ML Report

Expectation Maximization Algorithm

Course No: CSE 472

Course Name: Machine Learning Sessional

Name: Md. Saqib Hasan

Student ID: 1305057

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

Answer: We can use Gaussian mixture model (GMM) in the above scenario because the above scenario consists of a number of data points in 2 dimensional space which form clusters around certain fixed points (mean). So basically, they form clusters of continuous data which are Gaussian distribution with different means and standard deviations.

2) How will you model your data for GMM?

Answer: I will model my data using GMM as a list of (x,y) variables where x and y represent the two features of each sample of data. Then I will select k, the number of the Gaussian distributions that the data points are coming from. The EM algorithm iteratively predicts the probability of which data point comes from which distribution as well the mean and covariance of the data clusters. As such, we need to initialize the means and covariance as well for the k distributions.

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

Answer: In the EM algorithm, our first task has been to initialize the mean, covariance and the weights with some random initialization values. This is the first step of our algorithm. Using these random values, we can move our algorithm. In the E-step of the algorithm, we use these mean, covariance and weight vectors to calculate the probability that a certain sample or data point is coming from which Gaussian distribution. Thus, we have used the model parameters to predict and update the possibility of a point coming from one of the k distributions. Therefore, in the M-step, we use the probabilities that gives the maximum likelihood that jth data point is coming from the ith distribution to update the parameters mean, covariance and weights for the next iteration.

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

Answer:

The derivation of the log-likelihood function of step-4 is given below:

~~ln p~~

$$\ln p(X \mid \mu, \Sigma, W) = \sum_{j=1}^{N} \ln \{ p(x_j \mid \mu, \Sigma, W)$$

$$= \sum_{j=1}^{N} \ln \left(\sum_{l=1}^{K} w_i N_i (x_j \mid \mu_i \Sigma_i) \right)$$

The above is the function we wish to derive.

Now, first we will try to maximize the function

$$l = P(X)$$

We know that the data points from the distributions that we are using in our algorithm are independent of each other. The data points are generated under this assumption.

Therefore, we can write that:

$$l = P(x_1) \times P(x_2) \times P(x_3) \dots \times P(x_N)$$

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

Taking the Logarithm,

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

$$= \sum_{j=1}^{N} \ln N(x_j | \mu, \Sigma)$$

Considering for the k distributions and their weights, we have:

$$L = \sum_{j=1}^{N} \ln \left\{ \sum_{i=1}^{k} N_i (x_j | \mu_i, \Sigma_i) w_i \right\}$$

which is the required function we needed to desrive