Mixture Models and Expectation-Maximization

Justus H. Piater

Lecture at ENSIMAG, May 2002
Version 1.5 of 13 July 2013

Abstract

This tutorial attempts to provide a gentle introduction to EM by way of simple examples involving maximum-likelihood estimation of mixture-model parameters. Readers familiar with ML paramter estimation and clustering may want to skip directly to Sections 5.2 and 5.3.

Contents

1 Review: ML Parameter Estimation

Suppose we have a set of M example vectors $S = \{X_m\}$ that are drawn independently from an unknown probability distribution. We now want to fit a parametric model $p_\theta(x)$ to these data. To do this, we identify the most probable parameter vector $\hat{\theta}$ given the data S:

$$\begin{aligned}
\hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid S) \\
&= \underset{\theta}{\operatorname{argmax}}\, \frac{p(S \mid \theta)p(\theta)}{p(S)} \\
&= \underset{\theta}{\operatorname{argmax}}\, p(S \mid \theta) &&= \underset{\theta}{\operatorname{argmax}} \prod_{m=1}^{M} p(X_m \mid \theta) \\
&= \underset{\theta}{\operatorname{argmax}} \log p(S \mid \theta) &&= \underset{\theta}{\operatorname{argmax}} \sum_{m=1}^{M} \log p(X_m \mid \theta) \\
&= \underset{\theta}{\operatorname{argmax}}\, \ell(\theta)
\end{aligned} \tag{1}$$

This holds if the prior probabilities over the values of θ are uniform. This maximization can often be solved by finding roots of the gradient of the log-likelihood function. $\hat{\theta}$ is a vector that satisfies

$$\nabla_\theta \ell(\theta) = \sum_{m=1}^{M} \nabla_\theta \log p(X_m \mid \theta) = \sum_{m=1}^{M} \frac{1}{p(X_m \mid \theta)} \nabla_\theta p(X_m \mid \theta) = 0 \tag{2}$$

Consider, for example, a univariate Gaussian model:

$$\begin{aligned}
p(X_m \mid \mu, \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(X_m-\mu)^2}{2\sigma^2}} \\
\log p(X_m \mid \mu, \sigma) &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(X_m-\mu)^2}{2\sigma^2} \\
\nabla_{\mu,\sigma} \log p(X_m \mid \mu, \sigma) &= \begin{bmatrix} \frac{X_m-\mu}{\sigma^2} \\ -\frac{1}{\sigma} + \frac{(X_m-\mu)^2}{\sigma^3} \end{bmatrix}
\end{aligned}$$

Closed-form solution for $\hat{\mu}$:

$$\sum_{m=1}^{M} \frac{X_m - \hat{\mu}}{\sigma^2} = 0 \;\Rightarrow\; \sum_{m=1}^{M} X_m = M\hat{\mu} \;\Rightarrow\; \hat{\mu} = \frac{1}{M} \sum_{m=1}^{M} X_m \tag{3}$$

Closed-form solution for $\hat{\sigma}$:

$$\begin{aligned}
&\sum_{m=1}^{M} \left(-\frac{1}{\hat{\sigma}} + \frac{(X_m - \hat{\mu})^2}{\hat{\sigma}^3} \right) = 0 \\
&\Rightarrow\; \frac{M}{\hat{\sigma}} = \frac{1}{\hat{\sigma}^3} \sum_{m=1}^{M} (X_m - \hat{\mu})^2 \;\Rightarrow\; \hat{\sigma}^2 = \frac{1}{M} \sum_{m=1}^{M} (X_m - \hat{\mu})^2
\end{aligned} \tag{4}$$

2 Mixture Models

Suppose now that we have a set of M example vectors $S = \{X_m\}$ that were drawn from K independent, unknown probability distributions. Now, the probability of a data point given a model parameterization is

$$p_{\text{mix}}(X_m \mid \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} p_k(X_m \mid \theta_k) P(k) \tag{5}$$

where $P(k)$ denotes the prior probability that a data point is generated by mixture component k, with $\sum_{k=1}^{K} P(k) = 1$. Analogously to Eqn. 1, the log-likelihood function to be maximized over all the θ_k is

$$\ell(\theta_1, \ldots, \theta_K) = \sum_{m=1}^{M} \log p_{\text{mix}}(X_m \mid \theta_1, \ldots, \theta_K) \tag{6}$$

This is a multi-dimensional optimization problem with $\sum_{k=1}^{K} V_k + K - 1$ free parameters: For each of the K mixture components, a V_k-dimensional parameter vector is to be determined. The K mixture proportions $P(k)$ give rise to only $K - 1$ free parameters, since they add up to one.

If the parametric models p_k are differentiable, this maximization problem can in principle be solved by finding roots of the gradient, computed with respect to all scalar parameters of all mixture components k, and for $P(k)$, $k = 1, \ldots, K - 1$:

$$\nabla_{\theta_k} \ell(\theta_1, \ldots, \theta_K) \; = \; \sum_{m=1}^{M} \frac{P(k)}{p_{\text{mix}}(X_m \mid \theta_1, \ldots, \theta_K)} \nabla_{\theta_k} p_k(X_m \mid \theta_k) \; = \; 0 \tag{7}$$

These are the partial derivatives with respect to the $P(k)$, for $k = 1, \ldots, K - 1$:

$$\begin{aligned} & \frac{\partial}{\partial P(k)} \ell(\theta_1, \ldots, \theta_K) \\ = \; & \frac{\partial}{\partial P(k)} \sum_{m=1}^{M} \log \left(\sum_{k=1}^{K-1} p_k(X_m \mid \theta_k) P(k) \; + \; p_K(X_m \mid \theta_K) \left(1 - \sum_{k=1}^{K-1} P(k) \right) \right) \\ = \; & \sum_{m=1}^{M} \frac{p_k(X_m \mid \theta_k) - p_K(X_m \mid \theta_K)}{p_{\text{mix}}(X_m \mid \theta_1, \ldots, \theta_K)} \end{aligned} \tag{8}$$

Equations 7 and 8 define a system of $\sum_{k=1}^{K} V_k + K - 1$ simultaneous equations. Due to the presence of the mixture probability (5), this system is non-linear in all practical cases, and closed-form solutions usually do not exist. Therefore, one needs to resort to numerical optimization problems, using appropriate constraints on the θ_k and the $P(k)$.

3 The K-Means Problem and an EM solution

Often, an elegant way to estimate the parameters of a mixture model is Expectation-Maximization (EM) [1]. To illustrate this, we will begin with a simplified version of the above problem, known as K-Means.

Suppose we are given M data points S that we want to fit using a mixture of K univariate Gaussian distributions with identical and known variance σ^2, and non-informative component priors $P(k)$. If we knew which distribution generated which data point, this problem would be easy to solve. For this purpose, let us represent the data points X_m as $(K+1)$-tuples $\langle Y_m, w_{m1}, \ldots, w_{mK} \rangle$, where $w_{mk} = 1$ if Y_m was generated by component distribution k, otherwise 0. Then, from Eqn. 3, the maximum-likelihood solution is simply given by

$$\mu_k = \frac{1}{M_k} \sum_{m=1}^{M} w_{mk} Y_m \tag{9}$$

where $M_k = \sum_{m=1}^{M} w_{mk}$, and $k = 1, \ldots, K$.

However, the values of the w_{mk} are not known. On the other hand, if we knew the K means μ_k, we could easily compute maximum-likelihood estimates of the w_{mk}, i.e., those that maximize $p(S \mid \mu_k, w_{mk})$, the likelihood of the data, for all k and all m:

$$w_{mk} = \operatorname*{argmax}_k \, p(Y_m \mid \mu_k) P(k) \tag{10}$$

Unfortunately, we have neither the w_{mk} nor the μ_k.

The idea of the EM algorithm is to estimate both simultaneously by iterating between the above two calculations. We start by initializing our μ_k to arbitrary initial values, and then iterate the following two steps:

Expectation (E) Calculate the expected value of the w_{mk} based on the current estimates of the μ_k.

Maximization (M) Calculate the new maximum-likelihood estimate for the μ_k based on the current expected values of the w_{mk}.

At the **E** step, the expected value of w_{mk} is simply the probability that Y_m was generated by component k, which we compute using Bayes' Rule:

$$E[w_{mk}] = p(k \mid Y_m) = \frac{p(Y_m \mid k) P(k)}{p(Y_m)} = \frac{p(Y_m \mid \mu_k) P(k)}{\sum_{j=1}^{K} p(Y_m \mid \mu_j) P(j)} = \frac{\mathrm{e}^{-\frac{(Y_m - \mu_k)^2}{2\sigma^2}}}{\displaystyle\sum_{j=1}^{K} \mathrm{e}^{-\frac{(Y_m - \mu_j)^2}{2\sigma^2}}} \tag{11}$$

The $P(k)$ cancel out with the $P(j)$ since, as stated above, we are assuming equal component priors.

At the **M** step, we need to find the parameters μ_k that maximize the likelihood function

$$
\begin{aligned}
& p(S \mid \mu_k, w_{mk} \text{ for } k = 1, \ldots, K \text{ and } m = 1, \ldots, M) \\
= & \prod_{m=1}^{M} \sum_{k=1}^{K} \frac{w_{mk}}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(Y_m - \mu_k)^2}{2\sigma^2}} \qquad (12) \\
= & \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{k=1}^{K} w_{mk}(Y_m - \mu_k)^2} \qquad (13)
\end{aligned}
$$

where the second equality holds because in reality, each data point has been generated by exactly one random process, i.e., all w_{mk} are either zero or one.[1]

Equivalently, we can minimize the negative log-likelihood, here simplified by dropping irrelevant terms:

$$\ell(\mu_k, w_{mk} \text{ for } k = 1, \ldots, K \text{ and } m = 1, \ldots, M) = \sum_{m=1}^{M} \sum_{k=1}^{K} w_{mk}(Y_m - \mu_k)^2 \qquad (15)$$

Since $\ell(\cdot)$ is a random variable governed by the distribution that generates S, or, equivalently, by the distribution governing the unobserved variables w_{mk}, we must consider its expected value $E[\ell(\cdot)]$. Since $\ell(\cdot)$ is linear in the w_{mk}, we have

$$E[\ell(\cdot)] = E\left[\sum_{m=1}^{M} \sum_{k=1}^{K} w_{mk}(Y_m - \mu_k)^2\right] = \sum_{m=1}^{M} \sum_{k=1}^{K} E[w_{mk}](Y_m - \mu_k)^2 \qquad (16)$$

For a closed-form solution, we set the derivatives with respect to the μ_k to zero:

$$
\begin{aligned}
\frac{\partial}{\partial \mu_k} E[\ell(\cdot)] &= -2 \sum_{m=1}^{M} E[w_{mk}](Y_m - \mu_k) \\
0 &= \sum_{m=1}^{M} E[w_{mk}](Y_m - \mu_k) \\
\mu_k &= \frac{\sum_{m=1}^{M} E[w_{mk}] Y_m}{\sum_{m=1}^{M} E[w_{mk}]} \qquad (17)
\end{aligned}
$$

Thus, Equations 11 and 17 define the EM algorithm for the K-Means problem.

[1] A different way to use this fact is to replace Eqn. 12 by

$$p(S \mid \cdot) = \prod_{m=1}^{M} \prod_{k=1}^{K} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(Y_m - \mu_k)^2}{2\sigma^2}} \right)^{w_{mk}}, \qquad (14)$$

from which Eqn. 13 follows directly.

4 The EM Algorithm

The EM algorithm is a general method for solving the following class of problems:

Given: A set $Y = \{Y_m\}$, $m = 1, \ldots, M$, of observation vectors.

Assumption: The Y are the observable part of data points $X = \{X_m\}$ from a higher-dimensional space. In other words, $Y = Y(X)$ via a many-to-one mapping. The complete data X follow a parametric probability density function $p(X \mid \theta)$ (or, for discrete X, a probability mass function $P(X \mid \theta)$).

Wanted: An explanation of the observed data Y in terms of a parametric description of the full data X. Formally, we seek a maximum-likelihood estimate of the paramter vector θ:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \log p_Y(Y \mid \theta) \tag{18}$$

The incomplete-data specification p_Y is related to the complete-data specification p – for which we have a parametric model – by

$$p_Y(Y \mid \theta) \;=\; \int_{X(Y)} p(X \mid \theta)\,\mathrm{d}X \tag{19}$$

where $X(Y)$ denotes all values of X for which $Y(X) = Y$. Since we do not have the full data X to compute the solution (19) directly, we maximize instead its expectation $E[\log p(X \mid \theta)]$. This expectation is taken over the probability distribution governing X, which is determined by the known values Y and the probability density function describing the unobserved portion of X.

Unfortunately, we do not have the parameter vector θ that defines the probability distribution governing X (this vector is exactly what we set out to find in the first place). Therefore, we use an estimate of it, that we iteratively improve. Let us define a function Q that expresses the sought expectation of the likelihood as a function of the parameters θ that we are trying to estimate, given the observed data Y and a current estimate $\hat{\theta}$ of the parameters:

$$Q(\theta \mid \hat{\theta}) = E[\log p(X \mid \theta) \mid Y, \hat{\theta}] \tag{20}$$

This Q function will allow us to compute the expected log-likelihood of the complete data X for any parameterization θ, while the expectations are computed using a fixed probability distribution defined by the observed data Y and a given parameterization $\hat{\theta}$.

The general EM algorithm specifies an iterative procedure for improving the estimate $\hat{\theta}$:

1. Choose an initialization for $\hat{\theta}$.
2. **(E)** Construct a computable representation for Eqn. 20, using the current $\hat{\theta}$.
3. **(M)** Find a new parameterization $\hat{\theta}$ that maximizes the current Q function:

$$\hat{\theta} \leftarrow \underset{\theta}{\operatorname{argmax}}\, Q(\theta \mid \hat{\theta}) \tag{21}$$

4. If $\hat{\theta}$ has barely changed, stop. Otherwise, continue at Step 2.

This algorithm will improve the estimate $\hat{\theta}$, increasing the value of Q at every M step until it reaches a local maximum.

In practice, the E step involves the computation of some parameters defining Q. Although the EM algorithm is conceptually simple, both E and M steps may be quite difficult to compute. However, in many practical cases there exist closed-form solutions for both E and M steps.

5 Examples

5.1 The K-Means Problem Revisited

In the case of the K-Means problem, we have $X_m = Y_m \cup Z_m$, where the $Z_m = \{w_{mk}\}$ are the hidden variables, and $\theta = [\mu_1, \ldots, \mu_K]$. The Q function (20) is (cf. Eqn. 13)

$$\begin{aligned} Q(\theta \mid \hat{\theta}) &= E\left[\log \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{k=1}^{K} w_{mk}(Y_m-\mu_k)^2} \middle| Y, \hat{\theta}\right] \\ &= \sum_{m=1}^{M}\left(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{k=1}^{K} E[w_{mk} \mid Y_m, \hat{\theta}](Y_m - \mu_k)^2\right) \end{aligned} \tag{22}$$

Thus, at the E step, specifying the Q function amounts to computing the expected values of the unknown variables $Z_m = \{w_{mk}\}$ as shown in Eqn. 11, using the current parameter estimates $\hat{\theta} = \{\hat{\mu}_k\}$.

At the M step, the Q function is maximized as shown in Eqns. 16–17 after dropping constant terms.

5.2 Mixture Models

A typical application of EM is the estimation of the parameters of a mixture model

$$p_{\mathrm{mix}}(Y_m \mid \Theta) = \sum_{k=1}^{K} p(Y_m \mid \theta_k)\, P(k) \tag{23}$$

to fit an observed set of data points $\{Y_m\}$. The mixing proportions $P(k)$ and the components k_m that generated each data point Y_m are unknown. The objective is to find the parameter vector θ_k describing each component density $p(Y \mid \theta_k)$.

For distributions of the exponential family whose logarithms are linear in the w_{mk}, the Expectation step essentially computes, as shown in Eqn. 22 above, the expected values of the indicators w_{mk} that each data point Y_m was generated by component k, given the current parameter estimates θ_k and $P(k)$, using Bayes' Rule:

$$E[w_{mk}] = \frac{p(Y_m \mid \theta_k)\, P(k)}{\sum_{j=1}^{K} p(Y_m \mid \theta_j)\, P(j)} = \frac{p(Y_m \mid \theta_k)\, P(k)}{p_{\mathrm{mix}}(Y_m \mid \Theta)} \tag{24}$$

At the Maximization step, a new set of parameters θ_k, $k = 1, \ldots, K$, is computed to maximize the log-likelihood of the observed data:

$$\ell(\Theta) = \sum_{m=1}^{M} \log p_{\mathrm{mix}}(Y_m \mid \Theta) \tag{25}$$

At the maximum, the partial derivatives with respect to all parameters vanish:

$$\begin{aligned} 0 \;=\; \nabla_{\theta_k} \ell(\Theta) \;&=\; \sum_{m=1}^{M} \frac{P(k)}{p_{\mathrm{mix}}(Y_m \mid \Theta)}\, \nabla_{\theta_k} p(Y_m \mid \theta_k) \\ &=\; \sum_{m=1}^{M} \frac{w_{mk}}{p(Y_m \mid \theta_k)}\, \nabla_{\theta_k} p(Y_m \mid \theta_k) \qquad (26) \end{aligned}$$

where the second line (26) follows from substituting Eqn. 24. The Maximization is then computed by solving this system (26) for all θ_k. Moreover, the estimates of the component priors are updated by averaging the data-conditional component probabilities computed at the Expectation step:

$$P(k) = \frac{1}{M} \sum_{m=1}^{M} w_{mk} \tag{27}$$

5.3 Gaussian Mixture Models

The one-dimensional K-Means problem arguably constitutes the simplest special case of Gaussian mixture fitting. We will now derive an EM algorithm for the similar problem of a one-dimensional Gaussian mixture, where we do not know the variances σ_k^2 or the mixture proportions $P(k)$ either. The parameter vector for mixture component k is thus $\theta_k = [\mu_k, \sigma_k]^T$:

$$p_k(y \mid \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \mathrm{e}^{\frac{(y-\mu_k)^2}{2\sigma_k^2}} \tag{28}$$

The Expectation step is easily defined by plugging Eqn. 28 into Eqn. 24.

For the Maximization, we plug Eqn. 28 into Eqn. 26:

$$\begin{aligned}
0 = \frac{\partial}{\partial \mu_k} \ell(\Theta) &= \sum_{m=1}^{M} \frac{w_{mk}}{p_k(y_m \mid \mu_k, \sigma_k)} \frac{-2(y_m - \mu_k)}{2\sigma_k^2} p_k(y_m \mid \mu_k, \sigma_k) \\
&= \sum_{m=1}^{M} w_{mk}(y_m - \mu_k) \\
\mu_k &= \frac{\sum_{m=1}^{M} w_{mk} y_m}{\sum_{m=1}^{M} w_{mk}} \\
0 = \frac{\partial}{\partial \sigma_k} \ell(\Theta) &= \sum_{m=1}^{M} \frac{w_{mk}}{p_k(y_m \mid \mu_k, \sigma_k)} \left(\frac{-1}{\sigma_k} + \frac{-2(y_m - \mu_k)^2}{2\sigma_k^3} \right) p_k(y_m \mid \mu_k, \sigma_k) \\
&= \sum_{m=1}^{M} w_{mk} \left(\sigma_k^2 + (y_m - \mu_k)^2 \right) \\
\sigma_k^2 &= \frac{\sum_{m=1}^{M} w_{mk}(y_m - \mu_k)^2}{\sum_{m=1}^{M} w_{mk}}
\end{aligned}$$

Finally, we recompute the mixture proportions using Eqn. 27.

5.4 ET Image Reconstruction

In emission tomography (ET), body tissues are stimulated to emit photons, that are detected by D detectors surrounding the tissue. The body is modeled as a block of B equally-sized boxes. Given the number $y(d)$ of photons detected by each detector d, we want to know the number $n(b)$ of photons emitted at each box b. The emission of photons from box b is modeled as a Poisson process with mean $\lambda(b)$:

$$p(n(b) \mid \lambda(b)) = \mathrm{e}^{-\lambda(b)} \frac{\lambda(b)^n}{n!} \tag{29}$$

The $\lambda = \{\lambda(b),\ b = 1, \ldots, B\}$ are thus the unknown parameters we need to estimate, using the measurements $\mathbf{y} = \{y(d),\ d = 1, \ldots, D\}$.

A photon emitted from box b is detected by detector d with probability $p(b, d)$, and we assume that all photons are detected by exactly one detector:

$$\sum_{d=1}^{D} p(b, d) = 1 \tag{30}$$

The $p(b, d)$ are known, as they can be determined from the geometry of the detectors. The number $y(d)$ of photons detected by detector d is Poisson distributed

$$p(y \mid \lambda(d)) = \mathrm{e}^{-\lambda(d)} \frac{\lambda(d)^y}{y!} \tag{31}$$

and it is intuitive and provable that

$$\lambda(d) = E[y(d)] = \sum_{b=1}^{B} \lambda(b) p(b, d). \tag{32}$$

Let $x(b, d)$ be the number of photons emitted from box b detected by detector d. Thus, $\mathbf{x} = \{x(b, d),\ b = 1, \ldots, B,\ d = 1, \ldots, D\}$ constitute the complete data. Each $x(b, d)$ is Poisson distributed with mean

$$\lambda(b, d) = \lambda(b) p(b, d). \tag{33}$$

Assuming independce between all boxes and between all detectors, the likelihood function of the complete data is

$$p(x \mid \lambda) = \prod_{\substack{b=1,\ldots,B \\ d=1,\ldots,D}} \mathrm{e}^{-\lambda(b,d)} \frac{\lambda(b, d)^{x(b,d)}}{x(b, d)!} \tag{34}$$

and, using Eqn. 33, the log-likelihood is

$$\begin{aligned} \log p(\mathbf{x} \mid \lambda) \ &= \\ \sum_{\substack{b=1,\ldots,B \\ d=1,\ldots,D}} \Big(&- \lambda(b) p(b, d) + x(b, d) \log \lambda(b) + x(b, d) \log p(b, d) - \log x(b, d)! \Big) \end{aligned} \tag{35}$$

For the E step, we set up the function

$$Q(\lambda \mid \hat{\lambda}) = E[\log p(\mathbf{x} \mid \lambda) \mid \mathbf{y}, \hat{\lambda}]. \tag{36}$$

Since the Poisson distribution belongs to the exponential family, this once more boils down to estimating

$$E[x(b,d) \mid \mathbf{y}, \hat{\lambda}] \;=\; E[x(b,d) \mid y(d), \hat{\lambda}] \tag{37}$$

where the simplifying equality comes from the fact that all boxes are independent.

At the M step, we maximize Eqn. 35 by setting $\nabla_{\lambda(b)} \log p(\mathbf{x} \mid \lambda) = 0$. The remaining details are omitted here.

6 Bibliographical Remarks

The K-Means problem and its EM solution are borrowed from Mitchell's excellent textbook [2]. The ET image reconstruction example is from Moon [3], where the full solution is given. He also explains the general EM procedure quite clearly, and gives other examples as well.

References

[1] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)*, 39(1):1–38, 1977.

[2] T. Mitchell. *Machine Learning*. WCB/McGraw-Hill, 1997.

[3] T. Moon. The Expectation-Maximization Algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, 11 1996.