

Machine Learning: Advanced Probabilistic Methods (2015), P.Marttinen
 [The example is modified from lecture slides for the course "Laskentaintensiiviset tilastolliset menetelmät (Computational statistics)" by Petri Koistinen]

Suppose that we have N independent observations $\mathbf{x} = (x_1, \dots, x_N)$ from a two-component mixture of univariate Gaussian distributions

$$p(x_n|\theta) = \frac{1}{2}N(x_n|0, 1) + \frac{1}{2}N(x_n|\theta, 1). \quad (1)$$

This means that with probability $1/2$ the observation x_n is generated from the first component $N(x_n|0, 1)$, and with probability $1/2$ from the second component $N(x_n|\theta, 1)$. The model (1) has one unknown parameter, θ , representing the mean of the second component, and we would like to estimate it using maximum likelihood

$$\hat{\theta} = \arg \max_{\theta} \{\log p(\mathbf{x}|\theta)\}.$$

We do this by the EM-algorithm (although direct numerical optimization would also be straightforward for this simple model).

First we formulate the model using the **latent variable representation**, and introduce variables $\mathbf{z} = (z_1, \dots, z_N)$ which explicitly specify the component responsible for generating observation x_n . In detail:

$$z_n = (z_{n1}, z_{n2})^T = \begin{cases} (1, 0)^T, & (x_n \text{ is from } N(x_n|0, 1)) \\ (0, 1)^T, & (x_n \text{ is from } N(x_n|\theta, 1)) \end{cases}.$$

When we define the distributions for the latent variable model as follows

$$p(z_{n1} = 1) = p(z_{n2} = 1) = 0.5$$

and

$$p(x_n|z_n, \theta) = \begin{cases} N(x_n|0, 1), & \text{if } z_{n1} = 1 \\ N(x_n|\theta, 1), & \text{if } z_{n2} = 1 \end{cases}$$

it is easy to see that the marginal distribution of x_n obtained by summing over the latent variables

$$p(x_n|\theta) = \sum_{z_n} p(x_n|z_n, \theta)p(z_n)$$

is equal to the original distribution (1).

In the EM-algorithm we will maximize the expectation of the **log-likelihood of the complete data** (\mathbf{x}, \mathbf{z}) :

$$\begin{aligned} \log p(\mathbf{x}, \mathbf{z}|\theta) &= \log \left\{ \prod_{n=1}^N p(x_n, z_n|\theta) \right\} = \sum_{n=1}^N \log p(x_n, z_n|\theta) \\ &= \sum_{n=1}^N \log [0.5 \times N(x_n|0, 1)^{z_{n1}} \times N(x_n|\theta, 1)^{z_{n2}}] \\ &= \sum_{n=1}^N \{z_{n1} \log [N(x_n|0, 1)] + z_{n2} \log [N(x_n|\theta, 1)]\} + \text{const} \quad (2) \end{aligned}$$

E-step ¹⁰: Compute the posterior distribution of the latent variables, given the current estimate θ_0 of θ :

$$\begin{aligned} p(z_{n1} = 1|x_n, \theta_0) &\propto p(z_{n1} = 1)p(x_n|z_n, \theta_0) \\ &= 0.5 \times N(x_n|0, 1) \end{aligned} \quad (3)$$

$$\begin{aligned} p(z_{n2} = 1|x_n, \theta_0) &\propto p(z_{n2} = 1)p(x_n|z_n, \theta_0) \\ &= 0.5 \times N(x_n|\theta_0, 1) \end{aligned} \quad (4)$$

By normalizing (3) and (4) we get

$$\gamma(z_{n2}) \equiv p(z_{n2} = 1|x_n, \theta_0) = \frac{N(x_n|\theta_0, 1)}{N(x_n|0, 1) + N(x_n|\theta_0, 1)}. \quad (5)$$

E-step ²⁰: Evaluate the expectation of the complete data log-likelihood (2) over the posterior distribution of the latent variables (5):

$$\begin{aligned} Q(\theta, \theta_0) &= E_{\mathbf{z}|\mathbf{x}, \theta_0} [\log p(\mathbf{x}, \mathbf{z}|\theta)] \\ &= \sum_{n=1}^N \{E[z_{n1}] \log [N(x_n|0, 1)] + E[z_{n2}] \log [N(x_n|\theta, 1)]\} \\ &= \sum_{n=1}^N \{[1 - \gamma(z_{n2})] \log [N(x_n|0, 1)] + \gamma(z_{n2}) \log [N(x_n|\theta, 1)]\}. \end{aligned} \quad (6)$$

Note that in (6) we've discarded the term not dependent on θ in equation (2). As a matter of fact, the first term in each sum could also be discarded, but we retain it here for clarity.

M-step: Maximize $Q(\theta, \theta_0)$ with respect to θ . To differentiate $Q(\theta, \theta_0)$, we first note the following result, which can be verified by straightforward computation

$$\frac{d}{d\theta} N(x_n|\theta, 1) = N(x_n|\theta, 1)(x_n - \theta).$$

With this result at hand, we can write

$$\begin{aligned} \frac{d}{d\theta} Q(\theta, \theta_0) &= \frac{d}{d\theta} \sum_{n=1}^N \{[1 - \gamma(z_{n2})] \log [N(x_n|0, 1)] + \gamma(z_{n2}) \log [N(x_n|\theta, 1)]\} \\ &= \sum_{n=1}^N \frac{\gamma(z_{n2})}{N(x_n|\theta, 1)} N(x_n|\theta, 1)(x_n - \theta) = \sum_{n=1}^N \gamma(z_{n2})(x_n - \theta). \end{aligned}$$

Setting $\frac{d}{d\theta} Q(\theta, \theta_0) = 0$, we get

$$\begin{aligned} \theta &= \frac{\sum_{n=1}^N \gamma(z_{n2})x_n}{\sum_{n=1}^N \gamma(z_{n2})} \\ &= \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{n2})x_n, \end{aligned} \quad (7)$$

where we have defined $N_k = \sum_{n=1}^N \gamma(z_{n2})$, which can be interpreted as the effective number of observations assigned to component 2. Note that (7) has an intuitive interpretation: the mean of component (cluster) 2 is obtained as a weighted average of all points in the data set, in which the weighting factor for data point x_n is given by the posterior probability (or responsibility) $\gamma(z_{n2})$ that the 2nd component was responsible for generating x_n .

Code to run the EM-algorithm: *simple_em.m*