MACHINE LEARNING
ASSIGNMENT 5
AJAY KUMAR LOGANATHAN RAVICHANDRAN
USC ID - 1669468906

1) Clustering

a)

Given:

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

To find:

Show that if μ_k is the mean of all data points assigned to the cluster k, for any k, then the objective D is minimized

Solution :

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

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

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

For D to be minimized, it’s first order derivative w.r.t. μ_k should be equated to zero.

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

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

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

b)

Given:

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

To find:

μ_k *when D is minimum*

Solution:

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

D is minimized when $\frac{\partial D}{\partial \mu_k} = 0$

We know that $\frac{\partial \|x\|}{\partial x} = sign(x)$ *where,*

$sign(x) = -1 \; if \; x < 0$

$sign(x) = 1 \; if \; x > 0$

$sign(x) = 0 \; if \; x = 0$

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

Thus, for each cluster k,

$\sum_{n=1}^{N} sign(x_n - \mu_k) = 0$

Let n_1 denote the number of points for which $sign(x_n - \mu_k) = -1, \; i.e \; \mu_k > x_n$

Let n_2 denote the number of points for which $sign(x_n - \mu_k) = 1, \; i.e \; \mu_k < x_n$

From $\sum_{n=1}^{N} sign(x_n - \mu_k) = 0, \;\; n_1 - n_2 = 0, \;\; \Rightarrow \; n_1 = n_2$

For the above equation to be true, μ_k is the element-wise median of the points assigned to cluster k.

c)

i)

Given:

We apply a mapping φ(x) to map data points into feature space. Then, we define the objective function of kernel K-means as

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

i)

To Find:

$\overline{D}$ can be represented in terms of only one kernel $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$

Solution:

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

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

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

$$\Rightarrow \quad = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}(\Phi(x_n)^T\Phi(x_n) \ - 2\Phi(x_n)^T\overline{\mu_k} + \overline{\mu_k}^T\overline{\mu_k})$$

The center of the cluster k ($\overline{\mu_k}$) = $\frac{\sum_{i=1}^{N}\Phi(x_i)r_{ik}}{\sum_{i=1}^{N} r_{ik}}$

Substituting $\overline{\mu_k}$ in $\overline{D}$:

$$\Rightarrow \overline{D} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\Phi(x_n)^T\Phi(x_n) - \frac{2\Phi(x_n)^T\sum_{i=1}^{N}\Phi(x_i)r_{ik}}{\sum_{i=1}^{N} r_{ik}} + \frac{\sum_{i=1}^{N}\Phi(x_i)r_{ik}\sum_{j=1}^{N}\Phi(x_j)r_{jk}}{\sum_{i=1}^{N} r_{ik}\sum_{j=1}^{N} r_{jk}}]$$

Expressing $\Phi(x_n)^T\Phi(x_n)$ as a kernel function $K(x_i, x_j)$, we get:

$$\Rightarrow \overline{D} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}[\ K(x_n, x_n) \ - \frac{2\sum_{i=1}^{N} r_{ik}K(x_n, x_i)}{\sum_{i=1}^{N} r_{ik}} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik}\, r_{jk}\, K(x_i, x_j)}{(\sum_{i=1}^{N} r_{ik})^2}\ ] \text{ Since, } \sum_{i=1}^{N} r_{ik} = \sum_{j=1}^{N} r_{jk}$$

Therefore, $\overline{D}$ represented in terms of only kernel $K(x_i, x_j)$ is as shown below:

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

ii)

To Find: The equation of assigning a point to its cluster.

Solution:

For assigning a given data point to a cluster, we compute the distance to each cluster and assign it to the cluster for which the distance is minimum.

The distance formula can be shown as :

$$k(x) \ = \ argmin_c||\Phi(x) \ - \overline{\mu_c}||_2^2$$

We know that,

$$||\Phi(x) \ - \overline{\mu_c}||_2^2 \ = \ \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}[\ K(x_n, x_n) \ - \frac{2\sum_{i=1}^{N} r_{ik}K(x_n, x_i)}{\sum_{i=1}^{N} r_{ik}} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik}\, r_{jk}\, K(x_i, x_j)}{(\sum_{i=1}^{N} r_{ik})^2}\ ],$$

So for the data points(i) belonging to a particular cluster(c), only their r_{ic} will be 1 and the term for other points will be 0.

Thus, the above distance equation can be simplified to calculate distance of a datapoint x to a particular cluster(c) by:

$$||\Phi(x) - \overline{\mu_c}||_2^2 = [\, K(x,x) - \frac{2 \sum_{x_i \in c} K(x,x_i)}{|x_i \in C|} + \frac{\sum_{x_i \in c_k} \sum_{x_j \in c_k} K(x_i,x_j)}{|x_i \in C|^2} \,]$$

Thus,

$$k(x) = argmin_c[\, K(x,x) - \frac{2 \sum_{x_i \in c} K(x,x_i)}{|x_i \in C|} + \frac{\sum_{x_i \in c_k} \sum_{x_j \in c_k} K(x_i,x_j)}{|x_i \in C|^2} \,]$$

iii)

To Find: The pseudo-code of the complete kernel K-means algorithm including initialization of cluster centers.

Solution:

Pseudo code for complete kernel K means :

→Compute the kernel matrix $K(X_i, X_j)$ with values $\Phi^T(X_i)\Phi(X_j)$

→ Initialize clusters with random assignment of data points $(x_1, x_2, ..., x_n)$

→ For each data point **x**, find the distance of the point to each cluster **c** using the formula shown below and assign it to the cluster with the minimum distance.

$$||\Phi(x) - \overline{\mu_c}||_2^2 = [\, K(x,x) - \frac{2 \sum_{x_i \in c} K(x,x_i)}{|x_i \in C|} + \frac{\sum_{x_i \in c_k} \sum_{x_j \in c_k} K(x_i,x_j)}{|x_i \in C|^2} \,]$$

→ Repeat the above step till the cluster assignment doesn't change.

2) Gaussian Mixture Model

To Find:

Likelihood function of x_1 as a function of α and determine the maximum likelihood estimation of α

Given:

Gaussian distributions:

$f(x_1/\theta_1)$ *with* $\mu_1 = 0$ *and* $\sigma_1^2 = 1$

$\Rightarrow f(x_1/\theta_1) = \frac{1}{\sqrt{2\pi}} exp\, (\frac{-x_1^2}{2})$

$f(x_1/\theta_2)$ *with* $\mu_2 = 0$ *and* $\sigma_2^2 = 0.5$

$\Rightarrow f(x_1/\theta_2) = \frac{1}{\sqrt{\pi}}exp\,(-x_1^2)$

Prior probability of $\theta_1(f(\theta_1)) = \alpha$

Prior probability of $\theta_2(f(\theta_2)) = 1 - f(\theta_1) = 1 - \alpha$

Therefore,

Log likelihood :

$P(x_1) = P(x_1/\alpha) \quad = f(x_1/\theta_1)f(\theta_1) + f(x_1/\theta_2)f(\theta_2)$

$= \alpha\frac{1}{\sqrt{2\pi}}exp\,(\frac{-x_1^2}{2}) + (1-\alpha)\frac{1}{\sqrt{\pi}}exp\,(-x_1^2)$

$= [\frac{1}{\sqrt{2\pi}}exp\,(\frac{-x_1^2}{2}) - \frac{1}{\sqrt{\pi}}exp\,(-x_1^2)]\,\alpha + \frac{1}{\sqrt{\pi}}exp\,(-x_1^2)$

To find the maximum likelihood estimator of α, we differentiate the above equate w.r.t to α,

$\frac{\partial P(x_1)}{\partial\alpha} = \frac{1}{\sqrt{2\pi}}exp\,(\frac{-x_1^2}{2}) - \frac{1}{\sqrt{\pi}}exp\,(-x_1^2) = slope\,(m)$

Slope is positive when,

$\frac{1}{\sqrt{2\pi}}exp\,(\frac{-x_1^2}{2}) > \frac{1}{\sqrt{\pi}}exp\,(-x_1^2)$

$\Rightarrow \frac{exp\,(\frac{-x_1^2}{2})}{exp\,(-x_1^2)} >= \sqrt{2}$

$\Rightarrow exp\,(\frac{x_1^2}{2}) >= \sqrt{2}$

$\Rightarrow x_1^2 >= 2log(\sqrt{2})$

When slope is negative,

$f(x_1|\theta_1) < f(x_1\,|\theta_2)$,

Therefore set $\alpha = 0$

Thus whenever $x_1^2 >= 2log(\sqrt{2})$, we assign $\alpha = 1$ *and* $\alpha = 0$ *otherwise*

3) EM Algorithm

Given:

$p(x_i) = \{\pi + (1 - \pi)e^{-\lambda}\}$ *if* $x_i = 0$

$= \{(1 - \pi)\frac{\lambda^{x_i}e^{-\lambda}}{x_i!}\}$ *if* $x_i > 0$

a)

To Find: Complete likelihood function by using a proper hidden variable for the observations
Solution:
Since $p(x_i) = 0$ can be observed either from the poisson distribution or from the degenerate distribution, We define a hidden variable z_i to observe from which distribution $x_i = 0$ is coming from, where:

$z_i = 1 \; if \; x_i = 0 \; from \; zero-valued \; distribution$
$z_i = 0 \; if \; x_i = 0 \; from \; poisson's \; distribution$

Therefore ,
$p(x_i = 0, z_i = 1) = \pi$
$p(x_i = 0, z_i = 0) = (1-\pi)e^{-\lambda}$

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

$$log\, L = \sum_{x_i=0} z_i(log\,\pi) + (1-z_i)[log(1-\pi)-\lambda] + \sum_{x_i>0} [log(1-\pi)-\lambda+x_i log\lambda] - logx_i!$$

b)
Update equation for E-STEP
Updated parameter (θ) = (π, λ)
Old parameter (θ_0) = (π_0, λ_0)
$E_{P(z|x)}(z_i)$ = The expectation of probability of z given x_i

E-STEP:

$$Q(\theta, \theta_0) = \sum_{x_i=0} E_{P(z|x)}(z_i)(log\,\pi) + (1-E_{P(z|x)}(z_i))[log(1-\pi)-\lambda]$$
$$+ \sum_{x_i>0} [log(1-\pi)-\lambda+x_i log\lambda] - logx_i!$$

Where $E_{P(z|x)}(z_i) = \frac{P(x_i|z_i=1)P(z_i=1)}{P(x_i|z_i=0)P(z_i=0)+P(x_i|z_i=1)P(z_i=1)} = \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}$

Therefore,

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

Update equation for M-STEP:

Differentiate the E-Step function with respect to the parameter and equate it to 0 in order to get the new update parameter

$\rightarrow$ *To calculate* λ , *differentiate Q w.r.t* λ

$$\frac{\partial Q}{\partial \lambda} = \sum_{x_i=0} (1 - \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}})(-1) + \sum_{x_i>0} (-1) + \frac{x_i}{\lambda} = 0$$

$$\Rightarrow -\sum_{x_i=0} 1 + \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}} - \sum_{x_i>0} (1) + \sum_{x_i>0} \frac{x_i}{\lambda} = 0$$

$$\Rightarrow \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}} + \sum_{x_i>0} \frac{x_i}{\lambda} = \sum_{x_i=0} 1 + \sum_{x_i=0} 1$$

$$\Rightarrow \sum_{x_i>0} \frac{x_i}{\lambda} = n - \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}$$

$$\Rightarrow \lambda = \frac{\sum_{x_i>0} x_i}{n - \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}}$$

$\rightarrow$ *To calculate* π, *differentiate Q w.r.t* π

$$\frac{\partial Q}{\partial \pi} = \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}(\frac{1}{\pi}) + (1 - \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}})(\frac{1}{1-\pi})(-1) + \sum_{x_i>0} (\frac{1}{1-\pi})(-1) = 0$$

$$\Rightarrow \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}(\frac{1}{\pi}) + (\frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}})(\frac{1}{1-\pi}) - \sum_{x_i=0} (\frac{1}{1-\pi}) - \sum_{x_i>0} (\frac{1}{1-\pi}) = 0$$

$$\Rightarrow \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}(\frac{1}{\pi}) + (\frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}})(\frac{1}{1-\pi}) - \frac{n}{1-\pi} = 0$$

$$\Rightarrow \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}(\frac{1}{\pi} + \frac{1}{1-\pi}) - \frac{n}{1-\pi} = 0$$

$$\Rightarrow \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}(\frac{1}{\pi(1-\pi)}) - \frac{n}{1-\pi} = 0$$

$$\Rightarrow \sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}(\frac{1}{\pi}) = n$$

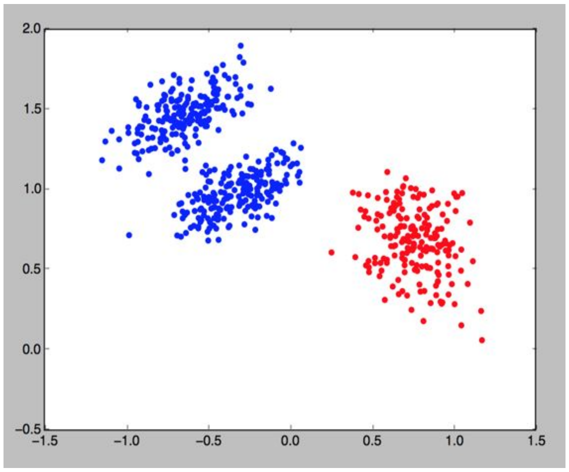
$$\Rightarrow \pi = \frac{\sum_{x_i=0} \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}}{n}$$

4. Programming

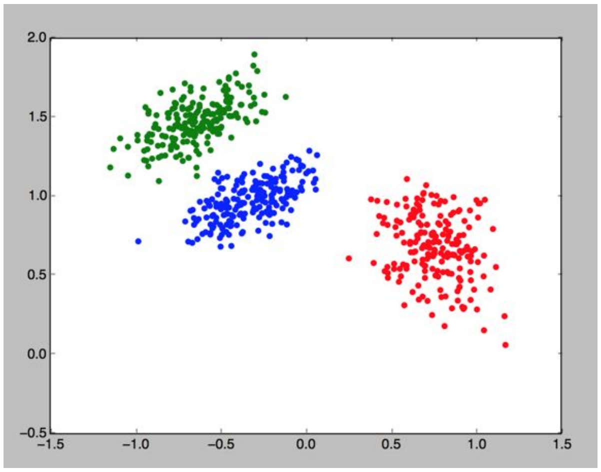
4.2 K-means

a)

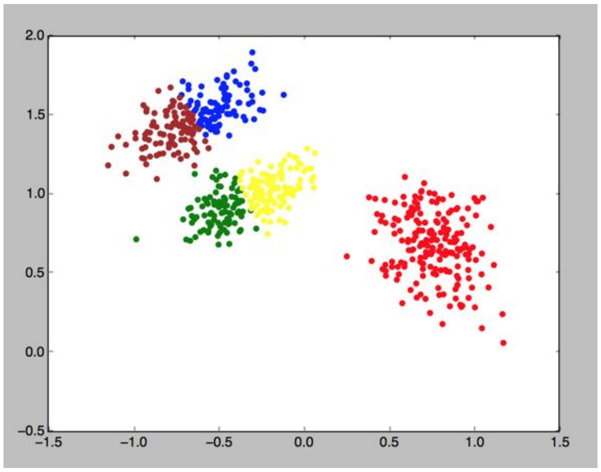
K = 2 for hw5_blob.csv



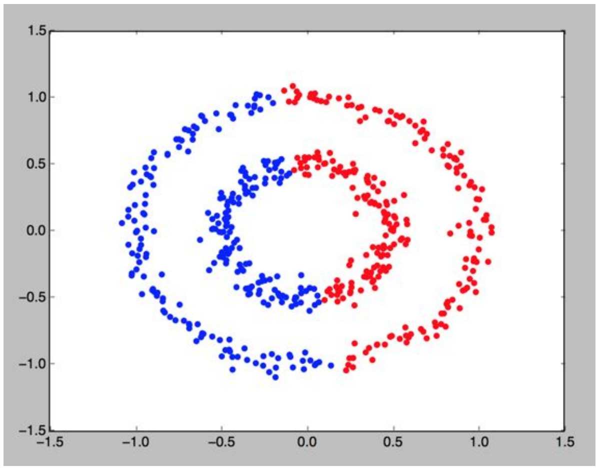
K = 3 for hw5_blob.csv



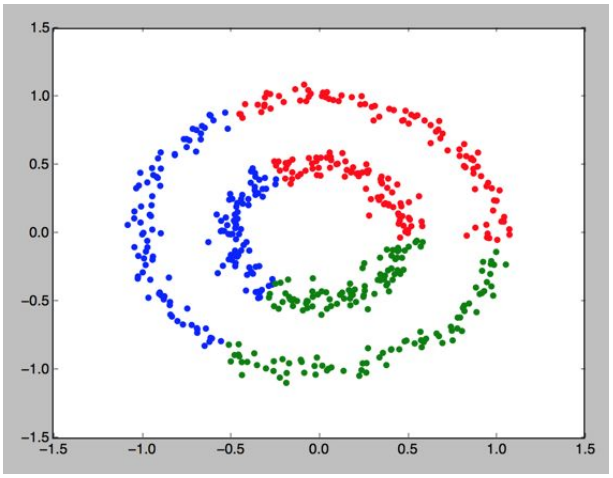
K = 5 for hw5_blob.csv



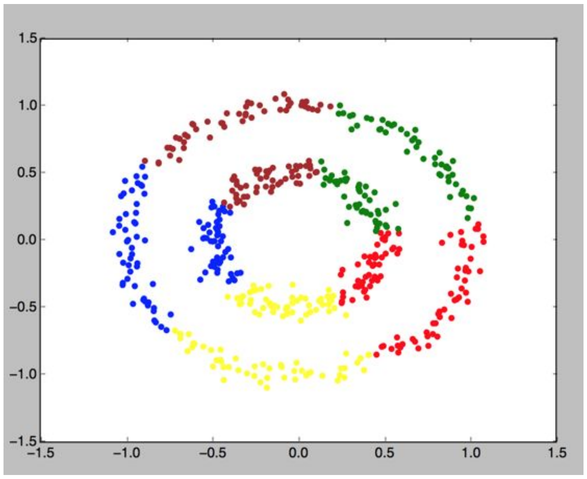
K = 2 for hw5_circle.csv



K = 3 for hw5_circle.csv



K = 5 for hw5_circle.csv



b)

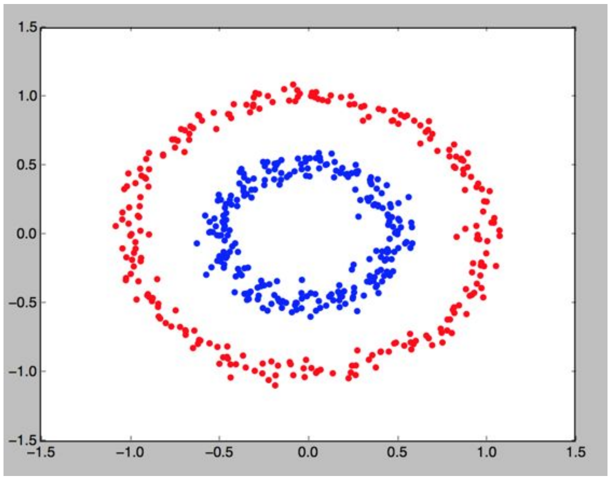
- The k-means algorithm separates data with linear decision boundaries. In our circle dataset, the data is not linearly separable in two dimensional space. Thus, we need to transform our features into higher dimensions to make them separable.
- Transforming into higher dimensions will help the data points falling in two concentric circles into two separate clusters.

4.3) Kernel K-means

a) Kernel function = Polynomial kernel

$$\begin{aligned} K(X_i, X_j) &= X_i^T X_j + (6(X_i^T X_i)).\ (6(X_j^T X_j)) \\ &= \|X_i X_j\| + (6\|X_i\|_2^2).\ (6\|X_j\|_2^2) \end{aligned}$$

b) Cluster Plot:



K = 2 for hw5_circle.csv

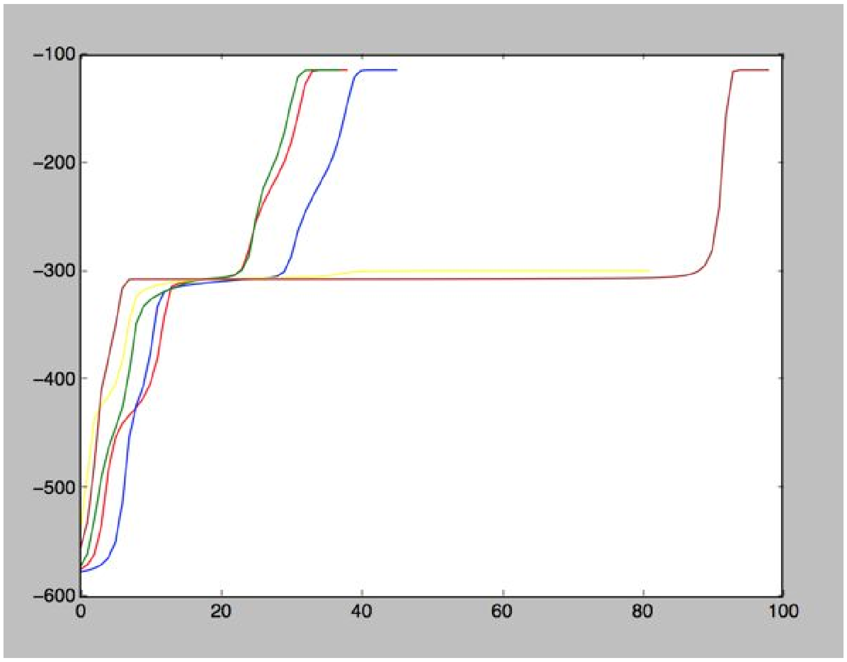
4.4)

a)

```
 -----------------Run   1-----------------

 CLUSTER CENTER
[array([-0.32592     ,   0.97133511]), array([-0.63946276,   1.47460509]), array([ 0.75896032,
0.67976982])]

 CLUSTER COVARIANCE
[array([[ 0.03604936,  0.0146391 ],
       [ 0.0146391 ,  0.01629116]]), array([[ 0.03596748,  0.01549305],
       [ 0.01549305,  0.01935205]]), array([[ 0.02717056, -0.00840045],
       [-0.00840045,  0.040442  ]])]

 PRIOR PROBABILITY
[0.33550529649107846, 0.33115952335791699, 0.3333351801510045]

 LOG LIKELIHOOD
-113.670867941

 -----------------Run   2-----------------

 CLUSTER CENTER
```

```
[array([ 0.75896032,   0.67976982]), array([-0.63946282,   1.47460567]), array([-0.32592047,
0.97133538])]

 CLUSTER COVARIANCE
[array([[ 0.02717056, -0.00840045],
      [-0.00840045,  0.040442  ]]), array([[ 0.03596754,  0.01549309],
      [ 0.01549309,  0.01935189]]), array([[ 0.03604944,  0.014639  ],
      [ 0.014639  ,  0.01629118]])]

 PRIOR PROBABILITY
[0.33333517940301871, 0.33115896021413554, 0.33550586038284569]

 LOG LIKELIHOOD
-113.670867858

 -----------------Run   3-----------------

 CLUSTER CENTER
[array([-0.32592049,   0.9713354 ]), array([ 0.75896032,   0.67976982]), array([-0.63946282,
1.4746057 ])]

 CLUSTER COVARIANCE
[array([[ 0.03604944,  0.01463899],
      [ 0.01463899,  0.01629118]]), array([[ 0.02717056, -0.00840045],
      [-0.00840045,  0.040442  ]]), array([[ 0.03596754,  0.0154931 ],
      [ 0.0154931 ,  0.01935188]])]

 PRIOR PROBABILITY
[0.33550588811782628, 0.33333517936619772, 0.33115893251597606]

 LOG LIKELIHOOD
-113.670867856

 -----------------Run   4-----------------

 CLUSTER CENTER
[array([ 0.73452602,   0.3830169 ]), array([ 0.76342942,   0.72463715]), array([-0.48102884,
1.22093723])]
```

```
 CLUSTER COVARIANCE
[array([[ 0.05134616, -0.02639723],
       [-0.02639723,  0.01883768]]), array([[ 0.02311226, -0.0069754 ],
       [-0.0069754 ,  0.02838428]]), array([[ 0.06106409, -0.02467728],
       [-0.02467728,  0.08128329]])]

 PRIOR PROBABILITY
[0.043773785068823255, 0.28903782255356664, 0.66718839237761018]

 LOG LIKELIHOOD
-298.877800036

 -----------------Run   5-----------------

 CLUSTER CENTER
[array([ 0.75896032,   0.67976982]), array([-0.32592018,   0.97133521]), array([-0.63946278,
1.4746053 ])]

 CLUSTER COVARIANCE
[array([[ 0.02717056, -0.00840045],
       [-0.00840045,  0.040442  ]]), array([[ 0.03604939,  0.01463906],
       [ 0.01463906,  0.01629117]]), array([[ 0.0359675 ,  0.01549307],
       [ 0.01549307,  0.01935199]])]

 PRIOR PROBABILITY
[0.33333517987692846, 0.33550550320134009, 0.33115931692173151]

 LOG LIKELIHOOD
-113.670867906
```



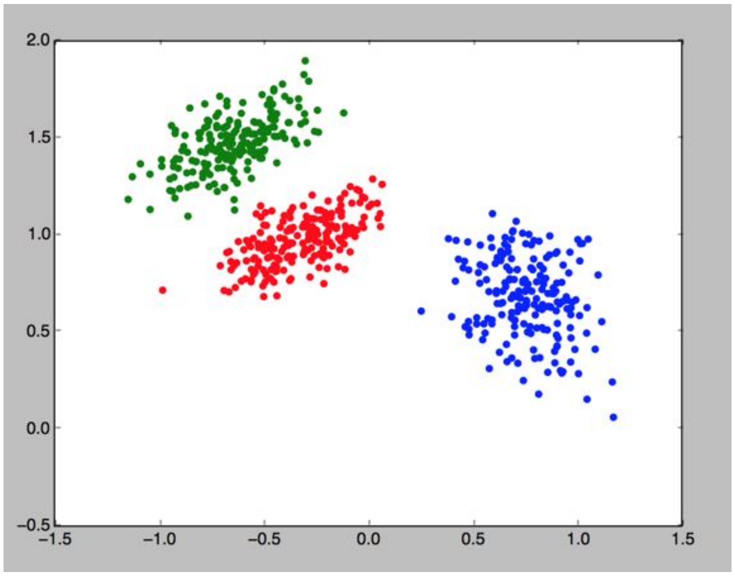
No of iterations (vs) Log likelihood

b)

1)

BEST LOG-LIKELIHOOD : -113.670867856

MOST LIKELY CLUSTER ASSIGNMENT



K = 3

2)

```
 BEST CLUSTER CENTERS

 CLUSTER CENTER FOR CLUSTER  1

  X :  -0.325920492145
  Y :  0.971335397

 CLUSTER CENTER FOR CLUSTER  2

  X :  0.758960323797
  Y :  0.679769821025

 CLUSTER CENTER FOR CLUSTER  3

  X :  -0.639462823433
  Y :  1.47460569574

 BEST CLUSTER COVARIANCE

 COVARIANCE FOR CLUSTER  1

[[ 0.03604944  0.01463899]
 [ 0.01463899  0.01629118]]

 COVARIANCE FOR CLUSTER  2

[[ 0.02717056 -0.00840045]
 [-0.00840045  0.040442  ]]

 COVARIANCE FOR CLUSTER  3

[[ 0.03596754  0.0154931 ]
 [ 0.0154931   0.01935188]]

 BEST PRIOR PROBABILITY


 PRIOR PROBABILITY FOR CLUSTER  1 :  0.335505888118
```

PRIOR PROBABILITY FOR CLUSTER 2 : 0.333335179366

PRIOR PROBABILITY FOR CLUSTER 3 : 0.331158932516

COLLABORATORS:

SINDHUJHA SETHURAMAN
LAVANYA KUMAR