CSE 472
Assignment 4: Expectation-Maximization Algorithm for Gaussian Mixture Model

Kazi Ashik Islam
1205007

May 15, 2017

Why should we use a Gaussian mixture model (GMM) in the given scenario?

The reasons are:

- A ship can only be at a single place at a certain time. So probability of the ships being in that place should be maximum. Probability of being in any other place should decrease as we go away from this position. This scenario is similar to a gaussian distribution.
- In the given scenario we are unable to pinpoint the pirate ships exactly, because of signal interference. We may assume that this interference is similar to gaussian noise. Then we can use gaussian distribution for each of the ships.
- As we may have multiple ships, so we use a gaussian mixture model.

How will you model your data for GMM?

x_i = location data (2 dimensional co-ordinate)
X = $(x_1\ x_2\ x_3\ ...\ x_N)^T$ (Complete location data)
N = number of location data
K = number of ships

Position of each ship is fitted to a gaussian distribution. For distribution of k^{th} ship:
mean position = μ_k
Covariance matrix = Σ_k

For the mixture model:
θ_k = prior probability of picking k^{th} ship

Derive the loglikelihood function

Gaussian mixture distribution can be written in the form:

$$
\begin{aligned}
p(x_i|\mu, \Sigma, \theta) &= \sum_{k=1}^{K} \theta_k N(x_i|\mu_k, \Sigma_k) \\
\implies p(X|\mu, \Sigma, \theta) &= \prod_{i=1}^{N} \sum_{k=1}^{K} \theta_k N(x_i|\mu_k, \Sigma_k) \\
\implies ln(p(X|\mu, \Sigma, \theta)) &= \sum_{i=1}^{N} ln(\sum_{k=1}^{K} \theta_k N(x_i|\mu_k, \Sigma_k))
\end{aligned}
$$

Derive the update equations in M step

Let us introduce a K-dimensional binary random variable z having a 1-of-K representation in which a particular element z_k is equal to 1 and all other elements are equal to zero. The values of z_k therefore satisfy $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$. Then:

$$
\begin{aligned}
p(z_k = 1) &= \theta_k \\
\implies p(Z) &= \prod_{k=1}^{K} \theta_k^{z_k}
\end{aligned}
$$

$$
\begin{aligned}
p(x_i|z_k = 1) &= N(x_i|\mu_k, \Sigma_k) \\
\implies p(x_i|Z) &= \prod_{k=1}^{K} N(x_i|\mu_k, \Sigma_k)^{z_k}
\end{aligned}
$$

Now likelihood for complete dataset:

$$
\begin{aligned}
p(X, Z|\mu, \Sigma, \Theta) &= p(X|Z, \mu, \Sigma, \Theta)p(Z|\mu, \Sigma, \Theta) \\
&= \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{z_{ik}} (N(x_i|\mu_k, \Sigma_k))^{z_{ik}}
\end{aligned}
$$

So, loglikelihood of complete dataset:

$$
\begin{aligned}
l_c &= ln(p(X, Z|\mu, \Sigma, \Theta)) \\
&= ln(\prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{z_{ik}} (N(x_i|\mu_k, \Sigma_k))^{z_{ik}}) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik}[ln(\theta_k) + ln(N(x_i|\mu_k, \Sigma_k))]
\end{aligned}
$$

Expectation of complete loglikelihood:

$$
\langle l_c \rangle = \sum_{i=1}^{N} \sum_{k=1}^{K} \langle z_{ik} \rangle [ln(\theta_k) + ln(N(x_i|\mu_k, \Sigma_k))] \tag{1}
$$

Updating μ_k

Differentiating (1) with respect to μ_k:

$$\frac{\partial \langle l_c \rangle}{\partial \mu_k} = \sum_{i=1}^{N} \langle z_{ik} \rangle [\frac{\partial}{\partial \mu_k} ln(N(x_i|\mu_k, \Sigma_k))]$$

Now,

$$\begin{aligned}
\frac{\partial}{\partial \mu_k} ln(N(x_i|\mu_k, \Sigma_k))] &= \frac{\partial}{\partial \mu_k} ln[\frac{1}{(2\pi)^{d/2}|\Sigma_k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} x_i - \mu_k\}] \\
&= -\frac{1}{2}\frac{\partial}{\partial \mu_k}[(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)] \\
&= \frac{1}{2}(x_i - \mu_k)^T (\Sigma_k^{-1} + (\Sigma_k^{-1})^T)[as \frac{\partial}{\partial X} X^T A X = X^T(A + A^T)] \\
&= (x_i - \mu_k)^T \Sigma_k^{-1}
\end{aligned}$$

So,

$$\frac{\partial \langle l_c \rangle}{\partial \mu_k} = \sum_{i=1}^{N} \langle z_{ik} \rangle (x_i - \mu_k)^T \Sigma_k^{-1}$$

Setting this value to zero:

$$\begin{aligned}
&\sum_{i=1}^{N} \langle z_{ik} \rangle (x_i - \mu_k)^T \Sigma_k^{-1} = 0 \\
&\Sigma_k^{-1} \sum_{i=1}^{N} \langle z_{ik} \rangle (x_i - \mu_k)^T = 0 \\
&\sum_{i=1}^{N} \langle z_{ik} \rangle (x_i - \mu_k)^T = 0 \\
&\mu_k = \frac{\sum_{i=1}^{N} \langle z_{ik} \rangle x_i}{\sum_{i=1}^{N} \langle z_{ik} \rangle}
\end{aligned}$$

Updating Σ_k

Differentiating (1) with respect to Σ_k^{-1}:

$$\frac{\partial l_c}{\partial \Sigma_k^{-1}} = \sum_{i=1}^{N} \langle z_{ik} \rangle [\frac{\partial}{\partial \Sigma_k^{-1}} ln(N(x_i|\mu_k, \Sigma_k))]$$

Now,

$$\begin{aligned}
\frac{\partial l_c}{\partial \Sigma_k^{-1}} ln(N(x_i|\mu_k, \Sigma_k))] &= \frac{\partial}{\partial \Sigma_k^{-1}} ln[\frac{1}{(2\pi)^{d/2}|\Sigma_k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\}] \\
&= -\frac{1}{2}\frac{\partial}{\partial \mu_k}[(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)] - \frac{1}{2}\frac{\partial}{\partial \Sigma_k^{-1}}[ln(|\Sigma_k|)] \\
&= -\frac{1}{2}(x_i - \mu_k)(x_i - \mu_k)^T + \frac{1}{2}\frac{\partial}{\partial \Sigma_k^{-1}} ln(|\Sigma_k^{-1}|)[as \frac{\partial}{\partial A} X^T A X = X X^T]] \\
&= -\frac{1}{2}(x_i - \mu_k)(x_i - \mu_k)^T + \frac{1}{2}(\Sigma_k)^T [as \frac{\partial}{\partial X} log|X| = (X^{-1})^T] \\
&= -\frac{1}{2}(x_i - \mu_k)(x_i - \mu_k)^T + \frac{1}{2}\Sigma_k
\end{aligned}$$

So,

$$\frac{\partial l_c}{\partial \Sigma_k^{-1}} = \sum_{i=1}^{N} \langle z_{ik} \rangle [-\frac{1}{2}(x_i - \mu_k)(x_i - \mu_k)^T + \frac{1}{2}\Sigma_k]$$

Setting this value to zero:

$$\sum_{i=1}^{N} \langle z_{ik} \rangle [-\frac{1}{2}(x_i - \mu_k)(x_i - \mu_k)^T + \frac{1}{2}\Sigma_k] = 0$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} \langle z_{ik} \rangle (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} \langle z_{ik} \rangle}$$

Updating θ_k

To maximize the loglikelihood with respect to θ_k in (1), we have to maintain the constraint $\sum_{k=1}^{K} \theta_k = 1$. To enforce this constraint, we use the lagrange multiplier λ and augment (1) as follows:

$$L_c = l_c - \lambda(\sum_{k=1}^{K} \theta_k - 1) \tag{2}$$

Differentiating (2) with respect to θ_k:

$$\frac{\partial L_c}{\partial \theta_k} = \sum_{i=1}^{N} \langle z_{ik} \rangle \frac{1}{\theta_k} - \lambda$$

Setting this value to zero:

$$\sum_{i=1}^{N} \langle z_{ik} \rangle \frac{1}{\theta_k} - \lambda = 0$$

$$\sum_{i=1}^{N} \langle z_{ik} \rangle - \lambda\theta_k = 0$$

$$\sum_{i=1}^{N} \langle z_{ik} \rangle = \lambda\theta_k \tag{3}$$

Taking sum over all k:

$$\sum_{k=1}^{K} \sum_{i=1}^{N} \langle z_{ik} \rangle = \lambda \sum_{k=1}^{K} \theta_k$$

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \langle z_{ik} \rangle = \lambda$$

$$\sum_{i=1}^{N} = \lambda$$

$$\lambda = N$$

Substituting λ in (3):

$$\theta_k = \frac{\sum_{i=1}^{N} \langle z_{ik} \rangle}{N}$$