Week 6 : Session 1

9. EM Algorithm

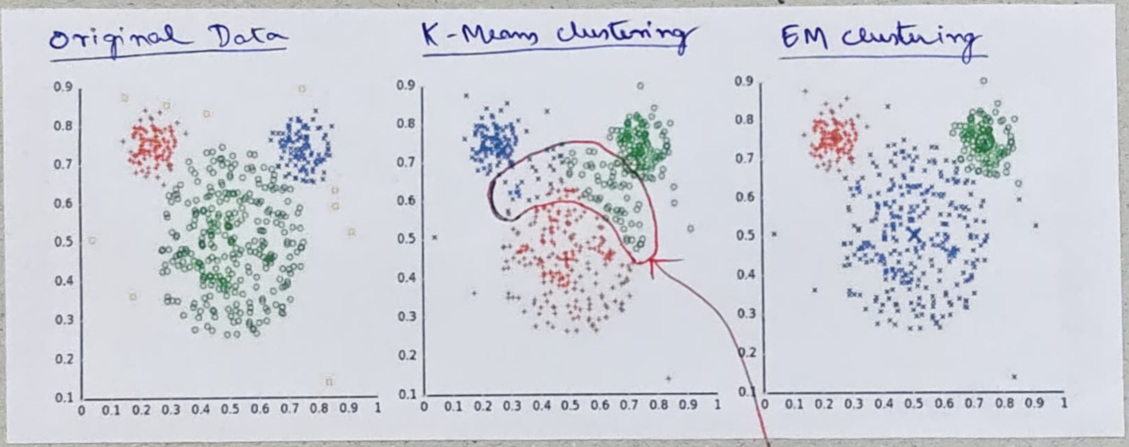
EM stands for Expectation Maximization. This can be used for Probabilistic clustering or Soft-clustering.

Previously we've seen the K-Means algorithm, which is used for Hard-clustering.

What is Probabilistic clustering ?

Previously, we assigned each point to a Unique cluster. But the problem is, there might be some points which cannot be assigned to any particular cluster, because those points seemingly belong to either clusters.

In Probabilistic clustering, we calculate the probability that a data point belongs to a cluster! This is why, we also call this as Soft clustering.



Plot: Different Cluster Analysis results on "mouse" data set.

In K-Means clustering, these data points are misclassified.

The Probabilistic clustering is clearly able to recover the original clusters to a great degree. Therefore, EM algorithm can yield a much better clustering performance than the K-Means algorithm.

To understand EM algorithm, let us consider a Gaussian cluster model. (ie) we assume that the clusters are generated by Gaussian distributions. With probability p_i (prior probability of i^{th} cluster), we generate a sample $\bar{x}$ from Gaussian cluster i.

(ie) $\mathcal{N}(\bar{\mu}_i, \sigma^2 I)$

Covariance Matrix

Centroid cluster

$$\Rightarrow \underbrace{\mathcal{N}(\bar{\mu}_1, \sigma^2 I), \mathcal{N}(\bar{\mu}_2, \sigma^2 I), \ldots\ldots, \mathcal{N}(\bar{\mu}_K, \sigma^2 I)}_{\text{Assuming there are K clusters}}$$

The Probability Density Function (PDF) is given as

$$f_X(\bar{x}) = \sum_{i=1}^{K} P_i \times \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\|\bar{x} - \bar{\mu}_i\|^2}$$

where, $P_1, P_2, \ldots P_K$ are Prior Probabilities of clusters.

This is termed as Gaussian Mixture model, as there are K Gaussians, one for each cluster.

Consider now M data points.

$\bar{x}(1), \bar{x}(2), \ldots, \bar{x}(M)$.

How do we cluster this data?

(ie) We desire to estimate Centroids $\bar{\mu}_i$

As well as the cluster Assignments $\alpha_i(j)$

Here, performing direct ML estimation is mathematically intractable (NOT possible)

In this problem, there are 2 unknowns.

① Centroid $\bar{\mu}_i$

② Cluster assignment $\alpha_i(j)$

However, if cluster assignment $\alpha_i(j)$ is known, problem is simple. For example: we have $M = 8$ data points, out of which

$$\bar{x}(1), \bar{x}(3), \bar{x}(5), \bar{x}(8) \quad \in \quad \text{Cluster } 1$$

$$\bar{x}(2), \bar{x}(4), \bar{x}(6), \bar{x}(7) \quad \in \quad \text{cluster } 2$$

Then the centroids are

$$\hat{\mu}_1 = \frac{\bar{x}(1) + \bar{x}(3) + \bar{x}(5) + \bar{x}(8)}{4} \qquad \left(\text{Average of points in cluster } 1\right)$$

$$\hat{\mu}_2 = \frac{\bar{x}(2) + \bar{x}(4) + \bar{x}(6) + \bar{x}(7)}{4} \qquad \left(\text{Average of points in cluster } 2\right)$$

Recall, clustering problem is Unsupervised Learning (No Labels). (i.e) there is No cluster information.

Here, we introduce the concept of missing data or latent information:

The cluster assignment variable

$$\alpha_i(j) = \begin{cases} 1 & \text{if } \bar{x}(j) \in \text{Cluster } i \\ 0 & \text{if } \bar{x}(j) \notin \text{cluster } i \end{cases}$$

is missing here.

Complete Data

$$\bar{x}(1), \bar{x}(2), \ldots, \bar{x}(M)$$
$$\alpha_i(j), \quad i = 1, 2, \ldots, K, \quad j = 1, 2, \ldots, M$$

We have to estimate the Missing data as well as the Centroids. To do these, we have to first start with cost function, and that cost function comes from the likelihood.

The Likelihood of the complete data

$$\underbrace{\prod_{j=1}^{M} \prod_{i=1}^{K} \left(p_i \times \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\|\bar{x}(j)-\bar{\mu}_i\|^2} \right)^{\alpha_i(j)}}_{\text{This is the joint PDF of the points, considering the availability of the missing data, which can be treated as the Likelihood function.}}$$

Product

Missing Data

The Log-Likelihood of the complete Data

$$\sum_{j=1}^{M} \sum_{i=1}^{K} \alpha_i(j) \left(p_i - \frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|\bar{x}(j) - \bar{\mu}_i\|^2 \right)$$

Product becomes Sum for Log.

EM Algorithm   Week 6 : Session 2

The EM algorithm proceeds iteratively.

(ie) in each iteration, we perform

Expectation step $\rightarrow$ E-step

Maximization step $\rightarrow$ M-Step

Let us consider the centroids in $(l-1)$ iterations.

$$\mu_0^{(l-1)}, \mu_1^{(l-1)}, \ldots, \mu_K^{(l-1)}$$

At the end of $(l-1)^{th}$ iteration, we have these centroids, which are nothing but cluster means. Recall, we have Gaussian clusters (ie) Gaussian Mixture model. We are modeling the clusters as Gaussian. So, the centers are nothing but the mean of the different Gaussian components in the mixture.

The <u>Expected value of the Log likelihood</u> in iteration l is,

$$\underbrace{\sum_{j=1}^{M}\sum_{i=1}^{K} \alpha_i^{(l)}(j) \left\{ \ln p_i - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\bar{x}(j) - \bar{\mu}_i\|^2 \right\}}_{Q(\bar{\mu})}$$

$$E\{\alpha_i(j)\} = \alpha_i^{(l)}(j)$$

Expected value of $\alpha_i(j)$ in l^{th} iteration.

How to calculate $\alpha_i^{(l)}(j)$?

Recall, $\alpha_i^{(l)}(j) = 1$ if $\bar{x}(j) \in C_i$

$$= P_r(\bar{x}(j) \in C_i) \times 1 + P_r(\bar{x}(j) \notin C_i) \times 0$$

$$= P_r(\bar{x}(j) \in C_i)$$

$$= P_r(C_i \mid \bar{x}(j))$$

$$= \frac{P_r(\bar{x}(j) \mid C_i) \cdot P(C_i)}{\sum_{k=1}^{K} P_r(\bar{x}(j) \mid C_k)\, P(C_k)}$$ (Bayes Rule)

Prior probability

$$= \frac{p_i \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\|\bar{x}(j) - \bar{\mu}_i^{(l-1)}\|^2}}{\sum_{k=1}^{K} p_k \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\|\bar{x}(j) - \bar{\mu}_k^{(l-1)}\|^2}}$$

$$= E\{\alpha_i(j)\}$$

M-Step

The M-step is the Maximization step i.e. Maximize the Expected value of Likelihood. To determine $\bar{\mu}_j$, differentiate w.r.t $\bar{\mu}_j$ and set equal to zero.

(ie) To maximize, take the Gradient of the Expected value of the Log Likelihood and set equal to zero.

$$\nabla_{\bar{\mu}_i} Q(\bar{u}) = 0$$

Solve for $\bar{\mu}_i^{(l)}$, we get

Centroid of i^{th} cluster in iteration l.

$$\bar{\mu}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\, \bar{x}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}$$

Now, compare the expressions in K-Means and EM,

$$\text{EM}: \bar{\mu}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\, \bar{x}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}$$

Same Expression!

$$\text{K-Means}: \bar{\mu}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\, \bar{x}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}$$

Both are Identical!

Then what is the difference?

EM	K-Means
EM performs Weighted average	K-Means performs Hard average
$\alpha_i^{(l)}(j)$ – Soft clustering	$\alpha_i^{(l)}(j)$ – Hard clustering
$\alpha_i^{(l)}(j) \in [0, 1]$ (ie) can take any value in the interval $[0, 1]$ Example : 0.95	$\alpha_i^{(l)}(j) \in \{0, 1\}$ (ie) can only take 2 possible values

Therefore, in EM, $\alpha_i^{(l)}(j)$ denote probabilities.

Previously, we assumed the Prior probabilities to be known.
But here, we compute Prior probabilities.

The Prior probabilities p_i can be computed as follows.

$$p_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}{M}$$

Sum of all Aposteriori probabilities of points in cluster i

Estimate of Prior probability in iteration l