Expectation-Maximization for the Gaussian Mixture Model

Thomas Bonald
Telecom ParisTech
thomas.bonald@telecom-paristech.fr

January 2019

In this note, we present a clustering technique based on the Gaussian mixture model. Data samples are assumed to be generated by a mixture of k Gaussian distributions, whose parameters are estimated by an iterative method known as Expectation-Maximization (the EM algorithm). We show that the k-means algorithm corresponds to the particular case where all Gaussian distributions are assumed to have the same diagonal covariance matrix, with infinitely small variance.

1 Gaussian mixture model

Gaussian model. The Gaussian distribution of some d-dimensional vector X is characterized by its mean μ and its covariance matrix Σ. We use the following notation:

$$X \sim \mathcal{N}(\mu, \Sigma).$$

The random vector X has a density f if and only if its covariance matrix is invertible, in which case:

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)},$$

where $|\Sigma|$ is the determinant of Σ.

Gaussian mixture model. Now consider k such distributions, with respective density functions $f_1, \dots, f_k$ and respective parameters $(\mu_1, \Sigma_1), \dots, (\mu_k, \Sigma_k)$. Let $\pi_1, \dots, \pi_k$ be any probability distribution on $\{1, \dots, k\}$. Select the j-th distribution with probability π_j, that is:

$$X \sim \mathcal{N}(\mu_Z, \Sigma_Z) \quad \text{with} \quad Z \sim \pi. \tag{1}$$

The vector X has density:

$$p_\theta(x) = \sum_{j=1}^{k} \pi_j f_j(x), \tag{2}$$

where the parameter $\theta = (\pi, \mu, \Sigma)$ consists of:

- the mixing distribution $\pi = (\pi_1, \dots, \pi_k)$,
- the set of means $\mu = (\mu_1, \dots, \mu_k)$,
- the set of covariance matrices $\Sigma = (\Sigma_1, \dots, \Sigma_k)$.

2 Maximum likelihood

Assume we seek to estimate the parameter θ based on n i.i.d. samples $x_1, \ldots, x_n$ of the Gaussian mixture model. Denoting by x the vector $(x_1, \ldots, x_n)$, we get the likelihood:

$$p_\theta(x) = \prod_{i=1}^{n} p_\theta(x_i),$$

and the log-likelihood:

$$\ell(\theta) = \log p_\theta(x) = \sum_{i=1}^{n} \log p_\theta(x_i).$$

In view of (2),

$$\ell(\theta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j f_j(x_i) \right). \tag{3}$$

The maximum likelihood estimator is:

$$\theta^\star = \arg\max_\theta \ell(\theta). \tag{4}$$

This problem is hard to solve in practice, even numerically, since the function $\theta \mapsto -\ell(\theta)$ is not convex in general.

3 Expectation-Maximization

The Expectation-Maximization (EM) algorithm is an iterative method for finding a local maximum of the likelihood. This technique applies to any mixture model (in fact, any model with latent variables). It is based on the observation that the problem (4) would be easy to solve given the latent variables $z_1, \ldots, z_n$ having been used implicitly in (1) to generate the data samples $x_1, \ldots, x_n$.

Latent variables. Let z the vector of latent variables $(z_1, \ldots, z_n)$. In view of (2), the Gaussian mixture model is the marginal distribution of the joint distribution:

$$p_\theta(x, z) = p_\theta(z) p_\theta(x|z),$$

with

$$p_\theta(z) = \prod_{i=1}^{n} p_\theta(z_i) = \prod_{i=1}^{n} \pi_{z_i} \quad \text{and} \quad p_\theta(x|z) = \prod_{i=1}^{n} p_\theta(x_i|z_i) = \prod_{i=1}^{n} f_{z_i}(x_i).$$

Given the latent variables, the log-likelihood becomes:

$$\ell(\theta; z) = \log p_\theta(x, z) = \sum_{i=1}^{n} \log \pi_{z_i} + \sum_{i=1}^{n} \log f_{z_i}(x_i),$$

so that each set of parameters $(\mu_1, \Sigma_1), \ldots, (\mu_k, \Sigma_k)$ can be estimated separately using the corresponding samples (we refer the reader to Appendix A for the maximum likelihood estimator of a Gaussian distribution). Specifically, the log-likelihood $\ell(\theta; z)$ is maximum for the empirical mixing distribution $(\hat{\pi}_1, \ldots, \hat{\pi}_k)$:

$$\forall j = 1, \ldots, k, \quad \hat{\pi}_j = \frac{n_j}{n}, \tag{5}$$

and the empirical means and covariance matrices $(\hat{\mu}_1, \hat{\Sigma}_1), \ldots, (\hat{\mu}_k, \hat{\Sigma}_k)$:

$$\forall j = 1, \ldots, k, \quad \hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n} 1_{\{z_i = j\}} x_i, \quad \hat{\Sigma}_j = \frac{1}{n_j} \sum_{i=1}^{n} 1_{\{z_i = j\}} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T, \tag{6}$$

where

$$n_j = \sum_{i=1}^{n} 1_{\{z_i=j\}}$$

is the number of samples generated according to the j-th Gaussian distribution (which we assume positive). The key problem is that the latent variables $z_1, \ldots, z_n$ are unknown.

Estimation of the latent variables. The conditional distribution of the latent variables given the data samples follows from:

$$p_\theta(x, z) = p_\theta(x) p_\theta(z|x). \tag{7}$$

Since:

$$p_\theta(x, z) = \prod_{i=1}^{n} \pi_{z_i} f_{z_i}(x_i),$$

we get:

$$p_\theta(z|x) = \prod_{i=1}^{n} p_\theta(z_i|x_i) \propto \prod_{i=1}^{n} \pi_{z_i} f_{z_i}(x_i).$$

In particular, the probability that sample i comes from distribution j is:

$$p_{ij} \propto \pi_j f_j(x_i). \tag{8}$$

Now given some initial parameter θ_0, we can use the corresponding distribution of the latent variables given by (8) to get the *expected* log-likelihood of x:

$$\begin{aligned} \ell_{\theta_0}(\theta) &= \sum_z p_{\theta_0}(z|x) \ell(\theta; z), \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} p_{ij} \left(\log \pi_j + \log f_j(x_i) \right). \end{aligned}$$

This expected log-likelihood is maximum for the empirical mixing distribution $(\hat{\pi}_1, \ldots, \hat{\pi}_k)$:

$$\forall j = 1, \ldots, k, \quad \hat{\pi}_j = \frac{n_j}{n}, \tag{9}$$

and the empirical means and covariance matrices $(\hat{\mu}_1, \hat{\Sigma}_1), \ldots, (\hat{\mu}_k, \hat{\Sigma}_k)$:

$$\forall j = 1, \ldots, k, \quad \hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n} p_{ij} x_i, \quad \hat{\Sigma}_j = \frac{1}{n_j} \sum_{i=1}^{n} p_{ij} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T, \tag{10}$$

where

$$n_j = \sum_{i=1}^{n} p_{ij}$$

is the *expected* number of samples generated according to the j-th Gaussian distribution.

Thus, starting from some initial parameter θ_0, we can compute the conditional distribution of the latent variables, given the data samples, and deduce a new estimate of the parameter, $\theta_1 = (\hat{\pi}, \hat{\mu}, \hat{\Sigma})$. By successive iterations, we obtain a sequence of parameters $\theta_0, \theta_1, \theta_2, \ldots$ which is expected to converge to a good approximation of the optimal parameter $\theta^\star$ (i.e., that solving the problem (4)). We shall prove in the next section that the corresponding sequence of log-likelihoods $\ell(\theta_0), \ell(\theta_1), \ell(\theta_2), \ldots$ is non-decreasing, which guarantees that the EM algorithm converges a local maximum of the likelihood. This is not the global maximum of the likelihook in general.

EM algorithm. A pseudo-code of the algorithm is shown below. The outcome is a *soft* clustering of the data samples, with p_{ij} the probability that sample i belongs to cluster k. A regular clustering can be obtained by selecting for each sample i the cluster j maximizing p_{ij}.

Algorithm 1: EM algorithm for the Gaussian mixture model

Input: Data samples $x_1, \ldots, x_n$; number of clusters k
Output: $p_1, \ldots, p_n$, probability distributions of samples $1, \ldots, n$ over the k clusters

Sample random values $\mu_1, \ldots, \mu_k$ from $x_1, \ldots, x_n$
$\bar{x} \leftarrow \frac{1}{n} \sum_{i=1}^{n} x_i$
$\sigma^2 \leftarrow \frac{1}{n} \sum_{i=1}^{n} ||x_i - \bar{x}||^2$
for $j = 1$ **to** k **do**
    $\Sigma_j \leftarrow \frac{\sigma^2}{k} I$
    $\pi_j \leftarrow 1$
while no convergence **do**
    `// Expectation`
    for $i = 1$ **to** n **do**
        $s \leftarrow 0$
        for $j = 1$ **to** k **do**
            $p_{ij} \leftarrow \frac{\pi_j}{\sqrt{|\Sigma_j|}} e^{-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)}$
            $s \leftarrow s + p_{ij}$
        for $j = 1$ **to** k **do**
            $p_{ij} \leftarrow p_{ij} / s$
    `// Maximization`
    for $j = 1$ **to** k **do**
        $\pi_j \leftarrow 0$
        $\mu_j \leftarrow 0$
        for $i = 1$ **to** n **do**
            $\pi_j \leftarrow \pi_j + p_{ij}$
            $\mu_j \leftarrow \mu_j + p_{ij} x_i$
        $\mu_j \leftarrow \mu_j / \pi_j$
        $\Sigma_j \leftarrow 0$
        for $i = 1$ **to** n **do**
            $\Sigma_j \leftarrow \Sigma_j + p_{ij} (x_i - \mu_j)(x_i - \mu_j)^T$
        $\Sigma_j \leftarrow \Sigma_j / \pi_j$

A key problem is in the choice of the initial parameter θ_0, and more specifically in the choice of the initial means $\mu_1, \ldots, \mu_k$, corresponding to the cluster centers. In the above pseudo-code, this is obtained by random sampling k values among the data samples $x_1, \ldots, x_n$, as in the k-means algorithm. Since this initial choice has a strong impact on the final result, several independent instances of the algorithm can be run, the best instance, i.e., that maximizing (3), being selected eventually. Another common strategy consists in selecting the cluster centers far from one another, as in the k-means++ algorithm.

The choice of the initial values of the covariance matrices is also critical. Here σ^2 is chosen as the average square distance between data points, in view of the equality

$$\frac{1}{n}\sum_{i=1}^{n}||x_i - \bar{x}||^2 = \frac{1}{n^2}\sum_{i,j=1}^{n}||x_i - x_j||^2.$$

If σ^2 is much larger than the typical square distance between data samples, then the probability distributions $p_1, \ldots, p_n$ tend to be close to uniform, the means $\mu_1, \ldots, \mu_k$ converging to $\bar{x}$, the center of the data samples.

The number of iterations depends on the stopping criterion. For instance, one may decide that convergence has occured whenever the main cluster of each data sample (that maximizing p_{ij} for data sample i) remains unchanged.

The complexity of the algorithm is in $O(nk)$ per iteration, which may be prohibitive for large values of k. The complexity may be reduced to $O(nm)$, for some integer m, by looking for the m nearest clusters of each data sample, using some appropriate data structure.

4 Convergence

In view of (7), the log-likelihood can be written:

$$\ell(\theta) = \log p_\theta(x, z) - \log p_\theta(z|x), \tag{11}$$

whenever $p_\theta(x, z) > 0$.

Now let θ_t be the estimate of θ at step t of the algorithm. Since the equality (11) holds for *each* value of z, provided $p_\theta(x, z) > 0$, we can take the expectation with respect to the corresponding conditional distribution of the latent variables, $p_{\theta_t}(z|x)$, and we obtain:

$$\ell(\theta) = \ell_{\theta_t}(\theta) - \sum_z p_{\theta_t}(z|x) \log p_\theta(z|x).$$

Now the difference in log-likelihood is:

$$\ell(\theta) - \ell(\theta_t) = \ell_{\theta_t}(\theta) - \ell_{\theta_t}(\theta_t) + D(\theta_t||\theta),$$

where

$$D(\theta_t||\theta) = \sum_z p_{\theta_t}(z|x) \log \frac{p_{\theta_t}(z|x)}{p_\theta(z|x)}$$

is the Kullback-Leibler divergence between the probability distributions $p_{\theta_t}(z|x)$ and $p_\theta(z|x)$. This quantity is non-negative (this is Gibbs' inequality, see Appendix B). Since:

$$\theta_{t+1} = \arg\max_\theta \ell_{\theta_t}(\theta),$$

we get:

$$\ell(\theta_{t+1}) - \ell(\theta_t) = \ell_{\theta_t}(\theta_{t+1}) - \ell_{\theta_t}(\theta_t) + D(\theta_t||\theta_{t+1}) \geq D(\theta_t||\theta_{t+1}) \geq 0,$$

showing that the corresponding sequence of log-likelihoods $\ell(\theta_0), \ell(\theta_1), \ell(\theta_2), \ldots$ is non-decreasing and thus converges.

5 Comparison with k-means

Consider the Gaussian mixture model with common covariance matrix $\sigma^2 I$, for some parameter $\sigma > 0$, and uniform mixing distribution:

$$X \sim \mathcal{N}(\mu_Z, \sigma^2 I) \quad \text{with} \quad Z \sim \mathcal{U}(\{1, \ldots, k\}).$$

The density becomes:

$$f(x) = \frac{1}{k} \sum_{j=1}^{k} f_j(x),$$

with:

$$f_j(x) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} e^{-\frac{||x-\mu_j||^2}{2\sigma^2}}.$$

The variance σ^2 is assumed to be known so that the only parameter is $\theta = \mu$, the vector of means. We refer to this model as the *symmetric* Gaussian mixture model.

Algorithm 2: EM algorithm for the symmetric Gaussian mixture model

Input: Data samples $x_1, \ldots, x_n$; distance σ; number of clusters k
Output: $p_1, \ldots, p_n$, probability distributions of samples $1, \ldots, n$ over the k clusters

Sample random values $\mu_1, \ldots, \mu_k$ from $x_1, \ldots, x_n$
while no convergence **do**
    `// Expectation`
    for $i = 1$ **to** n **do**
        $s \leftarrow 0$
        for $j = 1$ **to** k **do**
            $p_{ij} \leftarrow e^{-\frac{||x_i-\mu_j||^2}{2\sigma^2}}$
            $s \leftarrow s + p_{ij}$
        for $j = 1$ **to** k **do**
            $p_{ij} \leftarrow p_{ij}/s$
    `// Maximization`
    for $j = 1$ **to** k **do**
        $\pi_j \leftarrow 0$
        $\mu_j \leftarrow 0$
        for $i = 1$ **to** n **do**
            $\pi_j \leftarrow \pi_j + p_{ij}$
            $\mu_j \leftarrow \mu_j + p_{ij} x_i$
        $\mu_j \leftarrow \mu_j/\pi_j$

When $\sigma^2 \to 0$, the expectation step becomes:

$$p_{ij} = \begin{cases} 1 & \text{if } j = l, \\ 0 & \text{otherwise,} \end{cases}$$

where l is the index for which the distance $||x_i - \mu_l||$ is minimum (assuming this index is unique). The algorithm is exactly k-means.

In general, the algorithm provides a soft clustering, with the parameter σ controlling the spread of each cluster. When $\sigma^2 \to +\infty$, each sample belongs to each cluster with the same probability $1/k$.

6 A simple Gaussian mixture model

Finally, consider the Gaussian mixture model where all covariance matrices are diagonal. This is a good trade-off between the Gaussian mixture model (with $O(kd^2)$ parameters to be learned, where d is the dimension of the data samples) and the k-means algorithm, corresponding to the simplistic case of a uniform mixing distribution and covariance matrices equal to $\sigma^2 I$ for some fixed, small parameter σ^2. The EM algorithm is the same as Algorithm 1, with the update of the covariance matrices replaced by:

$$\Sigma_j \leftarrow \Sigma_j + p_{ij}\mathrm{diag}((x_i - \mu_j)^2),$$

where $(x_i - \mu_j)^2$ refers to the vector equal to the square of vector $x_i - \mu_j$ componentwise. The dependency across dimensions is no longer taken into account, but the algorithm is more robust in that the inversion of the covariance matrices is straightforward.

Further reading

- The initial paper on the EM algorithm [Dempster et al., 1977].
- A concise tutorial on the EM algorithm and variants [Roche, 2011].

Appendix

A Maximum likelihood for the Gaussian model

Consider the Gaussian model in dimension d:

$$p_\theta(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)},$$

with parameter $\theta = (\mu, \Sigma)$. For n i.i.d. samples $x_1, \ldots, x_n$ of the distribution, we get:

$$p_\theta(x) = p_\theta(x_1) \ldots p_\theta(x_n).$$

The log-likelihood is:

$$\ell(\theta) = \log p_\theta(x) = \sum_{i=1}^{n} \log p_\theta(x_i),$$

that is:

$$\ell(\theta) = c - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu).$$

with $c = -\frac{nd}{2} \log(2\pi)$.

The gradient in μ is the vector:

$$\frac{\partial \ell(\theta)}{\partial \mu} = -\sum_{i=1}^{n} \Sigma^{-1}(x_i - \mu),$$

which is equal to 0 for $\mu = \hat{\mu}$ with:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This is the empirical mean of the samples.

Now let $\Lambda = \Sigma^{-1}$. Since $|\Lambda| = |\Sigma|^{-1}$, we can rewrite the log-likelihood as:

$$\ell(\theta) = c + \frac{n}{2}\log|\Lambda| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Lambda(x_i - \mu).$$

We obtain[1]

$$\frac{\partial\ell(\theta)}{\partial\Lambda} = \frac{n}{2}\Lambda^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T,$$

which is equal to 0 for $\Lambda^{-1} = \hat{\Sigma}$, with

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T.$$

For $\mu = \hat{\mu}$, this is the empirical covariance matrix of the samples.

B Kullback-Leibler divergence and Gibbs' inequality

Let p and q be two probability distributions over $\{1, \ldots, k\}$. The Kullback-Leibler divergence between p and q is defined by:

$$D(p||q) = \sum_{j=1}^{k} p_j \log\frac{p_j}{q_j}.$$

It is a measure of how the probability distributions p and q differ. Observe that $D(p||q) = +\infty$ whenever $q_j = 0$ while $p_j > 0$ for some j. We have:

$$D(p||q) \geq 0,$$

with equality if and only if $p = q$. This is Gibbs' inequality, which follows from Jensen's inequality on observing that:

$$D(p||q) = \mathrm{E}\left(\log\frac{p_Z}{q_Z}\right) = -\mathrm{E}\left(\log\frac{q_Z}{p_Z}\right) \geq -\log\mathrm{E}\left(\frac{q_Z}{p_Z}\right) = -\log 1 = 0,$$

where the expectation is taken over Z, a random variable having distribution p. Since the log is strictly concave, the inequality is an equality if and only if $\frac{q_j}{p_j}$ is a constant, for each j such that $p_j > 0$, which means that $p = q$.

References

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological).*

[Roche, 2011] Roche, A. (2011). EM algorithm and variants: An informal tutorial. *arXiv:1105.1476.*

[1] For any matrix A, the gradient of the determinant $|A|$ in A is the comatrix of A. In particular, the gradient of $\log|A|$ in A is A^{-1} for any positive, symmetric matrix A. Moreover, for any vectors u, v, the gradient of u^TAv in A is the matrix uv^T.