Fitting Finite Mixtures of Gaussian Distribution using the EM-Algorithm

gustavoc@msu.edu

In a finite mixture model the density of a RV is modeled as the weighted sum of a finite number of densities. In this note we consider finite mixtures with Gaussian components. The density function of a mixture with 2 Gaussian components is:

$$p(x_i|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\alpha_1) = \alpha_1(2\pi\sigma_1^2)^{-\frac{1}{2}}e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + (1-\alpha_1)\,(2\pi\sigma_2^2)^{-\frac{1}{2}}e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}}$$

Above, $\mu_.$ and $\sigma_.^2$ and the means and variances of each of the components and $0 < \alpha_1 < 1$ is a mixture proportion. The model can be naturally extended to K components.

Likelihood Function

Assuming IID data, the joint density of the data is

$$p(x_1,\dots,x_n|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\alpha_1) = \prod_{i=1}^{n}\alpha_1 dnorm(x_i|\mu_1,\sigma_1^2) + (1-\alpha_1)\,dnorm(x_i|\mu_1,\sigma_1^2)$$

Where $dnorm(x_i|\mu_*,\sigma_*^2) = (2\pi\sigma_*^2)^{-\frac{1}{2}}e^{-\frac{(x_i-\mu_*)^2}{2\sigma_*^2}}$.

The model parameters, $\boldsymbol{\theta} = \{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\alpha_1\}$, can be estimated via maximum likelihood. Maximization could be done using general purpose optimization algorithms (e.g., those implemented in `optim()`); alternatively, we can maximize the likelihood using using the EM-algorithm.

Augmented likelihood

To facilitate the implementation of the EM-algorithm we introduce a latent variable $z_i \in \{1,2\}$ which indicates whether the ith observation comes from 1st or 2nd component of the mixture. Parameter α_1 can be interpreted as the proportion of the observations coming from the firsth component; therefore $p(z_i = 1|\alpha_1) = \alpha_1$ and $p(z_i = 2|\boldsymbol{\theta}) = 1 - \alpha_1$; thus $p(z_i = 1|\alpha_1) = \alpha_1^{1(z_i=1)}(1-\alpha_1\,)^{1(z_i=1)}$.

The augmented (or complete-data) likelihood is the joint distribution of the observed (x) and the missing (z) data given the parameters ($\boldsymbol{\theta}$). For the ith-data point the augmented likelihood is:

$$p(x_i, z_i|\boldsymbol{\theta}) = p(x_i|z_i,\theta) \times p(z_i|\boldsymbol{\theta})$$

$$= dnorm(x_i|\mu_1,\sigma_1^2)^{1(z_i=1)} dnorm(x_i|\mu_1,\sigma_1^2)^{1(z_i=2)} \times \alpha_1^{1(z_i=1)}(1-\alpha_1\,)^{1(z_i=1)}$$

Therefore, assuming IID data the complete-data likelihood becomes

$$p(\boldsymbol{x}, \boldsymbol{z} \,|\boldsymbol{\theta}) = \prod_{i=1}^{n} dnorm(x_i|\mu_1, \sigma_1^2)^{1(z_i=1)} dnorm(x_i|\mu_1, \sigma_1^2)^{1(z_i=2)} \times \alpha_1^{1(z_i=1)} (1-\alpha_1\,)^{1(z_i=1)}$$

E-step

To perform the E-step we must drive the expected value of the logarithm of the complete-data likelihood with respect to the distribution of the missing data given the observed data and the parameters, that is $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{\theta}$. The logarithm of the complete likelihood takes the form

$$l(\boldsymbol{x}, \boldsymbol{z} \,|\boldsymbol{\theta}) = \sum_{i=1}^{n} 1(z_i = 1) \log\left[dnorm(x_i|\mu_1, \sigma_1^2)\;\right] + 1(z_i = 2) \log\left[dnorm(x_i|\mu_1, \sigma_1^2)\;\right] + 1(z_i = 1) \log[\alpha_1] + 1(z_i = 1) \log[1-\alpha_1]$$

The expected value the above expression is

$$l(\boldsymbol{x}, \boldsymbol{z} \,|\boldsymbol{\theta}) = \textstyle\sum_{i=1}^{n} w_i \log\left[dnorm(x_i|\mu_1, \sigma_1^2)\;\right] + (1-w_i) \log\left[dnorm(x_i|\mu_1, \sigma_1^2)\;\right] + w_i \log[\alpha_1] + (1-w_i) \log[1-\alpha_1] \qquad [1]$$

Where $w_i = E(z_i = 1|x_i, \boldsymbol{\theta}) = p(z_i = 1|x_i, \boldsymbol{\theta})$ is the success probability of the ith latent variable. This probability is given by

$$p(z_i = 1|x_i, \boldsymbol{\theta}) = \frac{p(z_i=1|x_i,\boldsymbol{\theta})}{p(z_i=1|x_i,\boldsymbol{\theta}) + p(z_i=2|x_i,\boldsymbol{\theta})}$$

Using Bayes' rule

$$p(z_i = 1|x_i, \boldsymbol{\theta}) = \frac{p(x_i|z_i=1)p(z_i=1)}{p(x_i|z_i=1)p(z_i=1) + p(x_i|z_i=2)p(z_i=2)} = \frac{A}{A+B} \;[2]$$

Where $A = dnorm(x_i|\mu_1, \sigma_1^2)\;\; \times \alpha_1$ and $B = dnorm(x_i|\mu_1, \sigma_1^2)\;\; \times (1-\alpha_1)$

M-Step:

In the M-step we maximize [1] with respect to each of the parameters of the mixture. Note that [1] is a weighted log-likelihood. It can be shown that the ML estimates of the parameters are given by the following weighted means and weighted variances:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}\;;\; \hat{\mu}_2 = \frac{\sum_{i=1}^{n}(1-w_i) y_i}{\sum_{i=1}^{n}(1-w_i)};\; \hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n} w_i (y_i - \hat{\mu}_1)^2}{\sum_{i=1}^{n} w_i};\;\; \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{n}(1-w_i)(y_i - \hat{\mu}_2)^2}{\sum_{i=1}^{n}(1-w_i)}$$