# Assignment 4: Expectation-Maximization Algorithm for Gaussian Mixture Model

April 27, 2017

Description of the Assignment

Let's assume you have been hired by Bangladesh Navy as a data analyst for their newly acquired Ming-class submarines. They are planning to use these submarines to secure national maritime border. One of their potential targets is pirate ship. They use active acoustic sonar to find floating ships. By transmitting a sound pulse and receiving the echo on an array, the sonars can determine the direction of the echoes that return from objects hit by the sound. They can also measure the time it takes for echoes to return and calculate the distance to the object causing the echo. Unfortunately, if there are several ships nearby, the signals cause interference. Using periscope visuals, it is possible to determine the number of ships but exact location of each ship is difficult to locate. Without precise location, a torpedo hit is not guaranteed. The 2D plot of a sample sonar data is given in Figure 1. There are three ships in surface water, one inside each black circle. Each blue dot represents a single location measurement of one of the three ships. You have been instructed to build a system which takes the sonar data as well as the number of ships as input and provide the estimated location of each ship.
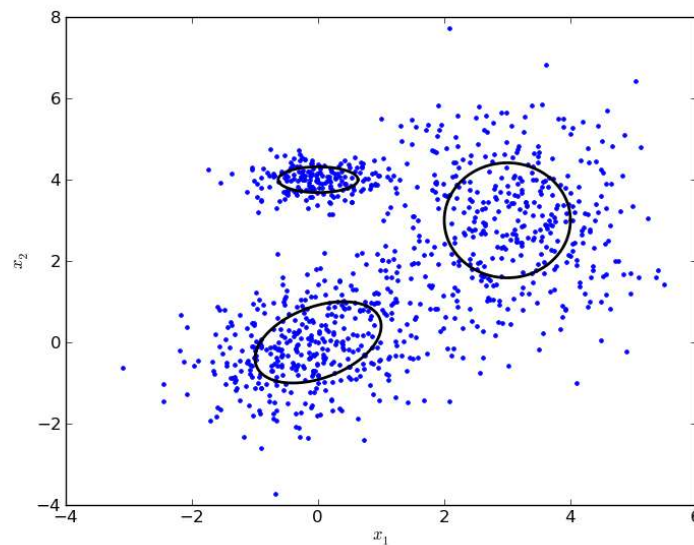
Since you are a new recruit, you are not yet well versed in applying machine learning model to real world problem. So, you plan to consult Dr. Google who is an expert of virtually every field. He informs you that you should use Gaussian mixture model in the above mentioned case. One possible implementation for this model is an EM (expectation-maximization) algorithm. He is also generous enough to give you a pseudocode for this which is given below.

Let a vector $\mathbf{x}$ with dimension D can be generated from any one of the K Gaussian distribution where the probability of selection of Gaussian distribution j is θ_j where $\sum_{k=1}^{K} \theta_k = 1$ and the probability of generation of $\mathbf{x}$ from Gaussian distribution j is given as

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_j|}} e^{\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right)}$$

To learn a Gaussian mixture model using EM algorithm, we need to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients). The steps are given below.

1. Initialize the means $\boldsymbol{\mu}_j$, covariances $\boldsymbol{\Sigma}_j$ and mixing coefficients θ_j, and evaluate the initial value of the log likelihood.

2. **E step**: Evaluate the conditional distribution of latent factors using the current parameter values

$$\pi_{ij} = p(z_i = j|\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}) = \frac{\theta_j \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{k=1}^{K} \theta_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

3. **M step**: Re-estimate the parameters using the conditional distribution of latent factors

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \pi_{ik} \mathbf{x}_i}{\sum_{i=1}^{N} \pi_{ik}}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} \pi_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} \pi_{ik}}$$

$$\theta_k = \frac{\sum_{i=1}^{N} \pi_{ik}}{N}$$

4. Evaluate the log likelihood and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} \theta_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

Now you are excited to get this detailed model and algorithm. However, since this is an issue of national security you want to make sure that the suggestion from Dr. Google is indeed correct. At the outset you want to answer the following questions in your report.

1. Why should you use a Gaussian mixture model (GMM) in the above scenario?
2. How will you model your data for GMM?
3. Derive the update equations in **M step**. (To make the derivations short you can use formulas from matrix calculus)
4. Derive the log-likelihood function in step 4.
5. Implement the above pseudocode and estimate the location of enemy ships.

Special Instructions

1. This time the report with the answer of first 4 questions is mandatory and will be graded. You have to explain those answers during viva. The report should be in .docx/.pdf and must be available at the time of viva. Write precisely in your own language and keep it as simple as possible. Show detailed derivations for questions 3 and 4.
2. Although you are allowed to consult Dr. Google, don't copy anything! If you do copy from internet or from any other person or from any other source, you will be severely punished and it is obvious. More than that, we expect fairness and honesty from you. Don't disappoint us!
3. Upload the code and report in Moodle within 2:00 P.M. of 8th May, 2017 (Monday). This is a strict system-imposed deadline for both section A and B.
4. For Python and Matlab, you may not get supporting software in the lab. If you do program in these languages, bring your computer in the sessional.
5. You are allowed to show the assignment in your own laptop during the final submission. But in that case, ensure an internet connection as you have to instantly download your code from the Moodle and show it.
6. Contact Sharif sir for any typos or discrepancies in this document.

Instructions for Moodle upload

1. Upload the assignment within the specified time. Otherwise, we can't accept it.
2. If you write code in a single file, then rename it as <Student id> <code>.<extension>. For example, if your student id is 1105123 and you have done in java, then your file name should be "1105123 code.java".
3. The report name should be <Student id> <report>.<extension>. For example, if your student id is 1105123 and it is in pdf format, then the report name should be "1105123 report.pdf".
4. Finally make a main folder, put the code and report in it, and rename the main folder as your <Student id> <Programming language>. For example, "1105123 Java". Then zip it and upload it.