The EM Algorithm

Ritwik Sahani/ Aayush Goyal

IIT Hyderabad

February 28, 2019

Suppose we have some data points $x_1 ...... x_n$

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

Can we find the parameters μ_b, σ^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.

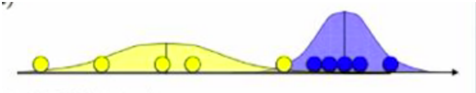
But what we do know is that these come from two gaussians, whose parameters μ, σ^2 is known

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

Yes, this can be done by claculating 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

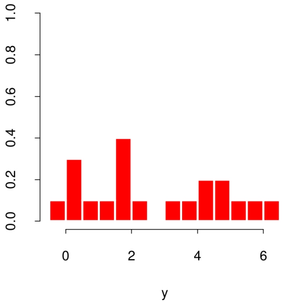
Which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.
⋆ **Basically, we will again calculate the parameters for the next E -step**

Introduction

The EM algorithm is used to find (local) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either missing values exist among the data, or the model can be formulated more simply by assuming the existence of further unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component to which each data point belongs.

Two-Component Mixture Model

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



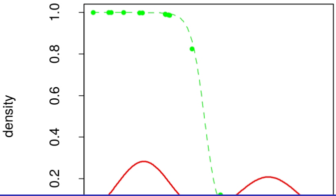


TABLE 8.1. *Twenty fictitious data points used in the two-component mixture example in Figure 8.5.*

-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 Bimodalty of data Gaussian distribution would be inappropriate. So we will model Y as a mixture of two distinct distributions.

$Y_1 \equiv N(\mu_1, \sigma_{12})$
$Y_2 \equiv N(\mu_2, \sigma_{22})$
$Y = (1 - \Delta) * Y_1 + \Delta * Y_2$
where $\Delta \in \{0, 1\}$ with $Pr(\Delta = 1) = \pi$

This generative representation is explicit: generate a $\Delta \in \{0, 1\}$ with probability π, and then depending on the outcome, deliver either $\mathrm{Y}_1 or Y_2. Let \phi$ θ (x) denote the normal density with parameters $\theta = (\mu, \sigma_2)$. Then the density of Y is $gY(y) = (1 - \pi)\phi\theta_1(y) + \pi\phi\theta_2(y)$.

Now suppose we wish to fit this model to the data in Figure 8.5 by maximum likelihood.The parameters are
$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_{12}, \mu_2, \sigma_{22})$
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 θ_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.

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

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

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

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, \; i = 1, 2, \ldots, N. \tag{8.42}$$

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

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i},$$

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

4. Iterate steps 2 and 3 until convergence.