Learning the Mixture of Gaussians

Prithwijit Guha

prithwijit.guha@gmail.com

1 Learning Gaussian Mixture Models

The GMM $\mathcal{G}(t) = \{\mathcal{C}_i(t)\}_{i=1}^{m}$ is a finite set of clusters of size m, where a cluster at the t^{th} instant is given by,

$$\mathcal{C}_i(t) = \{\mu_i(t), \Sigma_i(t), \pi_i(t)\} \tag{1}$$

Where, $\mu_i(t)$, $\Sigma_i(t)$ and $\pi_i(t)$ are the respective mean vector, co-variance matrix and the mixing parameter of $\mathcal{C}_i(t)$ at the t^{th} instant.

1.1 Initialization

The GMM is initialized with a single Cluster $C_1(1) = \{X_1, \Sigma_{init}, 1.0\}$, where X_1 is the data vector at $t = 1$ and Σ_{init} is the initial co-variance matrix whose values are assigned from the domain knowledge.

1.2 Update

In this sub-section we deduce the equations for updating the GMM $\mathcal{G}(t-1)$ learned till the $(t-1)^{th}$ instant to $\mathcal{G}(t)$ with the current data vector X_t. We consider the data vector to be belonging to the cluster $\mathcal{C}_j(t-1)$, if $(X_t - \mu_j(t-1))^T \Sigma_j(t-1)^{-1}(X_t - \mu_j(t-1)) < n\lambda$, where λ is a user defined threshold and n is the dimension of the data vector ($X \in \mathcal{R}^n$). Now, we consider the following cases.

In the first case, we assume that $\exists j : X_t \in \mathcal{C}_j(t-1)$. Let, $N_i(t)$ be the number of data vectors that has been assigned to $\mathcal{C}_i(t)$ till the t^{th} instant. Thus, we have,

$$\begin{aligned}
\pi_i(t) &= \frac{N_i(t)}{t} \\
&= \frac{(t-1)\pi_i(t-1) + \delta(i-j)}{t}; X_t \in \mathcal{C}_j(t-1) \\
&= (1-\alpha_t)\pi_i(t-1) + \alpha_t\delta(i-j); \alpha_t = \frac{1}{t}
\end{aligned} \tag{2}$$

Now, we update the mean and co-variance in $\mathcal{C}_j(t-1)$ only. To update the mean, we proceed as follows.

$$
\begin{aligned}
\mu_j(t) &= \frac{1}{N_j(t)} \sum_{X \in \mathcal{C}_j(t)} X \\
&= \frac{N_j(t-1)\mu_j(t-1) + X_t}{t\pi_j(t)} \\
&= \frac{(t\pi_j(t) - 1)\mu_j(t-1) + X_t}{t\pi_j(t)} \\
&= (1 - \beta_j(t))\mu_j(t-1) + \beta_j(t)X_t, \beta_j(t) = \frac{\alpha_t}{\pi_j(t)} \qquad (3)
\end{aligned}
$$

Similarly, we can update the co-variance matrix. From definition, we can compute the co-variance matrix at the t^{th} instant as,

$$
\begin{aligned}
\Sigma_j(t) &= \frac{1}{N_j(t)} \sum_{X \in \mathcal{C}_j(t)} (X - \mu_j(t))(X - \mu_j(t))^T \\
&= \frac{1}{N_j(t)} \sum_{X \in \mathcal{C}_j(t)} XX^T - \mu_j(t)\mu_j(t)^T \qquad (4)
\end{aligned}
$$

By further manipulating equation 4, we have the following.

$$
\begin{aligned}
N_j(t)(\Sigma_j(t) + \mu_j(t)\mu_j(t)^T) &= \sum_{X \in \mathcal{C}_j(t)} XX^T + X_t X_t^T \\
&= N_j(t-1)(\Sigma_j(t-1) + \mu_j(t-1)\mu_j(t-1)^T) + X_t X_t^T \\
\Sigma_j(t) + \mu_j(t)\mu_j(t)^T &= \frac{N_j(t-1)}{N_j(t)}(\Sigma_j(t-1) + \mu_j(t-1)\mu_j(t-1)^T) + \frac{1}{N_j(t)} X_t X_t^T \\
&= (1 - \beta_j(t))(\Sigma_j(t-1) + \mu_j(t-1)\mu_j(t-1)^T) + \beta_j(t)X_t X_t^T \qquad (5)
\end{aligned}
$$

Now, by substituting the update rule for $\mu_j(t)$ in the left hand side of 5, it can be shown that the updated co-variance matrix is given by,

$$
\Sigma_j(t) = (1 - \beta_j(t))[\Sigma_j(t-1) + \beta_j(t)\{X_t - \mu_j(t-1)\}\{X_t - \mu_j(t-1)\}^T] \qquad (6)
$$

In the second case, it may happen that $\forall j X_t \notin \mathcal{C}_j(t-1)$. In such cases, we intialize a new cluster $\mathcal{C}_k(t) = \{X_t, \Sigma_{init}, \alpha_t\}$. If $\mathcal{G}(t-1)$ contains less than m clusters, then we add $\mathcal{C}_k(t)$ to it. Otherwise, $\mathcal{C}_k(t)$ replaces the cluster with the lowest weight. More so, in this particualr case, the mixing parameters of all other clusters are penalized as $\pi_i(t) = (1 - \alpha_t)\pi_i(t-1), i \neq k$.