Gaussian Mixture Model

Oliver Zhao

1 Description

A mixture model is a probabilistic model used to describe sub-populations within a data set, without the sub-populations needing to be identified by a user. A Gaussian mixture model is a kernel method with the form:

$$f(x) = \sum_{m=1}^{M} \alpha_m \phi(x; \mu_m, \mathbf{\Sigma}_m), \tag{1.1}$$

Where α_m are the mixing proportions such that $\sum_m \alpha_m = 1$. Each Gaussian density has a mean μ_m and covariance matrix $\mathbf{\Sigma}_m$. The parameters are often fit by maximum likelihood with the expectation-maximization (EM) algorithm, an iterative method to find maximum likliehood estimates. Sometimes to simplify the problem, the covariance matrices are constrained to be scalar such that $\mathbf{\Sigma}_m = \sigma_m \mathbf{I}$.

The mixture model provides an estimate of the probability that observation i belongs to the component m, where

$$\hat{r}_{im} = \frac{\hat{\alpha}_m \phi(x_i; \hat{\mu}_m, \hat{\Sigma}_m)}{\sum_{k=1}^{M} \hat{\alpha}_k \phi(x_i; \hat{\mu}_k, \hat{\Sigma}_k)} \tag{1.2}$$

2 EM Algorithm

2.1 Bi-modal Distributions

Consider a bi-modal distribution of data points that we attempt to model as Y with two separate Gaussian distributions Y_1 and Y_2.

$$
\begin{aligned}
&Y_1\ N(\mu_1, \sigma_1^2) \\
&Y_2\ N(\mu_2, \sigma_2^2) \\
&Y = (1-\Delta)\cdot Y_1 + \Delta \cdot Y_2
\end{aligned} \tag{2.1}
$$

Where $\Delta\ \epsilon\ \{0,1\}$ with $\Pr(\Delta = 1) = \pi$. Let $\phi_\theta(x)$ denote the normal density with parameters $\theta = (\mu, \sigma^2)$. Then the density of Y is

$$
g_Y(y) = (1-\pi)\phi_{\theta_1}(y) + \pi\phi_{\theta_2}(y). \tag{2.2}
$$

Suppose we wish to fit this model by maximum likelihood. The parameters are

$$
\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2). \tag{2.3}
$$

The log-likelihood based on the N training cases is

$$
\ell(\theta; \mathbf{Z}) = \sum_{i=1}^{N} \log[(1-\pi)\phi_{\theta_1}(y_i) + \pi\phi_{\theta_2}(y_i)]. \tag{2.4}
$$

Direct maximization of $\ell(\theta; \mathbf{Z})$ is difficult due to the sum of terms in the logarithm. A simpler approach is to consider the unobserved latent variables Δ_i, taking values 0 or 1. If $\Delta_i = 1$ then Y_i comes from model 2, otherwise it comes from model 1. Suppose we knew the values of the Δ_i's. Then the log-likelihood would be

$$
\begin{aligned}
\ell_0(\theta; \mathbf{Z}, \boldsymbol{\Delta}) = &\sum_{i=1}^{N}[(1-\Delta_i)\log\phi_{\theta_1}(y_i) + \Delta_i\log\phi_{\theta_2}(y_i)] \\
&+ \sum_{i=1}^{N}[(1-\Delta_i)\log(1-\pi) + \Delta_i\log\pi],
\end{aligned} \tag{2.5}
$$

And the maximum likelihood estimates of μ_1 and σ_1^2 would be the sample mean and variance for those data with $\Delta_i = 0$, and similarly those for μ_2 and σ_2^2 would be the sample mean and variance for those data with $\Delta_i = 1$. The estimate of π would be the proportion of $\Delta_i = 1$. Since we do not know the values of the Δ_i's, we substitute for each Δ_i with its expected value in an iterative manner:

$$
\gamma_i(\theta) = \mathrm{E}(\Delta_i|\theta, \mathbf{Z}) = \Pr(\Delta_i = 1|\theta, \mathbf{Z}), \tag{2.6}
$$

Which is called the responsibility of model 2 for observation i.

2.2 EM Algorithm for Two-Component Gaussian Mixture

Description: ϕ is a normal PDF, while $\hat{\pi}$ is the estimated mixing probability.

---

1. Take the 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 for each data point:

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\theta_2}(y_i)}{(1-\hat{\pi})\phi_{\theta_1}(y_i) + \hat{\pi}\phi_{\theta_2}(y_i)},\ i = 1, 2, \ldots, N. \tag{2.7}$$

3. **Maximization Step:** Compute the weighted means, variances, and mixing probability:

$$\begin{aligned} \hat{\mu_1} &= \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu_1})^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \\ \hat{\mu_2} &= \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu_2})^2}{\sum_{i=1}^N \hat{\gamma}_i}, \\ \hat{\pi} &= \sum_{i=1}^N \hat{\gamma}_i / N. \end{aligned} \tag{2.8}$$

4. Iterate steps 2 and 3 until convergence.

2.3 General EM Algorithm for Gaussian Mixture Models

The algorithm in Section (2.2) can be intuitively expanded to three or more multivariate distributions:

Description: Consider a data set with N data points with p features, with the i data vector denoted as $\mathbf{y_i}$. To fit a GMM with k distributions:

---

1. Take the initial guesses for the parameters:

$$\begin{aligned} \text{Mean Vectors:} \quad & \hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_k \\ \text{Covariance Matrices:} \quad & \hat{\boldsymbol{\Sigma}}_1^2, \ldots, \hat{\boldsymbol{\Sigma}}_k^2 \\ \text{Weights:} \quad & \hat{\pi}_1, \ldots, \hat{\pi}_k \end{aligned}$$

2. **Expectation Step:** Compute the responsibilities for each data point:

$$\hat{\gamma}_{i,j} = \frac{\hat{\pi}_j \phi_{\theta_j}(\mathbf{y_i})}{\sum_{m=1}^{k} \hat{\pi}_m \phi_{\theta_m}(\mathbf{y_i})}, \; i = 1, 2, \ldots, N, \; j = 1, \ldots, k. \tag{2.9}$$

3. **Maximization Step:** Compute the weighted means, variances, and mixing probability:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_j &= \frac{\sum_{i=1}^{N} \hat{\gamma}_{i,j} \mathbf{y_i}}{\sum_{i=1}^{N} \hat{\gamma}_{i,j}}, \\ \hat{\boldsymbol{\Sigma}}_j &= \frac{1}{N} \sum_{i=1}^{N} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})^T, \\ \hat{\pi}_j &= \frac{1}{N} \sum_{i=1}^{N} \hat{\gamma}_{i,j}. \end{aligned} \tag{2.10}$$

   For $j = 1, \ldots, k$.

4. Iterate steps 2 and 3 until convergence.
