

# The EM Algorithm

Ritwik Sahani/ Aayush Goyal

IIT Hyderabad

March 4, 2019

Suppose we have some data points  $x_1, \dots, x_n$



and we also tell you that these points belong to two different probabilistic distributions, say, two gaussians such that we know which point belongs to which particular distribution.

**Can we find the parameters  $\mu_b, \sigma^2$  for the two gaussians?**

**Indeed, we can.**

$$\mu = \sum_{i=1}^{n_b} \frac{x_i}{n_b}$$

$$\sigma^2 = \sum_{i=1}^{n_b} \frac{(x_i - \mu_b)^2}{n_b}$$

and similarly for the other distribution.

Now, suppose data points are there but we don't know the as to which gaussian a particular point belongs.



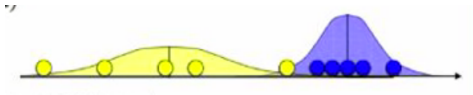
But what we do know is that these come from two gaussians, whose parameters  $\mu, \sigma^2$  is known

**Can we, for each point, decide which of the two gaussians it is more likely to belong to?**

Yes, this can be done by calculating a posteriori prob.

$$P(b|x_i) = \frac{P(x_i|b)P(b)}{P(x_i|b)P(b) + P(x_i|a)P(a)}$$

where  $P(x_i|b)$  can be obtained from the gaussian  $N(\mu_b, \sigma_b^2)$



**But what happens, if we neither knew the source gaussian, nor their parameters?**

Then we use the EM algorithm method, an iterative method, to find the same.

## Step-1: Expectation

Creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters.

★ **Basically we will assume some initial parameters and assign each point with its a- posteriori probability**

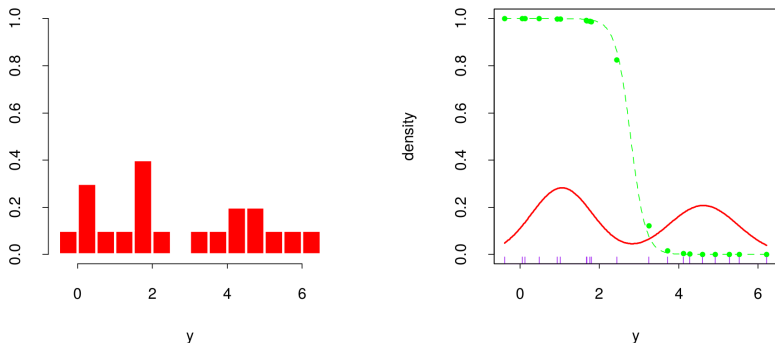
## Step-2: Maximization

Which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

★ **Basically, we will again calculate the parameters for the next E -step**

## Two-Component Mixture Model

We consider a simple mixture model for density estimation, and the corresponding EM algorithm for carrying out maximum likelihood estimation.



**Figure:** Mixture example. (Left panel:) Histogram of data. (Right panel:) Maximum likelihood fit of Gaussian densities (solid red) and responsibility (dotted green) of the left component density for observation  $y$ , as a function of  $y$ .

**TABLE 8.1.** Twenty fictitious data points used in the two-component mixture example in Figure 8.5.

-0.39	0.12	0.94	1.67	1.76	2.44	3.72	4.28	4.92	5.53
0.06	0.48	1.01	1.68	1.80	3.25	4.12	4.60	5.28	6.22

Due to Bimodality of data Gaussian distribution would be inappropriate. So we will model  $Y$  as a mixture of two distinct distributions.

$$Y_1 \equiv N(\mu_1, \sigma_1^2)$$

$$Y_2 \equiv N(\mu_2, \sigma_2^2)$$

$$Y = (1 - \Delta) * Y_1 + \Delta * Y_2$$

where  $\Delta \in \{0, 1\}$  with  $Pr(\Delta = 1) = \pi$

This generative representation is explicit: generate a  $\Delta \in \{0, 1\}$  with probability  $\pi$ , and then depending on the outcome, deliver either  $Y_1$  or  $Y_2$ . Let  $\phi_\theta(x)$  denote the normal density with parameters  $\theta = (\mu, \sigma^2)$ . Then the density of  $Y$  is

$$g_Y(y) = (1 - \pi)\phi_{\theta_1}(y) + \pi\phi_{\theta_2}(y).$$

Now suppose we wish to fit this model to the data in Figure 8.5 by maximum likelihood. The parameters are

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

The log-likelihood based on the  $N$  training cases is

$$l(\theta; Z) = \sum_{i=1}^N \log[(1 - \pi)\phi_{\theta_1}(y_i) + \pi\phi_{\theta_2}(y_i)]$$

Direct maximization of  $l(\theta; Z)$  is quite difficult numerically, because of the sum of terms inside the logarithm. There is, however, a simpler approach. We consider unobserved latent variables  $\theta_i$  taking values 0 or 1 as in (8.36): if  $\Delta_i = 1$  then  $Y_1$  comes from model 2, otherwise it comes from model 1.

$$l_0(\theta; Z, \Delta) = \sum_{i=1}^N [(1 - \Delta_i) \log \phi_{\theta_1}(y_i) + \Delta_i \log \phi_{\theta_2}(y_i)] + \sum_{i=1}^N [(1 - \Delta_i) \log(1 - \pi) + \Delta_i \log \pi]$$



## Now we apply the E-M Algorithm for Two-component Gaussian mixture

1. Take initial guesses for the parameters  $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$
2. *Expectation Step*: compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi} \phi_{\hat{\theta}_2}(y_i)}{(1 - \hat{\pi}) \phi_{\hat{\theta}_1}(y_i) + \hat{\pi} \phi_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \dots, N.$$

3. *Maximization Step*: compute the weighted means and variances:

$$\begin{aligned} \hat{\mu}_1 &= \frac{\sum_{i=1}^N (1 - \hat{\gamma}_i) y_i}{\sum_{i=1}^N (1 - \hat{\gamma}_i)}, & \hat{\sigma}_1^2 &= \frac{\sum_{i=1}^N (1 - \hat{\gamma}_i) (y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1 - \hat{\gamma}_i)}, \\ \hat{\mu}_2 &= \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, & \hat{\sigma}_2^2 &= \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i}, \end{aligned}$$

and the mixing probability  $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$ .

4. Iterate steps 2 and 3 until convergence.

Note that the actual maximization of the likelihood occurs when we put a spike of infinite height at any one data point, that is,  $\hat{\mu}_1 = y_i$  for some  $i$  and  $\hat{\sigma}_1^2 = 0$ . This gives infinite likelihood, but is not a useful solution. Hence we are actually looking for a good local maximum of the likelihood, one for which  $\hat{\sigma}_1^2, \hat{\sigma}_2^2 > 0$ . To further complicate matters, there can be more than one local maximum having  $\hat{\sigma}_1^2, \hat{\sigma}_2^2 > 0$ . In our example, we ran the EM algorithm with a number of different initial guesses for the parameters, all having  $\hat{\sigma}_k^2 > 0.5$ , and chose the run that gave us the highest maximized likelihood

Iteration	$\hat{\pi}$
1	0.485
5	0.493
10	0.523
15	0.544
20	0.546

Figure: Selected iterations for E-M algorithm

The final maximum likelihood estimates are  $\hat{\mu}_1 = 4.62$   $\hat{\sigma}_1^2 = 0.87$ ,  
 $\hat{\mu}_2 = 1.06$  ,  $\hat{\sigma}_2^2 = 0.77$ ,  $\hat{\pi} = 0.546$ .

## 8. Model Inference and Averaging

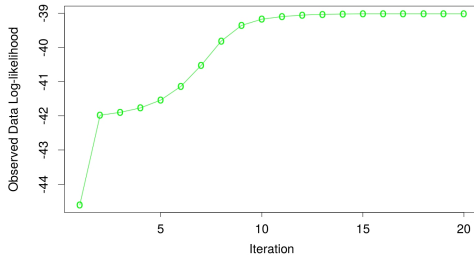


Figure: observed data log-likelihood as a function of the iteration number

## EM Algorithm in general

1. Start with initial guesses for the parameters  $\hat{\theta}^{(0)}$ .
2. *Expectation Step*: at the  $j$ th step, compute

$$Q(\theta', \hat{\theta}^{(j)}) = \mathbb{E}(\ell_0(\theta'; \mathbf{T}) | \mathbf{Z}, \hat{\theta}^{(j)})$$

as a function of the dummy argument  $\theta'$ .

3. *Maximization Step*: determine the new estimate  $\hat{\theta}^{(j+1)}$  as the maximizer of  $Q(\theta', \hat{\theta}^{(j)})$  over  $\theta'$ .
4. Iterate steps 2 and 3 until convergence.

## References

- 1) The Elements of Statistical Learning - Trevor Hastie, Robert Tibshirani, Jerome Friedman
- 2) <https://en.wikipedia.org/>
- 3) Victor Lavrenko - Youtube