Machine Learning Assignment 3

Liang Shuailong 1000829

Task1: Mixture of Gaussians & EM

1. The model parameters to be estimated is $\theta = \{\mu^{(1)}, \mu^{(2)}, \sigma_1^2, \sigma_2^2, p_1, p_2\}$, which specify the means, variances and weights of two Gaussian distributions. p_1, p_2 are hidden variables for the mixture model.

2. To use EM to solve the Gaussian Mixture Model problem, the $\mu^{(1)}, \mu^{(2)}$ are initialized as in the k-means algorithm. For this problem, random initialization is used and in order to make the result make sense, the random data point is guaranteed to lie with the bounding box of the samples. The variances σ_1^2, σ_2^2 are set all equal to the overall data variances, and weights p_1, p_2 both equal to 0.5.

   The log joint likelihood associated with the data increases as the algorithm iterates, illustrated Figure 1.

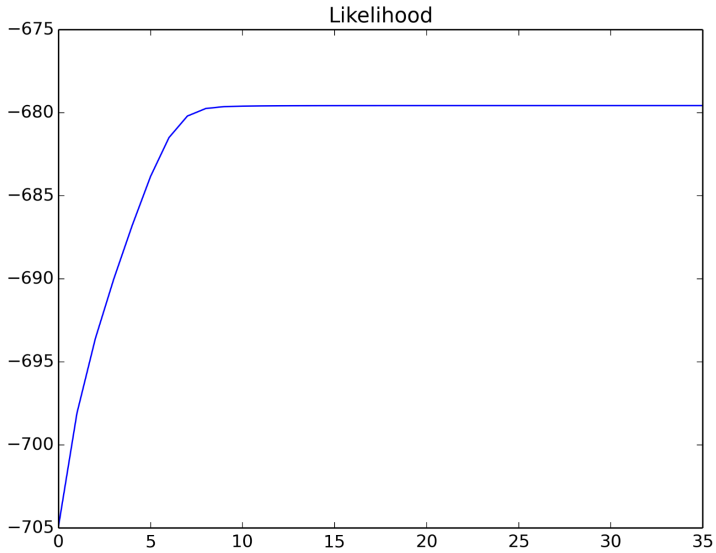

Figure 1 Soft EM: Likelihood vs. Number of iteration

As is analyzed in the class, the likelihood is guaranteed not to decrease after each iteration. After a while, the likelihood reaches its maximum, and the algorithm converges. The terminate condition is that the likelihood does not increase by a very tiny value(threshold).

After the convergence of the algorithm, the clustering result is shown in Figure 2.

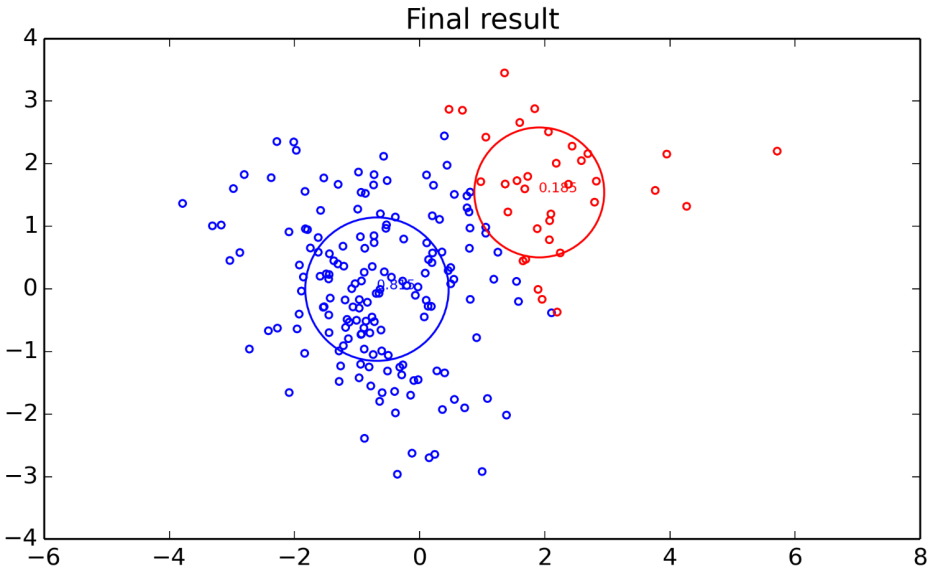

Figure 2 The membership after convergence of the soft EM algorithm

The color indicates the membership of the data point. The center and radius of the circle is specified by μ and σ, respectively, and the number near the center of the circle is the weight of the Gaussian distribution.

For the visualization of the evolving of the EM algorithm, you can open README.html in a browser (Safari preferred) to watch the GIF animation. The GIF image will be updated each time the algorithm is run if the corresponding option is set.

3. If hard EM is used, the algorithm will not change too much except that in E step, we no longer set $p(i|t)$ as the probability, but set it to 1 or 0 according to the probability.

   The log joint likelihood associated with the data increases as the algorithm iterates, illustrated Figure 3.

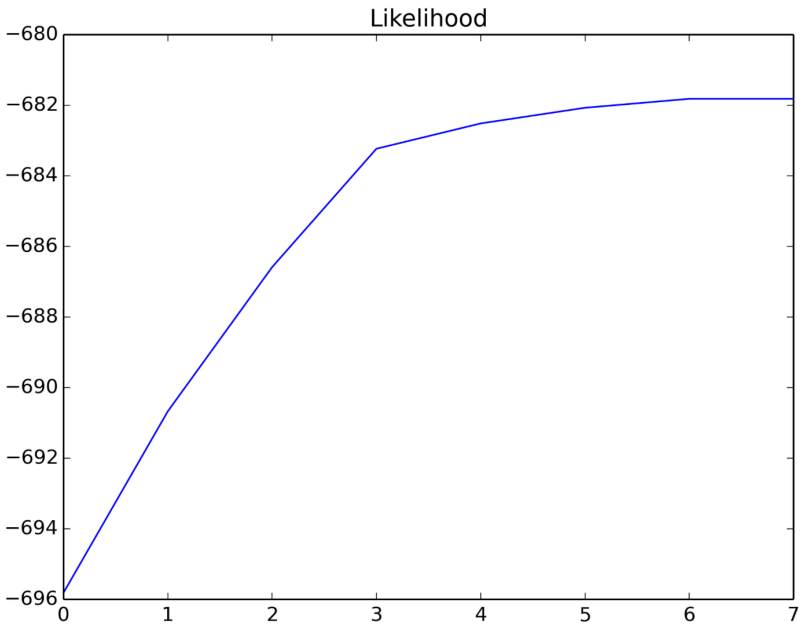

Figure 3 Hard EM: Likelihood vs. Number of iteration

As shown above, hard EM converges very quickly in a small number of iterations. However, it is not quite stable and converge to suboptimal more often than soft EM. The clustering result of hard EM is shown in Figure 4.

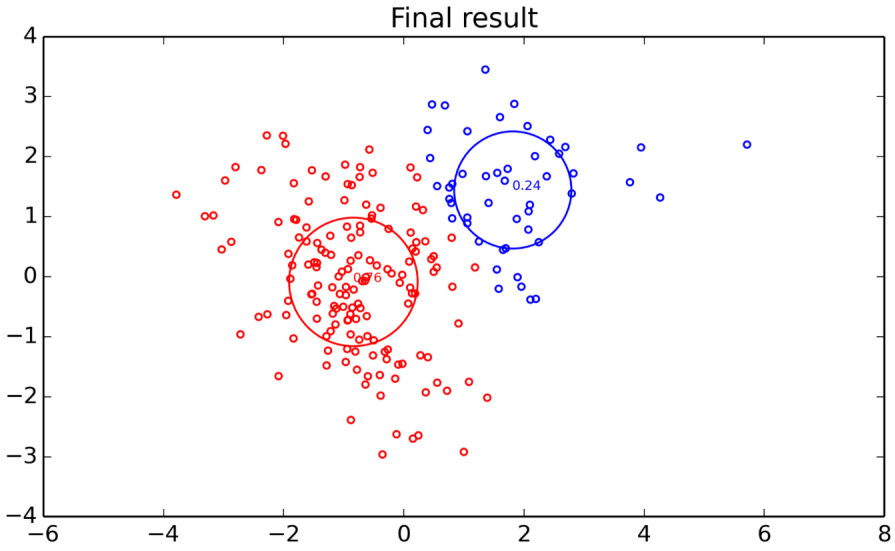

Figure 4 The membership after convergence of the hard EM algorithm

Task2: Theorem Proofing

Proof:

We prove from right to left. If we expand all the quadratic terms, and observe the coefficient of $\|v_i\|^2$ without considering cross terms, it is easy to find that the coefficient is

$$\frac{\lambda_i(\lambda_0 + \cdots + \lambda_{i-1} + \lambda_{i+1} + \cdots + \lambda_N)}{\lambda_0 + \cdots + \lambda_N}\|v_i\|^2 \quad (1)$$

for each $i = 0, 1, \ldots, N$.
If compared with left hand side of the equation, we can see that we lack

$$\frac{\lambda_i^2}{\lambda_0 + \cdots + \lambda_N}\|v_i\|^2 \quad (2)$$

for each $i = 0, 1, \ldots, N$.

Then we will prove that the cross terms are exactly (2).

The sum of all the cross terms of $\sum_{i=1}^{N} \eta_{0,i}\|v_i + v_0\|^2$ is

$$\frac{2\lambda_0^2\|v_0\|^2}{\lambda_0 + \cdots + \lambda_N} = \frac{\lambda_0^2\|v_0\|^2}{\lambda_0 + \cdots + \lambda_N} + \frac{(\|\lambda_1 v_1 + \cdots + \lambda_N v_N\|)^2}{\lambda_0 + \cdots + \lambda_N} \quad (3)$$

The sum of all the cross terms of $\sum_{1 \le j < k \le N} \eta_{j,k}\|v_j - v_k\|^2$ is

$$-2\sum_{1 \le j < k \le N} \frac{(\lambda_j v_j)(\lambda_k v_k)}{\lambda_0 + \cdots + \lambda_N} \quad (4)$$

(3) - (4), we can get exactly (2).
So the equation is proved.