Ananya Bhattacharjee
1305046

May 18, 2018

1. Why should you use a Gaussian mixture model (GMM) in the above scenario?

   Answer: We used gaussian distributions to generate points and we can see clusters in the plot. We can see that the points are very dense to some particular point. The density is decreasing with the increasing of Euclidean distance. These patterns are generated from some gaussian distributions. Hence, we use GMM to fit the data.

2. How will you model your data for GMM?

   Answer: GMM is defined by three parameters, i.e., mean (μ), covariance (Σ), and mixture coefficients (w). In the given scenario, the mean is the actual location of the ship. The reflection of sound from from ship generates other points. Each point is generated from a ship's body with the mixture probability. Therefore, we can model our data by following algorithm:

   - for $i = 1$ to N do
     $m \leftarrow$ index of one of the M models randomly selected according to mixing coefficients w.
     Randomly generate x_i according to the distribution $N(x_i|\mu_k, \Sigma_k)$
   - end

3. What are the intuitive meaning of the update equations in M step?

   We do not know the model parameters of the underlying distribution. So, we initialize them with some random values. We can then calculate the value of the likelihood function. In the E-step, we use our model equations to solve for the distribution of the missing data given our current guess of the model parameters and given the observed data. In the M-step, we use the probabilities that maximizes the log-likelihood to get a new estimate of our model parameters.

4. Derive the log-likelihood function in step 4.

Answer: We want to maximize

l $= \text{P(X)}$

All the generated data are independent of each other (identical and independent distribution). So

l $= \sum_{j=1}^{N} P(x_j)$

log l $= \sum_{j=1}^{N} \log P(x_j)$
$= \sum_{j=1}^{N} \log N(x_j|\mu, \sum)$

For k variables,

log l $= \sum_{j=1}^{N} \log[\sum_{i=1}^{K} N_i(x_j|\mu_i, \sum_i)w_i]$