EXPECTATION MAXIMIZATION: A GENTLE INTRODUCTION

MORITZ BLUME

1. Introduction

This tutorial was basically written for students/researchers who want to get into first touch with the Expectation Maximization (EM) Algorithm. The main motivation for writing this tutorial was the fact that I did not find any text that fitted my needs. I started with the great book "Artificial Intelligence: A modern Approach"of Russel and Norvig [6], which provides lots of intuition, but I was disappointed that the general explication of the EM algorithm did not directly fit the shown examples. There is a very good tutorial written by Jeff Bilmes [1]. However, I felt that some parts of the mathematical derivation could be shortened by another choice of the hidden variable (explained later on). Finally I had a look at the book "The EM Algorithm and Extensions" by McLachlan and Krishnan [4]. I can definitely recommend this book to anybody involved with the EM algorithm - though, for a beginner, some steps could be more elaborate.

So, I ended up by writing my own tutorial. Any feedback is greatly appreciated (error corrections, spelling/grammar mistakes, etc.).

2. Short Review: Maximum Likelihood

The goal of Maximum Likelihood (ML) estimation is to find parameters that maximize the probability of having received certain measurements of a random variable distributed by some probability density function (p.d.f.). For example, we look at a random variable Y and a measurement vector $\boldsymbol{y} = (y_1, ..., y_N)^\mathsf{T}$. The probability of receiving some measurement y_i is given by the p.d.f.

$$p(y_i|\boldsymbol{\Theta}) \quad , \tag{1}$$

where the p.d.f. is governed by the parameter $\boldsymbol{\Theta}$. The probability of having received the whole series of measurements is then

$$p(\boldsymbol{y}|\boldsymbol{\Theta}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{\Theta}) \quad , \tag{2}$$

if the measurements are independent. The likelihood function is defined as a function of $\boldsymbol{\Theta}$:

$$L(\boldsymbol{\Theta}) = p(\boldsymbol{y}|\boldsymbol{\Theta}) \quad . \tag{3}$$

The ML estimate of $\boldsymbol{\Theta}$ is found by maximizing L. Often, it is easier to maximize the log-likelihood

Date: March 15, 2002.

$$\log L(\boldsymbol{\Theta}) = \log p(\boldsymbol{y}|\boldsymbol{\Theta}) \tag{4}$$

$$= \log \prod_{i=1}^{N} p(y_i|\boldsymbol{\Theta}) \tag{5}$$

$$= \sum_{i=1}^{N} \log p(y_i|\boldsymbol{\Theta}) \quad . \tag{6}$$

Since the logarithm is a strictly increasing function, the maximum of L and $\log(L)$ is the same.

It is important to note that we do not include any *a priori* knowledge of the parameter by calculating the ML estimate. Instead, we assume that each choice of the parameter vector is equally likely and therefore that the p.d.f. of the parameter vector is a uniform distribution. If such a prior p.d.f. for the parameter vector is available then methods of Bayesian parameter estimation should be preferred. In this tutorial we will only look at ML estimates.

In some cases a closed form can be derived by just setting the derivative with respect to $\boldsymbol{\Theta}$ to zero. However, things are not always as easy, as the following example will show.

Example: mixture distributions. We assume that a data point y_j is produced as follows: first, one of I distributions/components which produces the measurement is chosen. Then, according to the parameters of this distribution, the actual measurement is produced. There are I components. The probability density function (p.d.f.) of the mixture distribution is

$$p(y_j|\boldsymbol{\Theta}) = \sum_{i=1}^{I} \alpha_i p(y_j|C = i, \beta_i) \quad , \tag{7}$$

where $\boldsymbol{\Theta}$ is composed as $\boldsymbol{\Theta} = (\alpha_1^\mathsf{T}, ..., \alpha_I^\mathsf{T}, \beta_1^\mathsf{T}, ..., \beta_M^\mathsf{T})^\mathsf{T}$. The parameters α_i define the weight of component i and must sum up to 1, and the parameter vectors $\boldsymbol{\beta_i}$ are associated to the p.d.f. of component i.

Based on some measurements $\boldsymbol{y} = (y_1, ..., y_J)^\mathsf{T}$ of the underlying random variable Y, the goal of ML (Maximum Likelihood) estimation is to find the parameters $\boldsymbol{\Theta}$ that maximize the probability of having received these measurements. This involves maximization of the log-likelihood for $\boldsymbol{\Theta}$:

$$\log L(\boldsymbol{\Theta}) = \log p(\boldsymbol{y}|\boldsymbol{\Theta}) \tag{8}$$

$$= \log \{p(y_1|\boldsymbol{\Theta}) \cdot ... \cdot p(y_J|\boldsymbol{\Theta})\} \tag{9}$$

$$= \sum_{j=1}^{J} \log p(y_j|\boldsymbol{\Theta}) \tag{10}$$

$$= \sum_{j=1}^{J} \log \left\{ \sum_{i=1}^{I} \alpha_i p(y_j|C = i, \boldsymbol{\beta_i}) \right\} \quad . \tag{11}$$

Since the logarithm is outside the sum, optimization of this term is not an easy task. There does not exist a closed form for the optimal value of $\boldsymbol{\Theta}$.

3. The EM algorithm

The goal of the EM algorithm is to facilitate maximum likelihood parameter estimation by introducing so-called *hidden* random variables which are not observed and therefore define the unobserved data. Let us denote this hidden random variable by some Z, and the unobserved data by a vector $\boldsymbol{z}$. Note that it does not play a role if this unobserved data is technically unobservable or just not observed due to other reasons. Often, it is just an artificial construct in order to enable an EM scheme.

We define the complete data $\boldsymbol{x}$ to consist of the incomplete observed data $\boldsymbol{y}$ and the unobserved data $\boldsymbol{z}$:

$$\boldsymbol{x} = (\boldsymbol{y}^\mathsf{T}, \boldsymbol{z}^\mathsf{T})^\mathsf{T} \quad . \tag{12}$$

Then, instead of looking at the log-likelihood for the *observed data*, we consider the log-likelihood function of the *complete data*:

$$\log L_C(\boldsymbol{\Theta}) = \log p(\boldsymbol{x}|\boldsymbol{\Theta}) \quad . \tag{13}$$

Remember that in the ML case we just had a likelihood function and wanted to find the parameter that maximizes the probability of having obtained the observed data. The problem in the present case of the complete data likelihood function is that we did not observe the complete data ($\boldsymbol{z}$ is missing!).

It may be surprising, but this problem of dealing with unobserved data in the end facilitates calculation of the ML parameter estimate.

The EM algorithm faces this problem by first finding an estimate for the likelihood function, and then maximizing the whole term. In order to find this estimate, it is important to note that each entry of $\boldsymbol{z}$ is a realization of a hidden random variable. However, since these realizations do not exist in reality, we have to consider each entry of $\boldsymbol{z}$ as a random variable itself. So, the whole likelihood function is nothing else than a function of a random variable and therefore by itself is a random variable (a quantity which depends on a random variable is a random variable). Its expectation given the observed data $\boldsymbol{y}$ is:

$$\mathbb{E}\left[\log L_C(\boldsymbol{\Theta})|\boldsymbol{y}, \boldsymbol{\Theta}^{(i)}\right] \quad . \tag{14}$$

The EM algorithm maximizes this expectation:

$$\boldsymbol{\Theta}^{(i+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\, \mathbb{E}\left[\log L_C(\boldsymbol{\Theta})|\boldsymbol{y}, \boldsymbol{\Theta}^{(i)}\right] \quad . \tag{15}$$

Continuation of the example. In the case of our mixture distributions, a good choice for $\boldsymbol{z}$ is a matrix whose entry z_{ij} is equal to one if and only if component i produced measurement j - otherwise it is zero. Note that each column contains exactly one entry equal to one.

By this choice, we introduce additional information into the process: $\boldsymbol{z}$ tells us the probability density function that underlies a certain measurement. Nobody knows if we can really measure this hidden variable, and it is not important. We just assume that we have it, do some calculous, and see what happens...

Accordingly, the log-likelihood function of the complete data is:

$$\log L_C(\boldsymbol{\Theta}) = \log p(\boldsymbol{x}|\boldsymbol{\Theta}) \tag{16}$$

$$= \log \prod_{j=1}^{J} \sum_{i=1}^{I} z_{ij} \alpha_i p(y_j|C = i, \beta_i) \tag{17}$$

$$= \sum_{j=1}^{J} \log \sum_{i=1}^{I} z_{ij} \alpha_i p(y_j|C = i, \beta_i) \quad . \tag{18}$$

This can now be rewritten as

$$\log L_C(\boldsymbol{\Theta}) = \sum_{j=1}^{J} \sum_{i=1}^{I} z_{ij} \log \alpha_i p(y_j|C = i, \beta_i) \quad , \tag{19}$$

since z_{ij} is zero for all but one term in the inner sum!

The choice of this hidden variable might look like magic. In fact this is the most difficult part for defining an EM scheme. The intuition behind the choice of z_{ij} might be that if the information about the assigned components were available, then it would be easily possible to estimate the parameters - one just has to do a simple ML estimation for each component.

One also should note that even if he decides to incorporate the right information (here: the assignment of measurements to components), the simplicity of the further steps strongly depends on how this is modeled. For example, instead of indicator variables z_{ij} one could define random variables $\boldsymbol{z}_j$ that represent the value of the respective component for measurement j. However, this heavily complicates further calculations (as can be seen in [1]).

Compared to the ML approach which involves just maximizing a log likelihood, the EM algorithm makes one step in between: the calculation of the expectation. This is denoted as the expectation step. Please note that in order to calculate the expectation, an initial parameter estimate $\boldsymbol{\Theta}^{(i)}$ is needed. But still, the whole term will depend on a parameter $\boldsymbol{\Theta}$. In the maximization step, a new update $\boldsymbol{\Theta}^{(i+1)}$ for $\boldsymbol{\Theta}$ is found in such manner, that it maximizes the whole term. This whole procedure is repeated until some stopping criteria, e. g. until the difference of change between the parameter updates becomes very small.

EM for mixture distributions. Now, back to our example. The parameter update is found by solving

$$\boldsymbol{\Theta}^{(i+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \, \mathbb{E}\left[\log L_C(\boldsymbol{\Theta})|\boldsymbol{y}, \boldsymbol{\Theta}^{(i)}\right] \tag{20}$$

$$= \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \, \mathbb{E}\left[\sum_{j=1}^{J} \sum_{i=1}^{I} z_{ij} \log\left\{\alpha_i p(y_j\right\}|C = i, \beta_i) \middle| \boldsymbol{y}, \boldsymbol{\Theta}^{(i)}\right] \quad . \tag{21}$$

3.1. **Expectation Step.** The expectation step consists of calculating the expected value of the complete data likelihood function. Due to linearity of expectation we have

$$\boldsymbol{\Theta}^{(i+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{j=1}^{J} \sum_{i=1}^{I} \mathbb{E}\left[z_{ij} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] \log \alpha_i p(y_j | C = i, \beta_i) \quad . \tag{22}$$

In order to complete this step, we have to see how $\mathbb{E}\left[z_{ij} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right]$ looks like:

$$\mathbb{E}\left[z_{ij} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] = 0 \cdot p(z_{ij} = 0 | \boldsymbol{\Theta}^{(i)}) + 1 \cdot p(z_{ij} = 1 | \boldsymbol{\Theta}^{(i)}) = p(z_{ij} = 1 | \boldsymbol{y}, \boldsymbol{\Theta}^{(i)}) \quad . \tag{23}$$

This is the probability that class i produced measurement j. In other words: The probability that class i was active while measurement j was produced:

$$p(z_{ij} = 1) = p(C = i | x_j, \boldsymbol{\Theta}^{(i)}) \quad . \tag{24}$$

Application of Bayes theorem leads to

$$p(C = i | x_j, \boldsymbol{\Theta}^{(i)}) = \frac{p(x_j | C = i, \boldsymbol{\Theta}^{(i)}) p(C = i | \boldsymbol{\Theta}^{(i)})}{p(x_j | \boldsymbol{\Theta}^{(i)})} \quad . \tag{25}$$

All the terms on the right side can be easily calculated. $p(x_j | C = i, \boldsymbol{\Theta}^{(i)})$ just comes from the i-th components density function. $p(C = i | \boldsymbol{\Theta}^{(i)})$ is the probability of choosing the ith component, which is nothing else than the parameter α_i (which are already given by $\boldsymbol{\Theta}^{(i)}$). Finally, $p(x_j | \boldsymbol{\Theta}^{(i)})$ is the whole distributions probability density function, which, based on $\boldsymbol{\Theta}^{(i)}$, is easy to calculate.

3.2. **Maximization Step.** Remember that our parameter $\boldsymbol{\Theta}$ is composed of the α_k and $\boldsymbol{\beta}_k$. We will perform the optimization by setting the respective partial derivatives to zero.

3.2.1. *Optimization with respect to α_k.* Optimization with respect to α_k in fact is a constrained optimization, since we suppose that the α_i sum up to one. So, the subproblem is: maximize (22) with respect to some α_k subject to $\sum_{i=1}^{I} \alpha_i = 1$. We do this by using the method of Lagrange multipliers (a very good tutorial explaining constrained optimization is [3]):

$$\frac{\partial}{\partial \alpha_k} \sum_{j=1}^{J} \sum_{i=1}^{I} \mathbb{E}\left[z_{ij} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] \log \left\{ \alpha_i p(y_j | C = i, \beta_i) \right\} + \lambda (\sum_{i=1}^{I} \alpha_i - 1) = 0 \tag{26}$$

$$\Leftrightarrow \sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] + \alpha_k \lambda = 0 \tag{27}$$

λ is found by getting rid of α_k, that is, by inserting the constraint. Therefore, we sum the whole equation up over i running from 1 to I:

$$\sum_{k=1}^{M}\sum_{j=1}^{J}\mathbb{E}\left[z_{ij}|\,\boldsymbol{y},\boldsymbol{\Theta}^{(i)}\right]+\lambda\sum_{k=1}^{I}\alpha_k=0 \tag{28}$$

$$\Leftrightarrow\lambda=-\sum_{k=1}^{I}\sum_{j=1}^{J}\mathbb{E}\left[z_{ij}|\,\boldsymbol{y},\boldsymbol{\Theta}^{(i)}\right] \tag{29}$$

$$\Leftrightarrow\lambda=-J\quad. \tag{30}$$

Now, (27) looks like

$$\alpha_k=\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[z_{ij}|\,\boldsymbol{y},\boldsymbol{\Theta}^{(i)}\right]\quad, \tag{31}$$

and α_k is our new estimate. As can be seen, it depends on the expectation of z_{ij}, which was already found in the expectation step!

3.2.2. *Maximization with respect to $\boldsymbol{\beta}_k$.* Maximization with respect to $\boldsymbol{\beta}_k$ depends on the components p.d.f.. In general, it is not possible to derive a closed form. For some, a closed form is sought by setting the first derivative to zero.

In this tutorial, we will take a closer look at Gaussian mixture distributions. In this case, $\boldsymbol{\beta}_k$ consists of the k-th mean value and the k-th variance:

$$\boldsymbol{\beta}_k=(\mu_k,\sigma_k)^{\mathsf{T}} \tag{32}$$

Their maxima are found by setting the derivative with respect to μ_k and σ_k equal to zero:

$$\frac{\partial}{\partial\beta_k}\sum_{j=1}^{J}\sum_{i=1}^{I}\mathbb{E}\left[z_{ij}|\,\boldsymbol{y},\boldsymbol{\Theta}^{(i)}\right]\log\left\{\alpha_i p(y_j|C=i,\beta_i)\right\}\quad. \tag{33}$$

$$=\frac{\partial}{\partial\beta_k}\sum_{j=1}^{J}\sum_{i=1}^{I}\mathbb{E}\left[z_{ij}|\,\boldsymbol{y},\boldsymbol{\Theta}^{(i)}\right]\log\alpha_i+\mathbb{E}\left[z_{ij}|\,\boldsymbol{y},\boldsymbol{\Theta}^{(i)}\right]\log p(y_j|C=i,\beta_i)\quad. \tag{34}$$

$$=\sum_{j=1}^{J}\mathbb{E}\left[z_{kj}|\,\boldsymbol{y},\boldsymbol{\Theta}^{(i)}\right]\frac{\partial}{\partial\beta_k}\log p(y_j|C=k,\boldsymbol{\beta}_k)\quad. \tag{35}$$

By inserting the definition of a Gaussian distribution we arrive at

$$\sum_{j=1}^{J}\mathbb{E}\left[z_{kj}|\,\boldsymbol{y},\boldsymbol{\Theta}^{(i)}\right]\cdot\frac{\partial}{\partial\beta_k}\log\left[\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(y_j-\mu_k)^2}{\sigma_k^2}\right)\right] \tag{36}$$

$$=\sum_{j=1}^{J}\mathbb{E}\left[z_{kj}|\,\boldsymbol{y},\boldsymbol{\Theta}^{(i)}\right]\cdot\frac{\partial}{\partial\beta_k}\left[-\log(\sigma_k)-\log\sqrt{2\pi}-\frac{1}{2}\frac{(y_j-\mu_i)^2}{\sigma_k^2}\right] \tag{37}$$

Maximization with respect to μ_k. For any maxima, the following condition has to be fulfilled:

$$\sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] \cdot \frac{\partial}{\partial \mu_k} \left[-\log(\sigma_k) - \log\sqrt{2\pi} - \frac{1}{2} \frac{(y_j - \mu_k)^2}{\sigma_k^2} \right] = 0 \quad . \tag{38}$$

The derivation for μ_k is straightforward:

$$\sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] \cdot \frac{\sigma_k^2 (\mu_k - y_j)}{\sigma_k^4} = 0 \tag{39}$$

$$\Leftrightarrow \sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] \cdot (\mu_k - y_j) = 0 \tag{40}$$

$$\Leftrightarrow \mu_k = \frac{\sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] y_j}{\sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right]} \tag{41}$$

Maximization with respect to σ_k. For any maxima, the following condition has to be fulfilled:

$$\sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] \cdot \frac{\partial}{\partial \sigma_k} \left[-\log(\sigma_k) - \log\sqrt{2\pi} - \frac{1}{2} \frac{(y_j - \mu_k)^2}{\sigma_k^2} \right] = 0 \quad . \tag{42}$$

Again, the derivation is straight forward:

$$\sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] \cdot \left[-\frac{1}{\sigma_k} - \frac{-(y_j - \mu_k)^2}{\sigma_k^3} \right] = 0 \tag{43}$$

$$\Leftrightarrow \sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] \cdot \left[-\sigma_k^2 + \frac{1}{2} (y_j - \mu_k)^2 \right] = 0 \tag{44}$$

$$\Leftrightarrow \sigma_k^2 = \frac{1}{2} \frac{\sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right] (y_j - \boldsymbol{\mu}_k)^2}{\sum_{j=1}^{J} \mathbb{E}\left[z_{kj} | \, \boldsymbol{y}, \boldsymbol{\Theta}^{(i)} \right]} \tag{45}$$

4. Conclusion

As can be seen, the EM Algorithm depends very much on the application. In fact, it has been criticized by some reviewers of the original paper by Dempster [2] that the term "Algorithm" is not adequate for this iteration scheme, since its formulation is too general. The most difficult task in the identification of the EM Algorithm for a new application is to find the hidden variable.

The reader is encouraged to look at different applications of the EM Algorithm in order to develop a better sense for where and how it can be applied.

References

1. J. Bilmes, *A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models*, 1998.

2. 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) **39** (1977), no. 1, 1–38.
3. Dan Klein, *Lagrange multipliers without permanent scarring*, University of California at Berkeley, Computer Science Division.
4. Geoffrey J. McLachlan and Thriyambakam Krishnan, *The EM algorithm and extensions*, Wiley, 1996.
5. Thomas P. Minka, *Old and new matrix algebra useful for statistics*, 2000.
6. Stuart J. Russell and Peter Norvig, *Artificial intelligence: A modern approach (2nd edition)*, Prentice Hall, 2002.

Moritz Blume, Technische Universität München, Institut für Informatik / I16, Boltzmannstrae 3, 85748 Garching b. München, Germany

E-mail address: blume@cs.tum.edu

URL: http://wwwnavab.cs.tum.edu/Main/MoritzBlume