Expectation Maximization applied to Gaussian Mixture Models

András Attila Sulyok

Pázmány Péter Catholic University
Faculty of Information Technology and Bionics

Gaussian Mixture Model

Latent variables: $z \sim \text{Categorical}(\phi)$, where
$\phi = (\phi_1, \ldots, \phi_K) \in \mathbb{R}^K$
(K clusters)

Visible variables: $x \mid z \sim \mathcal{N}(\mu_z, \Sigma_z)$
$x \in \mathbb{R}^d$
For one sample:
$$P(x \mid z) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_z|}} \exp\left[-\frac{1}{2}(x - \mu_z)^T \Sigma_z^{-1} (x - \mu_z)\right]$$

Parameters: $\theta = (\mu, \Sigma, \phi)$

Independence assumption: the data points are independent from each other

Expectation Maximization

In each iteration:

$$\begin{aligned}\theta^{(t+1)} &= \arg\max_{\theta} \mathbb{E}_{Z|X,\theta^{(t)}} \log P(X, Z \mid \theta) \\ &= \arg\max_{\theta} \sum_{Z} P(Z \mid X, \theta^{(t)}) \log P(X, Z \mid \theta) \\ &=: \arg\max_{\theta} E(\theta)\end{aligned}$$

- E-step: calculate the coefficients
- M-step: maximize

Elementwise formula

$$
\begin{aligned}
E(\theta) &= \mathbb{E}_{Z|X,\theta^{(t)}} \log P(X, Z \mid \theta) \\
&\quad \text{independence assumption:} \\
&= \sum_{i=1}^{N} \mathbb{E}_{z_i|x_i,\theta^{(t)}} \log P(x_i, z_i \mid \theta) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{K} P(z_i = j \mid x_i, \theta^{(t)}) \log P(x_i, z_i = j \mid \theta) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{K} \alpha_{ij} \log P(x_i, z_i = j \mid \theta)
\end{aligned}
$$

α_{ij}: responsibilities

Responsibilities

Using Bayes' Theorem:

$$\alpha_{ij} = \frac{P(x_i \mid z_i = j, \theta^{(t)}) P(z_i = j \mid \theta^{(t)})}{\sum_{l=1}^{K} P(x_i \mid z_i = l, \theta^{(t)}) P(z_i = l \mid \theta^{(t)})}$$

$$\propto P(x_i \mid z_i = j, \theta^{(t)}) P(z_i = j \mid \theta^{(t)})$$

$$\boxed{\alpha_{ij} \propto \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_j^{(t)}\right|}} \exp\left[-\frac{1}{2}\left(x_i - \mu_j^{(t)}\right)^T \left(\Sigma_j^{(t)}\right)^{-1} \left(x_i - \mu_j^{(t)}\right)\right] \phi_j}$$

You might need the logsumexp trick to avoid overflowing.

Optimization objective again

$$\arg\max_{\theta} E(\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{K} \alpha_{ij} \log P(x_i, z_i = j \mid \theta)$$

$$\begin{aligned} E(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} \alpha_{ij} \bigg[& -\frac{1}{2} \left(d \log(2\pi) + \log |\Sigma_j| \right) \\ & -\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \\ & + \log \phi_j \bigg] \end{aligned}$$

Maximization of a continuous function: set the gradient to 0.

Gradient w.r.t. μ_j

$$E(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} \alpha_{ij} \Big[-\frac{1}{2}\left(d \log(2\pi) + \log|\Sigma_j|\right) - \frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j) + \log \phi_j \Big]$$

$$\frac{\partial}{\partial \mu_j} E(\theta) = \sum_i \alpha_{ij} \Sigma_j^{-1}(x_i - \mu_j) = \Sigma_j^{-1} \sum_i \alpha_{ij}(x_i - \mu_j) = 0$$

$$\sum_i \alpha_{ij} x_i = \sum_i \alpha_{ij} \mu_j \implies \boxed{\mu_j = \frac{\sum_{i=1}^{N} \alpha_{ij} x_i}{\sum_i \alpha_{ij}}}$$

Gradient w.r.t. Σ_j

$$E(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} \alpha_{ij} \Big[-\frac{1}{2}\left(d \log(2\pi) + \log|\Sigma_j|\right) - \frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) + \log \phi_j \Big]$$

Good to know: $(|A|)' = |A|(A^{-1})^T$

$$\frac{\partial}{\partial \Sigma_j} E(\theta) = \sum_i \alpha_{ij} \Big[-\frac{1}{2} \cdot \frac{1}{|\Sigma_j|} \cdot |\Sigma_j| \Sigma_j^{-1} + \frac{1}{2} \Sigma_j^{-1} (x_i - \mu_j)(x_i - \mu_j)^T \Sigma_j^{-1} \Big] = 0$$

Gradient w.r.t. Σ_j

$$\sum_i \alpha_{ij}\left[-\frac{1}{2}\cdot\frac{1}{|\Sigma_j|}\cdot|\Sigma_j|\Sigma_j^{-1}+\frac{1}{2}\Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)^T\Sigma_j^{-1}\right]=0$$

$$\sum_i \alpha_{ij} = \sum_i \alpha_{ij}\Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)^T$$

$$\boxed{\Sigma_j = \frac{\sum_{i=1}^N \alpha_{ij}(x_i-\mu_j)(x_i-\mu_j)^T}{\sum_{i=1}^N \alpha_{ij}}}$$

Gradient w.r.t. ϕ_j

$$E(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} \alpha_{ij} \Big[-\frac{1}{2} \left(d \log(2\pi) + \log |\Sigma_j| \right) - \frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) + \log \phi_j \Big]$$

$$\frac{\partial}{\partial \phi_j} E(\theta) = \sum_i \alpha_{ij} \frac{1}{\phi_j} = 0 \quad ??$$

Not quite: remember, ϕ_j must sum to 1.

Gradient w.r.t. $\widetilde{\phi}_j$

$$\arg\max_{\phi} E(\theta) = \arg\max_{\phi} \sum_j \sum_i \alpha_{ij} \log \phi_i$$

s.t. $\sum_j \phi_j = 1$

Let $\sum_i \alpha_{ij} = \alpha_{.j}$ and $\widetilde{\phi} \in \mathbb{R}^K$ with $\widetilde{\phi}_j > 0 \; \forall j$

with this, the optimization above is equivalent to the unconstrained optimization:

$$\arg\max_{\widetilde{\phi}} \sum_j \alpha_{.j} \log \frac{\widetilde{\phi}_j}{\sum_l \widetilde{\phi}_l} = \sum_j \alpha_{.j} \left[\log \widetilde{\phi}_j - \log \sum_l \widetilde{\phi}_l \right]$$

Taking the gradient:

$$\frac{\partial}{\partial \widetilde{\phi}_k} (\dots) = \alpha_{.k} \frac{1}{\widetilde{\phi}_k} - \sum_j \alpha_{.j} \frac{1}{\sum_l \widetilde{\phi}_l} = 0$$

Gradient w.r.t. $\widetilde{\phi}_j$

$$\frac{d}{d\widetilde{\phi}_k}(\dots) = \alpha_{\cdot k}\frac{1}{\widetilde{\phi}_k} - \sum_j \alpha_{\cdot j}\frac{1}{\sum_l \widetilde{\phi}_l} = 0$$

$$\frac{\alpha_{\cdot k}}{\sum_j \alpha_{\cdot j}} = \frac{\widetilde{\phi}_k}{\sum_l \widetilde{\phi}_l} = \phi_k$$

This means that for the normalised version:

$$\boxed{\phi_j = \frac{\alpha_{\cdot j}}{\sum_{l=1}^{K} \alpha_{\cdot l}} = \frac{1}{N}\sum_{i=1}^{N} \alpha_{ij}}$$