CSCI 567 Fall 2016

Homework 3

MohammedJunaid Hundekar
2548-4069-96
hundekar@usc.edu
Collaborators: Sagar Makhwana, Dhawal Shah, Ankit Kothari

1 CLUSTERING

a) Assuming all r_{nk} are known, given

$$D = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \, || \, x_n - \mu_k \, ||_2^2$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \, (x_n - \mu_k)^T \, (x_n - \mu_k)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \, (x_n^T \, x_n - x_n^T \, \mu_k - \mu_k^T \, x_n + \mu_k^T \, \mu_k)$$

Taking derivative w.r.t μ_k and setting it to zero

$$\frac{\partial D}{\partial \mu_k} = \sum_{n=1}^{N} r_{nk} \, (2 \, \mu_k - 2 \, x_n) = 0$$

$$\sum_{n=1}^{N} r_{nk} \, \mu_k = \sum_{n=1}^{N} r_{nk} \, x_n$$

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}}$$

From the above equation we can clearly observe that to minimize this loss/ cost function, μ has to be the mean of the respective clusters.

b) $$D = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \,||\, x_n - \mu_k \,||_1$$

We observe that Taking derivative w.r.t μ_k and setting it to zero

$$\frac{\partial D}{\partial \mu_k} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}\, sign(x_n - \mu_k) = 0$$

Assume the class size is m for class k:

$$\sum_{m=1}^{M} sign(x_m - \mu_k) = 0$$

$$sign(x_m - \mu_k) = +1 \; if \; x_m - \mu_k > 0$$

$$= -1 \; if \; x_m - \mu_k < 0$$

Therefore: count($x_n - \mu_k > 0$) - count($x_n - \mu_k < 0$) = 0 where count() gives the number of elements

This becomes zero precisely at median.

c) By applying a mapping of $\Phi(x)$ to map data points into feature space, the objective function of kernel K-means is given as

$$\widehat{D} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \,||\, \Phi(x_n) - \widehat{\mu}_k \,||_2^2$$

where

$$\widehat{\mu}_k = \frac{\sum_{i=1}^{N} r_{ik}\, \Phi(x_i)}{\sum_{i=1}^{N} r_{ik}}$$

$$||\, \Phi(x_n) - \widehat{\mu}_k \,||_2^2 = (\Phi(x_n) - \widehat{\mu}_k)^T (\Phi(x_n) - \widehat{\mu}_k)$$

$$= \Phi(x_n)^T \Phi(x_n) - 2\, \widehat{\mu}_k^T \Phi(x_n) + \widehat{\mu}_k^T \widehat{\mu}_k$$

$$= \Phi(x_n)^T \Phi(x_n) - 2 \frac{\sum_{i=1}^{N} r_{ik}\, \Phi(x_i)^T \Phi(x_n)}{\sum_{i=1}^{N} r_{ik}} + \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} r_{ik} r_{jk}\, \Phi(x_i)^T \Phi(x_j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} r_{ik} r_{jk}}$$

Define a variable $n_k = \sum_{i=1}^{N} r_{ik}$:

$$||\, \Phi(x_n) - \widehat{\mu}_k \,||_2^2 = \Phi(x_n)^T \Phi(x_n) - 2 \frac{\sum_{i=1}^{N} r_{ik}\, \Phi(x_i)^T \Phi(x_n)}{n_k} + \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} r_{ik} r_{jk}\, \Phi(x_i)^T \Phi(x_j)}{n_k^2}$$

$$= K(x_n, x_n) - 2 \frac{\sum_{i=1}^{N} r_{ik} K(x_i, x_n)}{n_k} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik} r_{jk} K(x_i, x_j)}{n_k^2}$$

Thus,

$$\widehat{D} = \sum_{n=1}^{N} K(x_n, x_n) - 2 \frac{\sum_{i=1}^{N} r_{ik} K(x_i, x_n)}{n_k} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik} r_{jk} K(x_i, x_j)}{n_k^2}$$

1. For given point x_n calculate $K(x_n, x_n) - 2 \frac{\sum_{i=1}^{N} r_{ik} K(x_i, x_n)}{n_k} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik} r_{jk} K(x_i, x_j)}{n_k^2}$ for all possible clusters k
2. Assign cluster to point x_n using:

$$r_{nk} = \begin{cases} 1 & k = arg\ min_k \| \Phi(x_n) \ 2\ \hat{\mu}_k \|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

where

$$K(x_n, x_n) - 2 \frac{\sum_{i=1}^{N} r_{ik} K(x_i, x_n)}{n_k} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik} r_{jk} K(x_i, x_j)}{n_k^2}$$

and $n_k = \sum_{i=1}^{N} r_{ik}$

Algorithm Kernel k means

1: **procedure** KERNEL K MEANS

2: mu[i] = **x**(random(1..N)) for 1 σ *i* σ *k*

initialise cluster centroids[1..k] randomly choosing any *k* points of N (Sample without replacement)

3: **for** i **do**:1 to N

4: **for** j **do**:1 to N

K[i,j] = $\Phi(x_i)\Phi(x_j)$

5: end **for**

6: end **for**

r(n,k) ← 0

Until convergence repeat

7: **for** i **do**:1 to N

8: j = *arg* $min_k \| \Phi(x_n) - \hat{\mu}_k \|_2^2$

Use the above formula to calculate distances

9: r[i,j] = 1

10: Update μ_j

Recalculate centroids of assigned cluster *j*

11: end **for**

12:end **procedure**

2. Gaussian Mixture Models.

From the given information we can write the likelihood function $L(\alpha) = p(x_1 \mid \alpha)$ as:

$$p(x_1 \mid \alpha) = \frac{\alpha}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} + \frac{1-\alpha}{\sqrt{\pi}} e^{-x^2}$$

Consider that a single sample x_1 has been observed. Determine the maximum likelihood estimate of α.
We can write the likelihood as follows:

$$p(x_1 \mid \alpha) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} - \frac{1}{\sqrt{\pi}} e^{-x_1^2} \right) \alpha + \frac{1}{\sqrt{\pi}} e^{-x_1^2}$$

Thus, we see that the likelihood is simply a linear function of alpha where the sign of the slope is determined by which the Gaussian produces the larger response.
Since we know that $0 \leq \alpha \leq 1$,
if the slope is positive that we should choose $\alpha = 1$
if the slope is negative we should use $\alpha = 0$.
Using straightforward algebra one can show that the slope is positive whenever $x_1^2 \geq \log 2$ and we should set $\alpha = 1$ otherwise set $\alpha = 0$.
Alternatively, one could also apply Expectation maximization for this problem (not an efficient solution).
Starting with $\alpha = 0.5$ and applying EM, you would observe that in each iteration, your estimate of α will strictly increase or decrease depends on which of the two Gaussians fit x_1 better, eventually lead to 1 or 0 accordingly.

3 EM algorithm

Given:

$$p(x_i) = \begin{cases} \pi + (1 \ \lambda \ \pi)\, e^{-\lambda} & x_i = 0 \\ (1 - \pi)\ \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} & x_i > 0 \end{cases}$$

We can rewrite the above expressions in terms of X as below

$$X_i = \begin{cases} 0 & \text{probability} = \pi + (1 \ \lambda \ \pi)\, e^{-\lambda} \\ x_i & \text{probability} = (1 \ \lambda \ \pi)\ \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \end{cases}$$

We define a *latent* variable Z_i for all cases where $X_i = 0$, as Z_i can be 0 or 1

As we cannot observe the above, X_i comes out of a mixture of a degenerate distribution as follows:

$$Z_i = \begin{cases} 1 & X_i \text{ is from the degenerate distribution} \\ 0 & \text{otherwise} \end{cases}$$

Thus

$p(X_i = 0,\ Z_i = 1) = p(Z_i = 1) \quad p(X_i = 0 | Z_i = 1) = \pi \quad 1$

$P(X_i = 0,\ Z_i = 0) = p(Z_i = 0) \quad p(X_i = 0 | Z_i = 0) = (1 - \pi)\, e^{-\lambda} \quad 1$

$$L((\pi, \lambda)|(X, Z)) = \prod_{x_i = 0} \pi^{z_i} \ ((1 - \pi)\, e^{-\lambda})^{1 - z_i} * \prod_{x_i > 0} (1 - \pi)\ e^{\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}}$$

$$\log \mathrm{L} = \sum_{I(x_i = 0)} z_i \log(\pi) + (1 - z_i)\ (\log(1 - \pi) - \lambda)$$

$$+ \sum_{I(x_i = 0)} \log(1 - \pi) + x_i \log(\lambda_i) - \lambda - \log(x_i!))$$

$\theta = (\pi, \lambda)$; θ_0 represents a known parameter as estimated from previous step

E step:

$$Q(\theta, \theta_0) = \sum_{I(x_i = 0)} E_{P(Z|X)}[z_i] \log(\pi) + (1 - E_{P(Z|X)}[z_i]\ \log(1 - \pi) - \lambda)$$

$$+ \sum_{I(x_i > 0)} \log(1 - \pi) + x_i \log(\lambda_i) - \lambda - \log(x_i!))$$

Substituting the values we get

$$E_{P(Z|X)}[z_i] = 0 * p(Z_i = 0|X) + 1 * p(Z_i = 1|X_i = 0)$$

$$= \frac{p(X_i = 0|Z_i = 1)\, p(Z_i = 1)}{p(X_i = 0|Z_i = 0)\, p(Z_i = 0) + p(X_i = 0|Z_i = 1)\, p(Z_i = 1)}$$

$$= \frac{\pi_0}{\pi_0 + (1 - \pi_0) e^{-\lambda_0}}$$

Substituting in $Q(\theta, \theta_0)$

$$Q(\theta, \theta_0) = \sum_{I(x_i = 0)} \frac{\pi_0}{\pi_0 + (1-\pi_0)e^{-\lambda_0}} \log(\pi) + \left(\frac{(1-\pi_0)e^{-\lambda_0}}{\pi_0 + (1-\pi_0)e^{-\lambda_0}}\right)(\log(1 - \pi) - \lambda)$$

$$+ \sum_{I(x_i > 0)} \log(1 - \pi) + x_i \log(\lambda) - \lambda - \log(x_i!))$$

M step:

$$\frac{\partial Q}{\partial \lambda} = 0$$

$$= \sum_{I(x_i = 0)} (1 - E[z_i])(-1) + \sum_{I(x_i > 0)} \left(\frac{x_i}{\lambda} - 1\right) = 0$$

$$\Rightarrow \hat{\lambda} = \frac{\sum_{I(x_i > 0)} x_i}{n - \sum_{I(x_i = 0)} E[z_i]}$$

$$\hat{\lambda} = \frac{\sum_{I(x_i > 0)} x_i}{n - \sum_{I(x_i = 0)} \hat{z}_i}$$

where $\hat{z} = \frac{\pi_0}{\pi_0 + (1 - \pi_0)\, e^{-\lambda_0}}$

$$\frac{\partial Q}{\partial \pi} = 0$$

$$= \sum_{I(x_i = 0)} \left(\frac{E[z_i]}{\pi} - \frac{1 - E[z_i]}{1 - \pi}\right) - \sum_{I(x_i > 0)} \frac{1}{1 - \pi} = 0$$

$$= \sum_{I(x_i = 0)} \left(\frac{E[z_i]}{\pi} + \frac{E[z_i]}{1 - \pi}\right) - \frac{n}{1 - \pi} = 0$$

$$\Rightarrow \hat{\pi} = \sum_{I(x_i = 0)} \frac{\hat{z}_i}{n}$$

Thus parameter updates are as follows:

$$\hat{z}_1 = \frac{\pi_0}{\pi_0 + (1 - \pi_0)\, e^{-\lambda_0}}$$

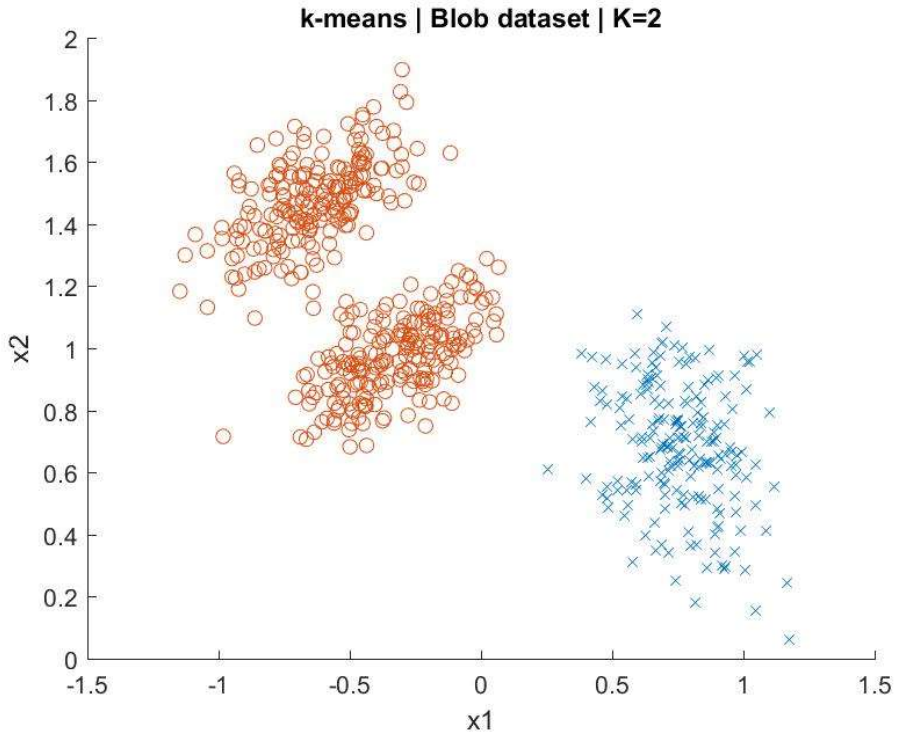
$$\hat{\lambda}_1 = \frac{\sum_{I(x_i > 0)} x_i}{n - \sum_{I(x_i = 0)} \hat{z}_1}$$

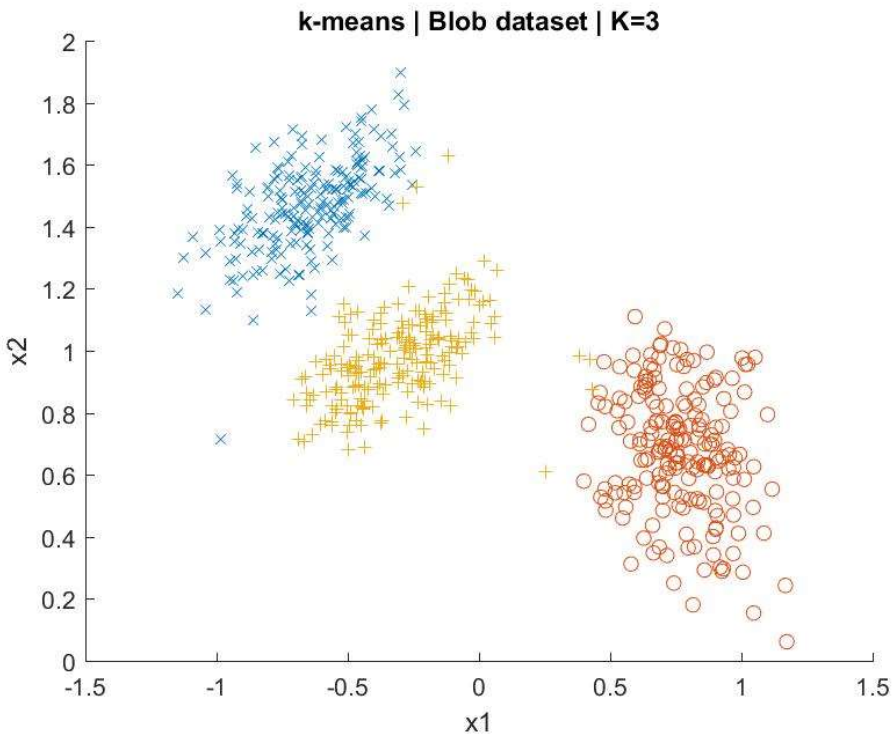
$$\hat{\pi} = \sum_{I(x_i = 0)} \frac{\hat{z}_1}{n}$$

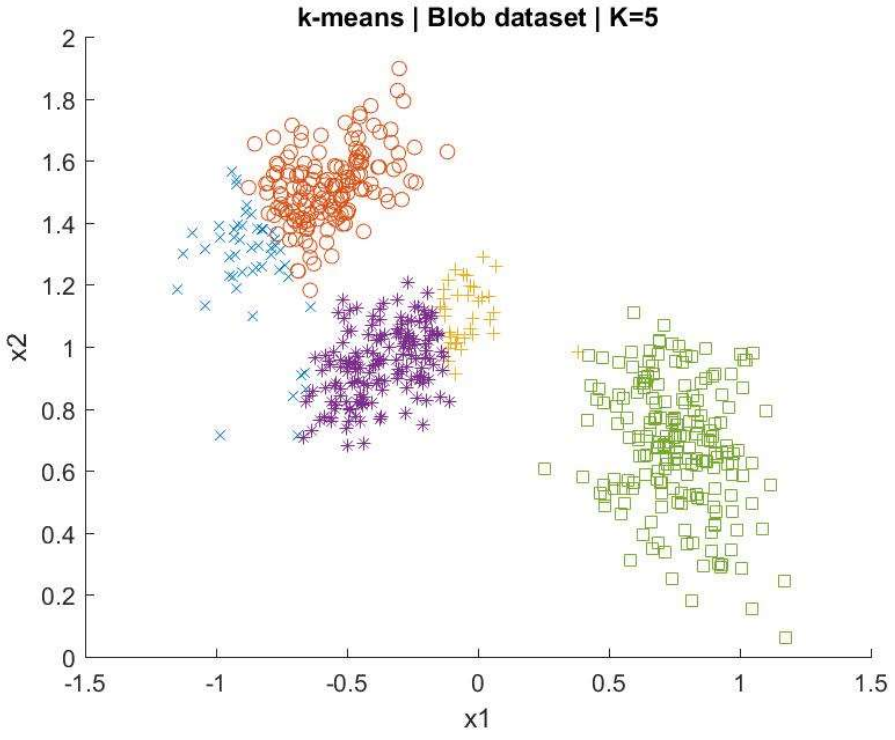
4 Programming

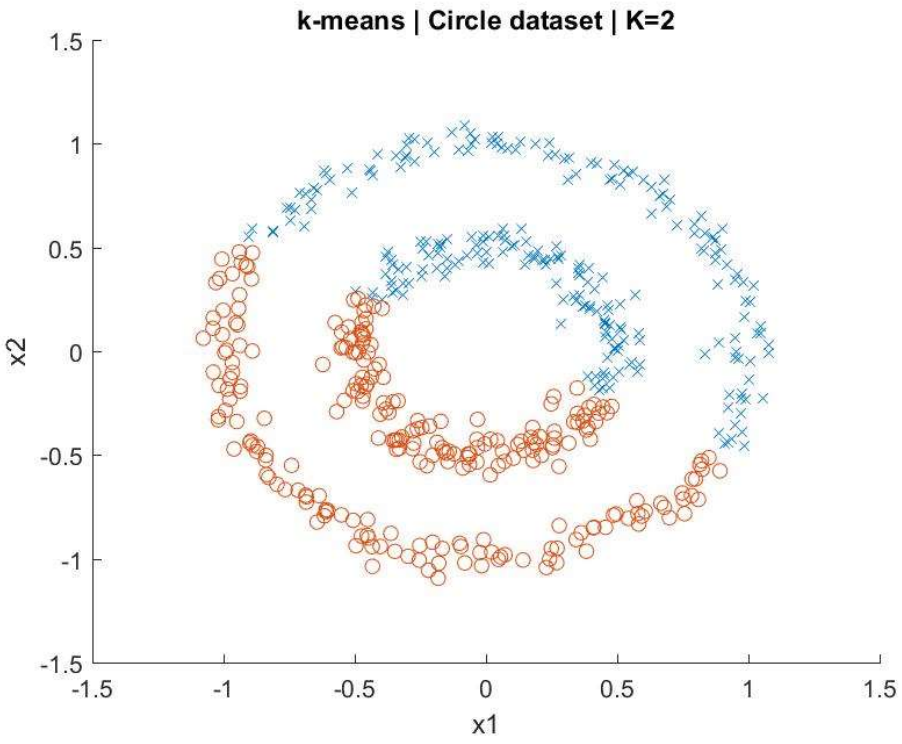
4.2 Implement k-means

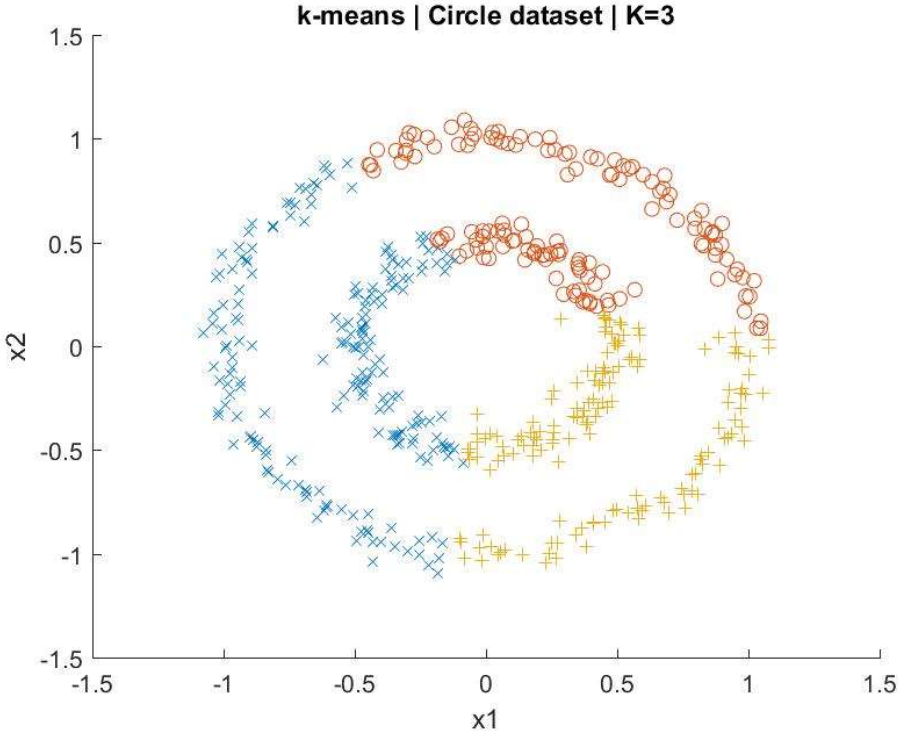
a)

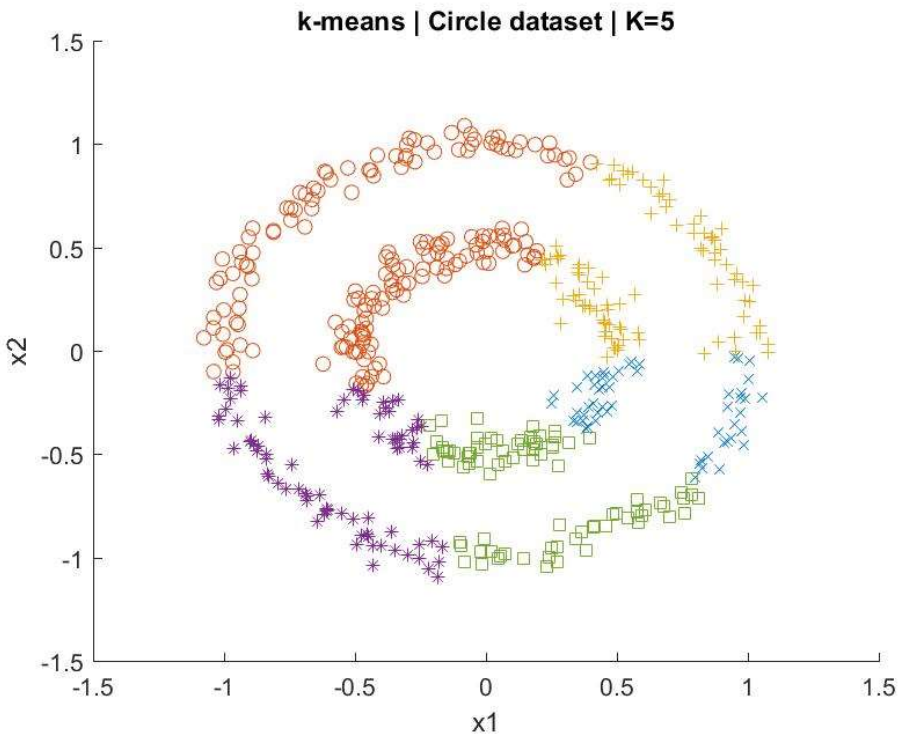



k-means | Blob dataset | K=3
x2
x1
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-1.5
-1
-0.5
0
0.5
1
1.5


k-means | Blob dataset | K=5
x2
x1
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-1.5
-1
-0.5
0
0.5
1
1.5


k-means | Circle dataset | K=2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-1.5
-1
-0.5
0
0.5
1
1.5
x1


k-means | Circle dataset | K=3
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-1.5
-1
-0.5
0
0.5
1
1.5
x1

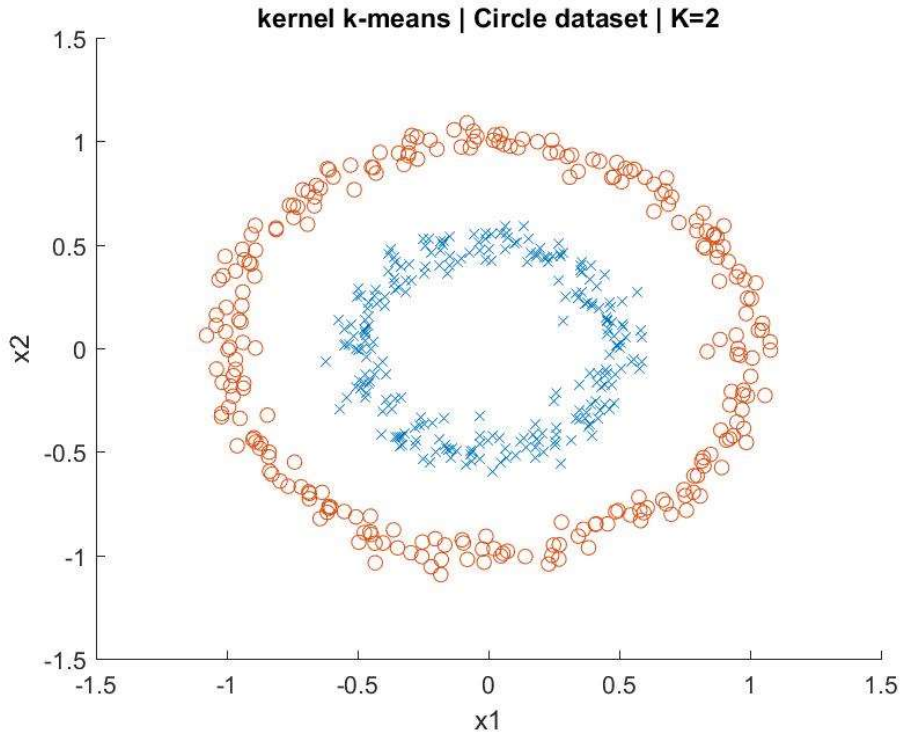


b)As we can see from the above plots K-means fails on the circular dataset, this is due to the fact that k means assumes equal variance on all data and, both the clusters have approximately the same mean with further exacerbates the problem of separating the data. Thus this data is not linearly separable in this feature space. It can be separated if mapped to another feature space, which makes it linearly separable.

4.3 Implement kernel k-means

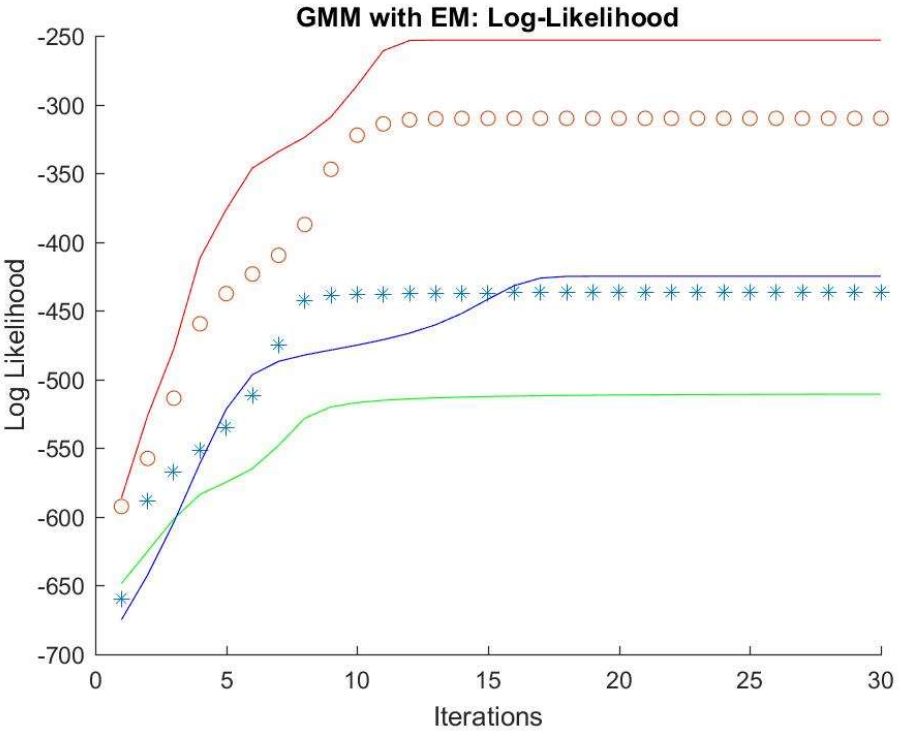
a) Choice of kernel: Polynomial with Degree = 2 and Constant = 0

b)

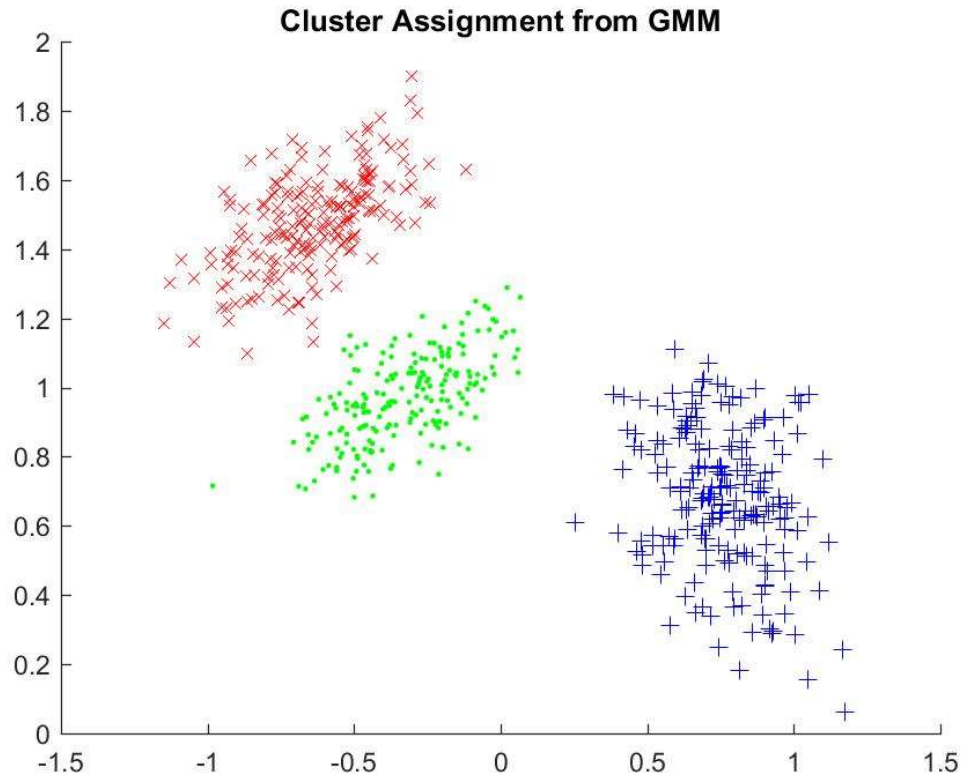


4.4 Implement Gaussian Mixture Model

a)



b)



The best case is: 1
The mean and covariance for first distribution is:
-0.6231 1.5797

0.0359 0.0173
0.0173 0.0305

The mean and covariance for second distribution is:
0.5275 0.8065

0.0971 -0.0464
-0.0464 0.0606

The mean and covariance for third distribution is:
-0.3285 0.9729

0.0334 0.0132
0.0132 0.0154