta)] - E_q[-\log p(\mathcal{X}|\theta)] - H(q) && (13)\\ &= E_q[-\log p(\mathcal{X}, Z|\theta)] + \log p(\mathcal{X}|\theta) - H(q) && (14)\\ &= -\mathcal{L}(q, \theta) + \text{const.} && (15) \end{aligned}$$

This yields a second proof of the evidence lower bound, following from the nonnegativity of relative entropy. In fact, this is the proof given in [5] and [3].

$$\log p(\mathcal{X}|\theta) = D_{KL}(q||p(Z|\mathcal{X}|\theta)) + \mathcal{L}(q, \theta) \geq \mathcal{L}(q, \theta) \tag{16}$$

Selecting a Proxy Distribution

The quality of our lower bound $\mathcal{L}(q, \theta)$ depends heavily on the choice of proxy distribution $q(Z)$. We now show that the evidence lower bound is *tight* in the sense that equality holds when the proxy distribution $q(Z)$ is chosen to be the hidden posterior $p(Z|\mathcal{X}, \theta)$. This will be useful later for proving that the Expectation Maximization algorithm converges.

Remark 1. *Maximizing $\mathcal{L}(q, \theta)$ with respect to q is equivalent to minimizing the relative entropy between q and the hidden posterior $p(Z|\mathcal{X}, \theta)$. Hence, the optimal choice for q is exactly the hidden posterior, for which $D_{KL}(q||p(Z|\mathcal{X}, \theta)) = 0$, and*

$$\log p(\mathcal{X}|\theta) = E_q[\log p(\mathcal{X}, Z|\theta)] + H(q) = \mathcal{L}(q, \theta)$$

1.2 Expectation Maximization

Recall that the maximum likelihood estimate of the parameters θ given observed data $\mathcal{X}$ in the presence of hidden variables Z is

$$\hat{\theta}_{ML} = \arg\max_{\theta} \ell(\theta|\mathcal{X}) = \arg\max_{\theta} \log p(\mathcal{X}|\theta) \tag{17}$$

Unfortunately, when reasoning about hidden variables, finding a global maximum is difficult. Instead, the **Expectation Maximization** algorithm is an iterative procedure for computing a local maximum of the likelihood function, under the assumption that the hidden posterior $p(Z|\mathcal{X}, \theta)$ is tractable. We will take advantage of the evidence lower bound

$$\ell(\theta|\mathcal{X}) \geq \mathcal{L}(q, \theta) \tag{18}$$

on the data likelihood. Consider only proxy distributions of the form $q_\vartheta(Z) = p(Z|\mathcal{X}, \vartheta)$, where ϑ is some fixed configuration of the variables Θ, possibly different from our estimate θ. The optimal value for ϑ, in the sense that $\mathcal{L}(q_\vartheta, \theta)$ is maximum, depends on the particular choice of θ. Similarly, the optimal value for θ depends on the choice of ϑ. This suggests an iterative scheme in which we alternate between maximizing with respect to ϑ and with respect to θ, gradually improving the log-likelihood.

Iterative Procedure

Suppose at time t we have an estimate θ_t of the parameters. To improve our estimate, we perform two steps of coordinate ascent on $\mathcal{L}(\vartheta, \theta) \equiv \mathcal{L}(q_\vartheta, \theta)$, as described in [4],

E-Step Compute a new lower bound on the observed log-likelihood, with

$$\vartheta_{t+1} = \arg\max_{\vartheta} \mathcal{L}(\vartheta, \theta_t) = \theta_t$$

M-Step Estimate new parameters by optimizing over the lower bound,

$$\theta_{t+1} = \arg\max_{\theta} \mathcal{L}(\vartheta_{t+1}, \theta) = \arg\max_{\theta} E_q[\log p(\mathcal{X}, Z|\theta)]$$

In the M-Step, the expectation is taken with respect to $q_{\vartheta_{t+1}}$.

Alternative Formulation

In the M-Step, the entropy term of the evidence lower bound $\mathcal{L}(\vartheta_{t+1}, \theta)$ does not depend on θ. The remaining term $Q(\theta_t, \theta) = E_q[\log p(\mathcal{X}, Z|\theta)]$ is sometimes called the **auxiliary function** or **Q-function**. To us, this is the **expected complete-data log-likelihood**.

Proof of Convergence

To prove convergence of this algorithm, we show that the data likelihood $\ell(\theta|\mathcal{X})$ increases after each update.

Theorem 1. *After a single iteration of Expectation Maximization, the observed data likelihood of the estimated parameters has not decreased, that is,*

$$\ell(\theta_t|\mathcal{X}) \leq \ell(\theta_{t+1}|\mathcal{X})$$

Proof. This result is a simple consequence of all the hard work we have put in so far:

$$\begin{aligned} \ell(\theta_t|\mathcal{X}) &= \mathcal{L}(q_{\vartheta_{t+1}}, \theta_t) && \text{(Remark 1)} \\ &\leq \mathcal{L}(q_{\vartheta_{t+1}}, \theta_{t+1}) && \text{(M-Step)} \\ &\leq \ell(\theta_{t+1}|\mathcal{X}) && \text{(ELBO)} \end{aligned}$$

□

It is also possible to show that Expectation-Maximization converges to something *useful*.

Theorem 2. *(Neal & Hinton 1998, Thm. 2) Every local maximum of the evidence lower bound $\mathcal{L}(q, \theta)$ is a local maximum of the data likelihood $\ell(\theta|\mathcal{X})$.*

Starting from an initial guess θ_0, We run this procedure until some stopping criterion is met and obtain a sequence $\{(\vartheta_t, \theta_t)\}_{t=1}^{T}$ of parameter estimates.

1.3 Example: Coin Flips

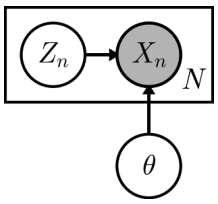
Now that we have a good grasp on the theory behind Expectation Maximization, let's get some intuition by means of a simple example. As usual, the simplest possible example involves coin flips!

Probabilistic Model

Suppose we have two coins, each with a different probability of heads, θ_A and θ_B, unknown to us. We collect data from a series of N trials in order to estimate the bias of each coin. Each trial k consists of flipping the same random coin Z_k a total of M times and recording only the total number X_k of heads.

This situation is best described by the following **generative probabilistic model**, which precisely describes our assumptions about how the data was generated. The corresponding graphical model and a set of sample data are shown in Figure 1.

$$\begin{aligned} \theta &= (\theta_A, \theta_B) && && \text{fixed coin biases} \\ Z_n &\sim \text{Uniform}\{A, B\} && \forall\, n = 1, \ldots, N && \text{coin indicators} \\ X_n | Z_n, \theta &\sim \text{Bin}[\theta_{Z_n}, M] && \forall\, n = 1, \ldots, N && \text{head count} \end{aligned} \tag{19}$$



#	Sequence	Heads
1	HTTTHHTHTH	5
2	HHHHTHHHHH	9
3	HTHHHHHTHH	8
4	HTHTTTHHTT	4
5	THHHTHHHTH	7

Figure 1: Sample data and graphical model representation for the coin flip example, with $N = 5$ trials and $M = 10$ flips per trial. Adapted from [2].

Complete Data Log-Likelihood

The complete data log-likelihood for a single trial (x_n, z_n) is

$$\log p(x_n, z_n | \theta) = \log p(z_n) + \log p(x_n | z_n, \theta) \tag{20}$$

In this model, $P(z_n) = \frac{1}{2}$ is uniform. The remaining term is

$$\log p(x_n | z_n, \theta) = \log \binom{M}{x_n} \theta_{z_n}^{x_n} (1 - \theta_{z_n})^{M - x_n} \tag{21}$$

$$= \log \binom{M}{x_n} + x_n \log \theta_{z_n} + (M - x_n) \log(1 - \theta_{z_n}) \tag{22}$$

Expectation Maximization

Now that we have specified the probabilistic model and worked out all relevant probabilities, we are ready to derive an Expectation Maximization algorithm.

The **E-Step** is straightforward. The **M-Step** computes a new parameter estimate θ_{t+1} by optimizing over the lower bound found in the E-Step. Let $\vartheta = \vartheta_{t+1} = \theta_t$. Then,

$$\theta_{t+1} = \arg\max_\theta \mathcal{L}(\theta, q_\vartheta) = \arg\max_\theta E_q[\log p(\mathcal{X}, Z|\theta)] \tag{23}$$

$$= \arg\max_\theta E_q[\log p(\mathcal{X}|Z,\theta)p(Z)] \tag{24}$$

$$= \arg\max_\theta E_q[\log p(\mathcal{X}|Z,\theta)] + \log p(Z) \tag{25}$$

$$= \arg\max_\theta E_q[\log p(\mathcal{X}|Z,\theta)] \tag{26}$$

Now, because each trial is conditionally independent of the others, given the parameters,

$$E_q[\log p(\mathcal{X}|Z,\theta)] = E_q\left[\log \prod_{n=1}^N p(x_n|Z_n,\theta)\right] = \sum_{n=1}^N E_q[\log p(x_n|Z_n,\theta)] \tag{27}$$

$$= \sum_{n=1}^N E_q\Big[x_n \log\theta_{z_n} + (M-x_n)\log(1-\theta_{z_n})\Big] + \sum_{n=1}^N \log\binom{M}{x_n} \tag{28}$$

$$= \sum_{n=1}^N E_q\Big[x_n \log\theta_{z_n} + (M-x_n)\log(1-\theta_{z_n})\Big] + \text{const. w.r.t. } \theta \tag{29}$$

$$= \sum_{n=1}^N q_\vartheta(z_n = A)\Big[x_n\log\theta_A + (M-x_n)\log\theta_A\Big] \tag{30}$$

$$+ \sum_{n=1}^N q_\vartheta(z_n = B)\Big[x_n\log\theta_B + (M-x_n)\log\theta_B\Big] + \text{const. w.r.t. } \theta \tag{31}$$

Let $a_k = q(z_k = A)$ and $b_k = q(z_k = B)$. Note $\sum_{k=1}^N a_k = \sum_{k=1}^N b_k = 1$. To maximize the above expression with respect to the parameters, we take derivatives with respect to θ_A and θ_B and set to zero:

$$\frac{\partial}{\partial\theta_A}\Big[E_q[\log p(\mathcal{X}|Z,\theta)]\Big] = \frac{1}{\theta_A}\sum_{n=1}^N a_n x_n + \frac{1}{1-\theta_A}\sum_{n=1}^N a_n(M-x_n) = 0 \tag{32}$$

$$\frac{\partial}{\partial\theta_B}\Big[E_q[\log p(\mathcal{X}|Z,\theta)]\Big] = \frac{1}{\theta_B}\sum_{n=1}^N b_n x_n + \frac{1}{1-\theta_B}\sum_{n=1}^N b_n(M-x_n) = 0 \tag{33}$$

$$\tag{34}$$

Solving for θ_A and θ_B, we obtain

$$\theta_A = \frac{\sum_{n=1}^N a_n x_n}{\sum_{n=1}^N a_n M} \qquad \theta_B = \frac{\sum_{n=1}^N b_n x_n}{\sum_{n=1}^N b_n M} \tag{35}$$

1.4 Example: Gaussian Mixture Model

Probabilistic Model

In a Gaussian Mixture Model, samples are drawn from a random *cluster*, each normally distributed with its own mean and variance. Our goal will be to estimate the following parameters:

$$\begin{aligned} \boldsymbol{\pi} &= (\pi_1, \ldots, \pi_K) && \text{mixing weights} \\ \boldsymbol{\mu} &= (\mu_1, \ldots, \mu_K) && \text{cluster centers} \\ \boldsymbol{\Sigma} &= (\Sigma_1, \ldots, \Sigma_K) && \text{cluster variance} \end{aligned} \tag{36}$$

The full model specification is below. A graphical model is shown in Figure 2.

$$\begin{aligned} \theta &= (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) && \text{model parameters} \\ z_n &\sim \text{Cat}[\pi] && \text{cluster indicators} \\ x_n | z_n, \theta &\sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n}) && \text{base distribution} \end{aligned} \tag{37}$$

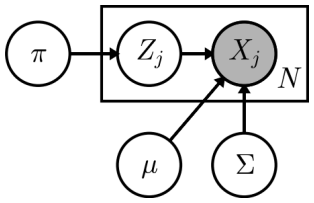


Figure 2: Gaussian Mixture Model.

Complete Data Log-Likelihood

The complete data log-likelihood for a single datapoint (x_n, z_n) is

$$\log p(x_n, z_n|\theta) = \log \prod_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)^{\mathbb{I}(z_n=k)} \tag{38}$$

$$= \sum_{k=1}^{K} \mathbb{I}(z_n = k) \log \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \tag{39}$$

Similarly, the complete data log-likelihood over all points $\{(x_n, z_n)\}_{n=1}^{N}$ is

$$\log p(X, Z|\theta) = \sum_{n=1}^{N} \log p(x_n, z_n|\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}(z_n = k) \log \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \tag{40}$$

Hidden Posterior

The hidden posterior for a single point (x_n, z_n) can be found using Bayes' rule:

$$p(z_n = k|x_n, \theta) = \frac{P(z_n = k|\theta) p(x_n|z_n = k, \theta)}{p(x_N|\theta)} \tag{41}$$

$$= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x_n|\mu_{k'}, \Sigma_{k'})} \tag{42}$$

Expectation Maximization

Our derivation will follow that of [3], adapted to our notation.

E-Step

Before the E-step, we have an estimate θ_t of the parameters, and seek to compute a new lower bound on the observed log-likelihood. Earlier, we showed that the optimal lower bound is

$$\mathcal{L}(q_{\theta_t}, \theta) = E_q[\log p(\mathcal{X}, Z|\theta)] + \text{const.} \tag{43}$$

where $q_{\theta_t}(z) \equiv p(z|\mathcal{X}, \theta_t)$ and the second term is constant with respect to θ. The E-Step requires us to derive an expression for the first term. Using Equation 40, the expected complete data log-likelihood is given by

$$Q(\theta_t, \theta) = E_q[\log p(\mathcal{X}, Z|\theta)] = \sum_{n=1}^{N}\sum_{k=1}^{K} E_q\big[\mathbb{I}(z_n = k)\log \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)\big] \tag{44}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} E_q\big[\mathbb{I}(z_n = k)\big]\log \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \tag{45}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_n = k \mid x_n, \theta_t)\log \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \tag{46}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log \pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \tag{47}$$

where $r_{nk} \equiv p(z_n = k \mid x_n, \theta_t)$ is the **responsibility** that cluster k takes for data point x_n after step t. During the E-Step, we compute these values explicitly with Equation 42.

M-Step

During the M-Step, we optimize our lower bound with respect to the parameters $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. For Gaussian Mixture Models, the lower bound is easy to maximize by taking derivatives and setting to zero. For the mixing weights $\boldsymbol{\pi}$,

$$\frac{\partial Q}{\partial \pi_k} = \sum_{n=1}^{N} \frac{r_{nk}}{\pi_k} = 0 \quad \Longrightarrow \quad \boxed{\pi_k = \frac{1}{N}\sum_{n=1}^{N} r_{nk} = \frac{r_k}{N}} \tag{48}$$

where $r_k \equiv \sum_{n=1}^{n} r_{nk}$ is the *effective* number of points assigned to cluster k. For the cluster centers $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, you should verify that the correct updates are

$$\boxed{\mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{r_k}} \qquad \boxed{\Sigma_k = \frac{\sum_{n=1}^{N} r_{nk} x_n x_n^T}{r_k} - \mu_k \mu_k^T} \tag{49}$$

1.5 Advice for Deriving EM Algorithms

The previous two examples suggest a general approach for deriving a new algorithm.

1. **Specify the probabilistic model.** Identify the observed variables, hidden variables, and parameters. Draw the corresponding graphical model to help determine the underlying independence structure.
2. **Identify the complete-data likelihood $P(X, Z|\theta)$.** For exponential family models, the complete-data likelihood will be convex and easy to optimize. In other models, other work may be required.
3. **Identify the hidden posterior $P(Z|X, \theta)$.** If this distribution is not tractable, you may want to consider variational inference, which we will discuss later.
4. **Derive the E-Step.** Write down an expression for $E_q[\log p(\mathcal{X}|Z, \theta)]$.
5. **Derive the M-Step.** Try taking derivatives and setting to zero. If this doesn't work, you may need to resort to gradient-based methods or variational inference.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN: 0387310738.

[2] Chuong B Do and Serafim Batzoglou. "What is the expectation maximization algorithm?" In: *Nature biotechnology* 26.8 (2008), pp. 897–899.

[3] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020, 9780262018029.

[4] Radford M Neal and Geoffrey E Hinton. "A view of the EM algorithm that justifies incremental, sparse, and other variants". In: *Learning in graphical models*. Springer, 1998, pp. 355–368.

[5] Dimitris G Tzikas, Aristidis C Likas, and Nickolaos P Galatsanos. "The variational approximation for Bayesian inference". In: *Signal Processing Magazine, IEEE* 25.6 (2008), pp. 131–146.