

EM algorithm example: A mixture of Gaussians with unknown means

Guy Lebanon

Spring 2001

Description of the EM algorithm

The EM algorithm maximize $\log p(D|\theta)$ (we will sometimes use D to represent the data y_1, \dots, y_n):

Step 0: Initialize

Start with $t = 0$, some initial guess $\theta^{(0)}$ and iterate over the following steps

E step

Compute the bound

$$Q(\theta'; \theta^{(t)}) = E\{\log p(D, h|\theta')|\theta^{(t)}, D\}$$

M step

Maximize the bound

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta'} Q(\theta'; \theta^{(t)})$$

Update $t = t + 1$ and go to E step.

Example - mixture of Gaussians, equal priors

We are given data $D = (y_1, \dots, y_n)$ in which each point is generated from one of two Gaussians (with equal prior) with different unknown means and variances one: $N(y; \mu_1, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_1)^2}{2}}$ or $N(y; \mu_2, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_2)^2}{2}}$. We introduce hidden variables $H = \{h_{i,j}, i = 1, \dots, n, j = 1, 2\}$. $h_{i,1}$ will get value 1 if y_i was generated from the first Gaussian and 0 otherwise. $h_{i,2}$ will get value 1 if y_i was generated from the second Gaussian and 0 otherwise.

The Q function in this case is

$$\begin{aligned}
Q(\boldsymbol{\mu}'; \boldsymbol{\mu}^{(t)}) &= E\{\log p(D, h|\boldsymbol{\mu}')|\boldsymbol{\mu}^{(t)}, D\} = E\left\{\sum_{i=1}^n \log p(y_i, h|\boldsymbol{\mu}')|\boldsymbol{\mu}^{(t)}, D\right\} \\
&= E\left\{\sum_{i=1}^n \log \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}h_{i,1}(y_i - \mu'_1)^2 - \frac{1}{2}h_{i,2}(y_i - \mu'_2)^2} |\boldsymbol{\mu}^{(t)}, D\right\} \\
&= E\left\{-\sum_{i=1}^n \frac{1}{2}h_{i,1}(y_i - \mu'_1)^2 - \sum_{i=1}^n \frac{1}{2}h_{i,2}(y_i - \mu'_2)^2 - \log 2\sqrt{2\pi} |\boldsymbol{\mu}^{(t)}, D\right\} \\
&= -c - \frac{1}{2} \sum_{i=1}^n E\{h_{i,1}|\boldsymbol{\mu}^{(t)}, D\}(y_i - \mu'_1)^2 - \frac{1}{2} \sum_{i=1}^n E\{h_{i,2}|\boldsymbol{\mu}^{(t)}, D\}(y_i - \mu'_2)^2
\end{aligned}$$

Computing $E\{h_{i,j}|\boldsymbol{\mu}^{(t)}, D\}$ results in

$$E\{h_{i,1}|\boldsymbol{\mu}^{(t)}, D\} = 1 \cdot Pr(h_{i,1} = 1|\boldsymbol{\mu}^{(t)}, D) + 0 \cdot Pr(h_{i,1} = 0|\boldsymbol{\mu}^{(t)}, D) = \frac{e^{-\frac{1}{2}(y_i - \mu_1^{(t)})^2}}{e^{-\frac{1}{2}(y_i - \mu_1^{(t)})^2} + e^{-\frac{1}{2}(y_i - \mu_2^{(t)})^2}} = F_{i,1}$$

$$E\{h_{i,2}|\boldsymbol{\mu}^{(t)}, D\} = 1 \cdot Pr(h_{i,2} = 1|\boldsymbol{\mu}^{(t)}, D) + 0 \cdot Pr(h_{i,2} = 0|\boldsymbol{\mu}^{(t)}, D) = \frac{e^{-\frac{1}{2}(y_i - \mu_2^{(t)})^2}}{e^{-\frac{1}{2}(y_i - \mu_1^{(t)})^2} + e^{-\frac{1}{2}(y_i - \mu_2^{(t)})^2}} = F_{i,2}$$

The M step consists of plugging the expressions for $E\{h_{i,j}|\boldsymbol{\mu}^{(t)}, D\}$ in the Q function and maximizing with respect to μ'_1, μ'_2 :

$$\operatorname{argmax}_{\mu'_1, \mu'_2} - \frac{1}{2} \sum_{i=1}^n \frac{e^{-\frac{1}{2}(y_i - \mu_1^{(t)})^2}}{e^{-\frac{1}{2}(y_i - \mu_1^{(t)})^2} + e^{-\frac{1}{2}(y_i - \mu_2^{(t)})^2}} (y_i - \mu'_1)^2 - \frac{1}{2} \sum_{i=1}^n \frac{e^{-\frac{1}{2}(y_i - \mu_2^{(t)})^2}}{e^{-\frac{1}{2}(y_i - \mu_1^{(t)})^2} + e^{-\frac{1}{2}(y_i - \mu_2^{(t)})^2}} (y_i - \mu'_2)^2$$

Maximizing with respect to μ'_1 and μ'_2 by computing derivatives and comparing to 0 yields

$$\mu_1^{(t+1)} = \mu'_1 = \frac{\sum_{i=1}^n F_{i,1} y_i}{\sum_i F_{i,1}}, \quad \mu_2^{(t+1)} = \mu'_2 = \frac{\sum_{i=1}^n F_{i,2} y_i}{\sum_i F_{i,2}}$$

The above update rule is simply a weighted average of the example points (a weighted centroid) where each weight is the relative confidence of the corresponding Gaussian generating this example (calculated from previous iteration means).

Example - mixture of Gaussians, nonequal priors

We now return to the previous example, when the priors of the Gaussian are not equal (the priors will be denoted by λ_1, λ_2). In this case the parameters are $\boldsymbol{\theta} = \mu_1, \mu_2, \lambda_1, \lambda_2$.

The Q function in this case is

$$\begin{aligned}
Q(\boldsymbol{\theta}'; \boldsymbol{\theta}^{(t)}) &= E\{\log p(D, h|\boldsymbol{\theta}')|\boldsymbol{\theta}^{(t)}, D\} = E\left\{\sum_{i=1}^n \log p(y_i, h|\boldsymbol{\theta}')|\boldsymbol{\theta}^{(t)}, D\right\} \\
&= E\left\{\sum_{i=1}^n \log(\lambda'_1)^{h_{i,1}} (\lambda'_2)^{h_{i,2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}h_{i,1}(y_i - \mu'_1)^2 - \frac{1}{2}h_{i,2}(y_i - \mu'_2)^2} |\boldsymbol{\theta}^{(t)}, D\right\} \\
&= E\left\{c + \sum_{i=1}^n \left(h_{i,1} \log \lambda'_1 + h_{i,2} \log \lambda'_2 - \frac{1}{2}h_{i,1}(y_i - \mu'_1)^2 - \frac{1}{2}h_{i,2}(y_i - \mu'_2)^2\right) |\boldsymbol{\theta}^{(t)}, D\right\} \\
&= c + \frac{1}{2} \sum_{i=1}^n E\{h_{i,1}|\boldsymbol{\theta}^{(t)}, D\} (2 \log \lambda'_1 - (y_i - \mu'_1)^2) + \frac{1}{2} \sum_{i=1}^n E\{h_{i,2}|\boldsymbol{\theta}^{(t)}, D\} (2 \log \lambda'_2 - (y_i - \mu'_2)^2)
\end{aligned}$$

Computing $E\{h_{i,j}|\boldsymbol{\theta}^{(t)}, D\}$ results in

$$E\{h_{i,1}|\boldsymbol{\theta}^{(t)}, D\} = 1 \cdot Pr(h_{i,1} = 1|\boldsymbol{\theta}^{(t)}, D) + 0 \cdot Pr(h_{i,1} = 0|\boldsymbol{\theta}^{(t)}, D) = \frac{\lambda_1^{(t)} e^{-\frac{1}{2}(y_i - \mu_1^{(t)})^2}}{\lambda_1^{(t)} e^{-\frac{1}{2}(y_i - \mu_1^{(t)})^2} + \lambda_2^{(t)} e^{-\frac{1}{2}(y_i - \mu_2^{(t)})^2}} = F_{i,1}$$

$$E\{h_{i,2}|\boldsymbol{\theta}^{(t)}, D\} = 1 \cdot Pr(h_{i,2} = 1|\boldsymbol{\theta}^{(t)}, D) + 0 \cdot Pr(h_{i,2} = 0|\boldsymbol{\theta}^{(t)}, D) = \frac{\lambda_2^{(t)} e^{-\frac{1}{2}(y_i - \mu_2^{(t)})^2}}{\lambda_1^{(t)} e^{-\frac{1}{2}(y_i - \mu_1^{(t)})^2} + \lambda_2^{(t)} e^{-\frac{1}{2}(y_i - \mu_2^{(t)})^2}} = F_{i,2}$$

The M step consists of plugging the expressions for $E\{h_{i,j}|\boldsymbol{\mu}^{(t)}, D\}$ in the Q function and maximizing with regards to μ'_1, μ'_2 :

$$\operatorname{argmax}_{\mu'_1, \mu'_2, \lambda'_1, \lambda'_2} \frac{1}{2} \sum_{i=1}^n F_{i,1} (2 \log \lambda'_1 - (y_i - \mu'_1)^2) + F_{i,1} (2 \log \lambda'_2 - (y_i - \mu'_2)^2)$$

Maximizing with respect to μ'_1 and μ'_2 by computing derivatives and comparing to 0 yields as before;

$$\mu_1^{(t+1)} = \mu'_1 = \frac{\sum_{i=1}^n F_{i,1} y_i}{\sum_i F_{i,1}}, \quad \mu_2^{(t+1)} = \mu'_2 = \frac{\sum_{i=1}^n F_{i,2} y_i}{\sum_i F_{i,2}}$$

Before maximizing with respect to λ'_1 and λ'_2 note that we need to enforce the constraint $\lambda'_1 + \lambda'_2 = 1$ and so we will substitute $1 - \lambda'_1$ for λ'_2 :

$$\operatorname{argmax}_{\lambda'_1} \sum_{i=1}^n F_{i,1} \log \lambda'_1 + F_{i,2} \log(1 - \lambda'_1).$$

Setting the derivative to 0 yields

$$\frac{\sum_{i=1}^n F_{i,1}}{\lambda'_1} - \frac{\sum_{i=1}^n F_{i,2}}{1 - \lambda'_2} = 0$$

and after some elementary manipulations, the update rule turns to be

$$\lambda_1^{(t+1)} = \lambda'_1 = \frac{\sum_{i=1}^n F_{i,1}}{\sum_{i=1}^n F_{i,1} + F_{i,2}} = \frac{\sum_{i=1}^n F_{i,1}}{n}$$

and $\lambda_2^{(t+1)} = 1 - \lambda_1^{(t+1)}$.

If the probability function is of the form $f(y_i|\theta) = c(y_i)e^{\sum_i h_i g_i(y_i)}$ where h_i are the hidden variables the log-likelihood will be linear in h_i . This will make the E step simply substitute h_i by $E\{h_i|\theta^{(t)}\}$ in the log-likelihood.

Continued on the next page

EM shortcut for exponential distributions

Guy Lebanon

If the probability function is of the form $f(y_i|\theta) = c(y_i, \theta)e^{\sum_i h_i g_i(y_i, \theta)}$ where h_i are the hidden variables the log-likelihood will be linear in h_i . This will make the E step simply substitute h_i by $E\{h_i|\theta^{(t)}\}$ in the log-likelihood. As a consequence, for this situation the EM algorithm boils down to:

- Write down the maximum likelihood estimator (using the log-likelihood of the data).
- Substitute h_i by $E\{h_i|\theta^{(t)}\}$ in the MLE to get $\theta^{(t+1)}$ and re-iterate until convergence.

As an example, consider the case of a mixture of Gaussians with unknown means and priors: We are given data $D = (y_1, \dots, y_n)$ in which each point is generated from k Gaussians with different unknown means and variances one: $N(y; \mu_j, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_j)^2}{2}}$.

We introduce hidden variables $H = \{h_{i,j}, i = 1, \dots, n, j = 1, \dots, k\}$. $h_{i,j}$ will get value 1 if y_i was generated from the j th Gaussian and 0 otherwise. The priors of the Gaussian will be denoted by λ_j , making the parameters $\theta = \{\mu_1, \dots, \mu_k, \lambda_1, \dots, \lambda_k\}$.

The MLE for λ_j is $\frac{1}{n} \sum_i h_{i,j}$ making the EM iteration

$$\lambda_j^{(t+1)} = \frac{1}{n} \sum_i E\{h_{i,j}|\theta^{(t)}\} = \frac{1}{n} \sum_i F_{i,j}.$$

where

$$F_{i,j} = \frac{\lambda_j^{(t)} e^{-\frac{1}{2}(y_i - \mu_j^{(t)})^2}}{\sum_j \lambda_j^{(t)} e^{-\frac{1}{2}(y_i - \mu_j^{(t)})^2}}.$$

The MLE for μ_j is $\frac{\sum_i h_{i,j} y_i}{\sum_i h_{i,j}}$ and so the EM iteration becomes

$$\mu_j^{(t+1)} = \frac{\sum_i E\{h_{i,j}|\theta^{(t)}\} y_i}{\sum_i E\{h_{i,j}\}} = \frac{\sum_i F_{i,j} y_i}{\sum_i F_{i,j}}.$$