1. Clustering

(a)

Given,

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

where μ_k is prototype of the k-th cluster, r_{nk} is a binary indicator variable. If x_n is assigned to the cluster k, r_{nk} is 1 otherwise r_{nk} is 0.

Now, assuming that all r_{nk} are known, we can simplify the above equation as follows,

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

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

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

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

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

(b)

The distortion measure can be changed to,

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

Differentiating with respect to mean we get,

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

The value of μ_k which will satisfy the above equation will ensure that there are equal number of points to the left and right for the cluster in consideration.

Calculating particular cluster of size m:

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

$$sign(x_m - \mu_k) = \begin{cases} +1 \; if \; x_m - \mu_k > 0 \\ -1 \; if \; x_m - \mu_k < 0 \end{cases}$$

So, $\psi(x_n | x_n - \mu_k > 0) - \psi(x_n | x_n - \mu_k < 0) = 0$ where ψ denotes number of elements.

This becomes zero at median.

(c)

(i) The objective function of kernel K-means by applying a mapping $\phi(x)$ to map points into feature space can be defined as,

$$\widetilde{D} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left|\left| \phi(x_n) - \tilde{\mu}_k \right|\right|_1$$

where $\tilde{\mu}_k$ is the centre of the cluster k in feature space,

$$\tilde{\mu}_k = \frac{\sum_{n=1}^{N} r_{ik} \phi(x_i)}{\sum_{n=1}^{N} r_{ik}}$$

Consider $\left|\left| \phi(x_n) - \tilde{\mu}_k \right|\right|^2$

$$\left|\left| \phi(x_n) - \tilde{\mu}_k \right|\right|^2 = (\phi(x_n) - \tilde{\mu}_k)^T (\phi(x_n) - \tilde{\mu}_k)$$

$$= \phi(x_n)^T \phi(x_n) - 2\tilde{\mu}^T \phi(x_n) + \tilde{\mu}^T \tilde{\mu}$$

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

Define $n_k = \sum_{i=1}^{N} r_{ik}$, so we get,

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

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

So, finally

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

(ii) For a point x_n, calculate $\widetilde{D}$ for all possible clusters k

Assigning cluster to x_n,

$$r_{nk} = \begin{cases} 1 & k = argmin_k \left|\left|\phi(x_n) - \tilde{\mu}_k\right|\right|_k^2 \\ 0 & otherwise \end{cases}$$

Where,

$$\left|\left|\phi(x_n) - \tilde{\mu}_k\right|\right|_k^2 = \widetilde{D} = \sum_{n=1}^{N} K(x_n, x_n) - 2\frac{\sum_{i=1}^{N} r_{ik} K(x_i, x_n)}{n_k} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{jk} r_{ik} K(x_i, x_j)}{n_k^2}$$

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

(iii)

Algorithm K Means

Procedure kernel k means

//selecting centroids

Centroid[i] = x(random(1..N)) for $1 \leq i \leq k$

//calculating the value of kernel fucntion

for i in range(N):

for j in range(N):

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

end for

end for

$r(n,k) < -[0]$

for i in range(N):

//distance and centroid calculations

$j = argmin_k k\left|\left|\phi(x_n) - \mu_k\right|\right|^2$

r[i,j] = 1

Update centroid[j]

end for

end procedure

2. Gaussian Mixture Model

Given that α is the mixing parameter for the two Gaussian distributions, we can write,

$$\alpha = f(x|\theta_1)$$

Using the fact that α is the mixing parameter, we can also write,

$$1 - \alpha = f(x|\theta_2)$$

Now calculating the likelihood function for α using the distributions, we get,

$$L(\alpha) = P(c_1) * P(x_1|c_1) + P(c_2) * P(x_1|c_2)$$

Given that $f(x|\theta_1)$ is a Gaussian with $\mu_1 = 0$ and $\sigma^2 = 1$, we get

$$P(c_1) = \alpha$$

$$P(x_1|c_1) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x_1^2}{2}}$$

Given that $f(x|\theta_2)$ is a Gaussian with $\mu_1 = 0$ and $\sigma^2 = 0.5$, we get

$$P(c_2) = 1 - \alpha$$

$$P(x_1|c_2) = \frac{1}{\sqrt{\pi}} \exp^{-x_1^2}$$

Therefore the likelihood can be written as,

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

Simplifying, we get,

$$\boxed{L(\alpha) = \left(\frac{1}{\sqrt{2\pi}} \exp^{-\frac{x_1^2}{2}} - \frac{1}{\sqrt{\pi}} \exp^{-x_1^2} \right) \alpha + \frac{1}{\sqrt{\pi}} \exp^{-x_1^2}}$$

Here, we see that the likelihood is a linear function of α, slope is determined by the Gaussian which gives a larger response. We find that the slope is positive whenever $x_1^2 \geq \log 2$, we set $\alpha = 1$ and we set $\alpha = 0$ for all the other conditions.

3. EM Algorithm

The observed data probability of observation i is given as,

$$p(x_i) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & if\ x_i = 0 \\ (1-\pi)\dfrac{\lambda_i^x e^{-\lambda}}{x_i!} & if\ x_i > 0 \end{cases}$$

The above probability function can be represented as a function of X_i in the following way,

$$X_i = \begin{cases} 0 \quad prob = \pi + (1-\pi)e^{-\lambda} \\ x_i \quad prob = (1-\pi)\dfrac{\lambda_i^x e^{-\lambda}}{x_i!} \end{cases}$$

(a)

By defining a latent variable Z_i for all cases when $X_i = 0$. It is latent as we do not know if X_i came from Poisson or degenerate distribution during observation. So, X_i, as a mixture of degenerate distribution,

$$Z_i = \begin{cases} 1 \;\; X_i \; is \; from \; degenerate \; distribution \\ 0 \qquad\qquad\qquad\qquad\qquad otherwise \end{cases}$$

$$P(X_i = 0, Z_i = 1) = P(Z_i = 1) \times P(X_i = 0|Z_i = 1) = \pi \times 1$$

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

The likelihood function can be written as,

$$L\big((\pi,\lambda)\big|(X,Z)\big) = \prod_{x_i=0} \pi^{Z_i} \times \left((1-\pi)e^{-\lambda}\right)^{1-z_i} \times \prod_{x_i>0} (1-\pi)e^{\frac{\lambda_i^x e^{-\lambda}}{x_i!}}$$

$$\boxed{\begin{aligned} \log L = &\sum_{I(x_i=0)} z_i \log(\pi) + (1-z_i)(\log(1-\pi) - \lambda) \\ &+ \sum_{I(x_i>0)} (\log(1-\pi) + x_i \log(\lambda_i) - \lambda - \log(x_i!)) \end{aligned}}$$

Notation: $\theta = (\pi, \lambda)\ and\ \theta_0$ represents a known parameter

(b)

E Step:

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

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

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

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

Hence,

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

M Step:

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

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

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

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

where

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

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

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

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

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

Thus the updates to the parameters are:

$$\widehat{z_1} = \frac{\pi_0}{\pi_0 + (1-\pi_0)e^{-\lambda_0}}$$

$$\widehat{\lambda_1} = \frac{\sum_{I(x_i>0)} x_i}{n - \sum_{I(x_i=0)} \widehat{z_1}}$$

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

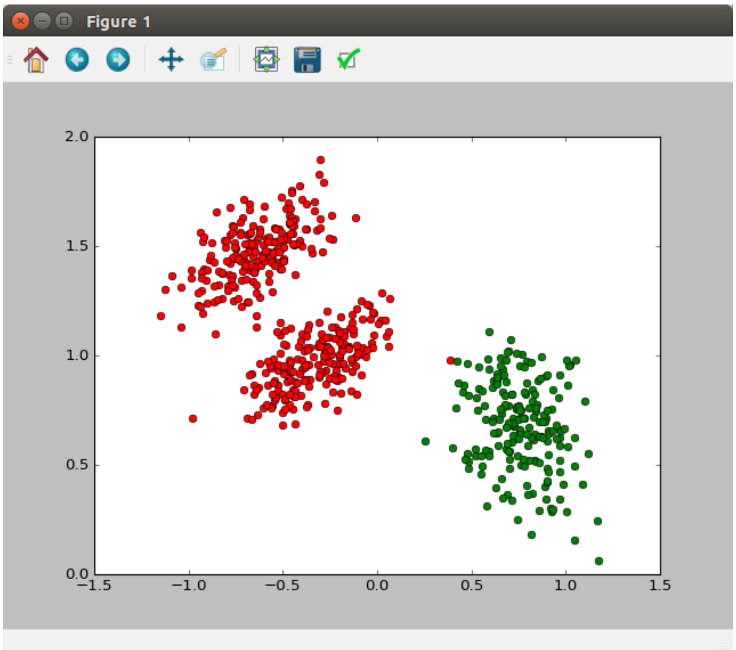
4. Programming

4.2 K-Means clustering

(a)

Blob Dataset

K=2



K=3

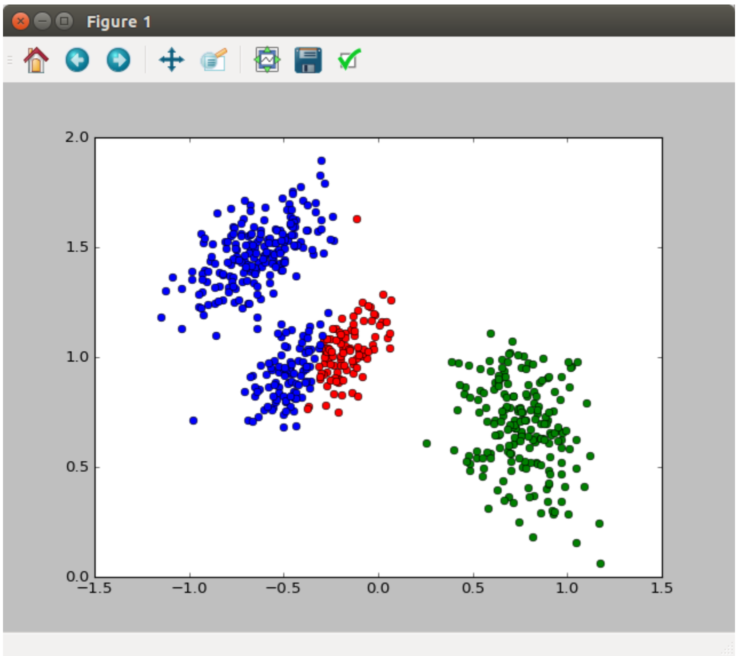


Figure 1
2.0
1.5
1.0
0.5
0.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

K=5

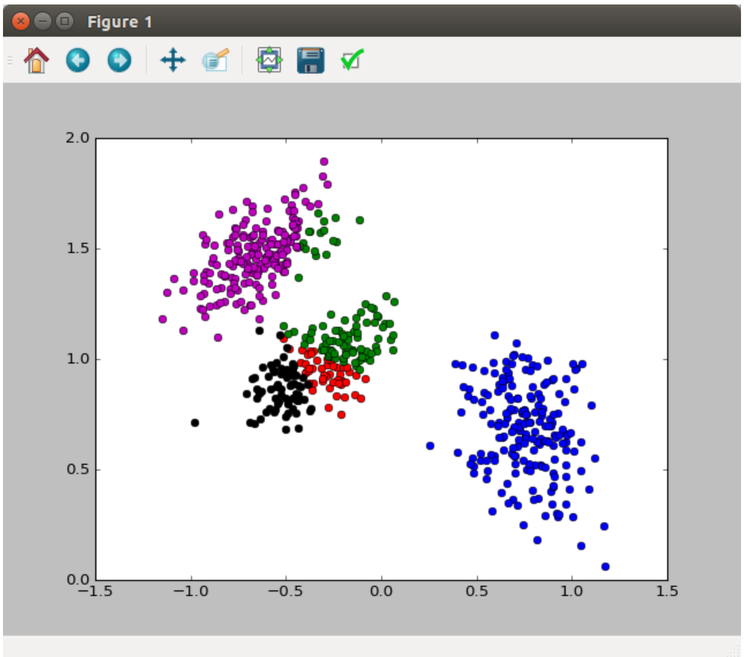
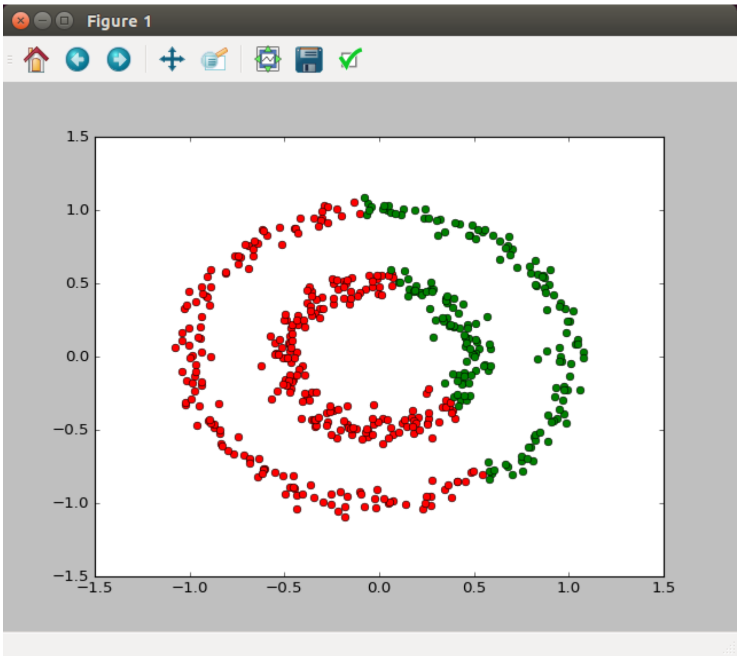


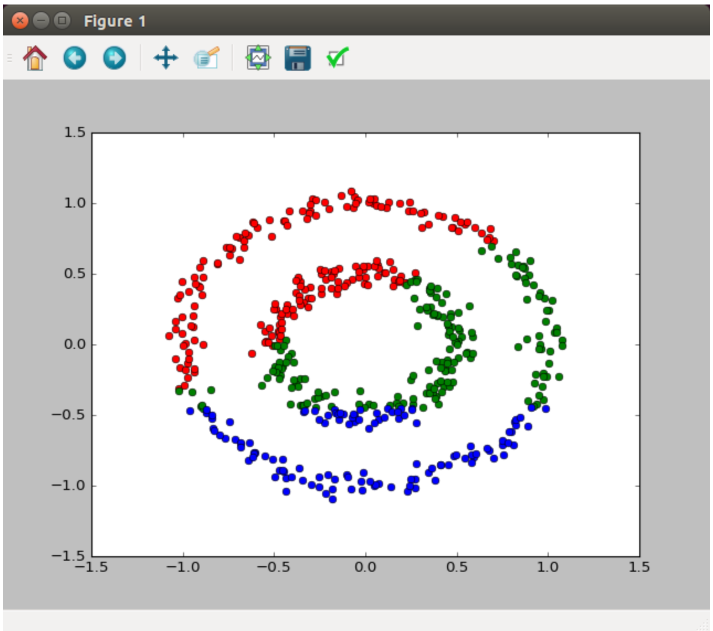
Figure 1
2.0
1.5
1.0
0.5
0.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Circle Dataset

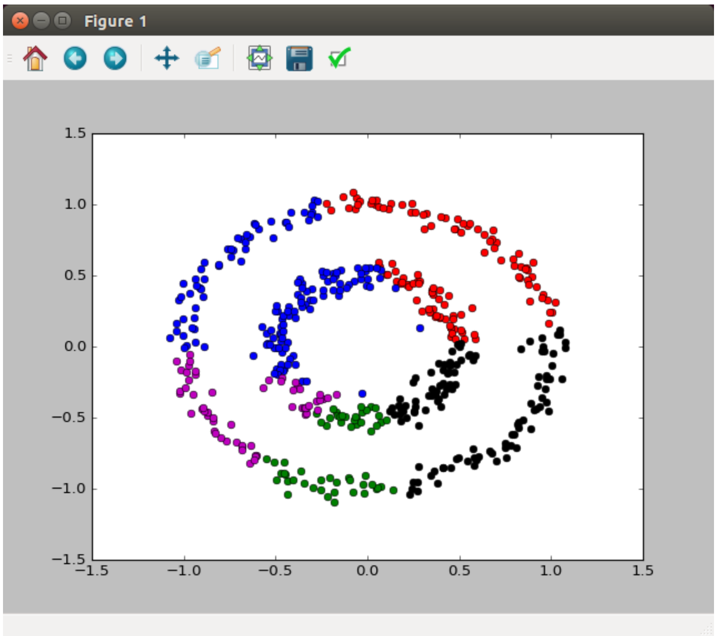
K=2



K=3



K=5



(b)

The decision boundary being used by regular K means clustering is linear which divides the circle into two halves (tries to make it a linear boundary). The circle data set isn't linearly separable. So, in the next section we try kernel techniques to try to classify it.

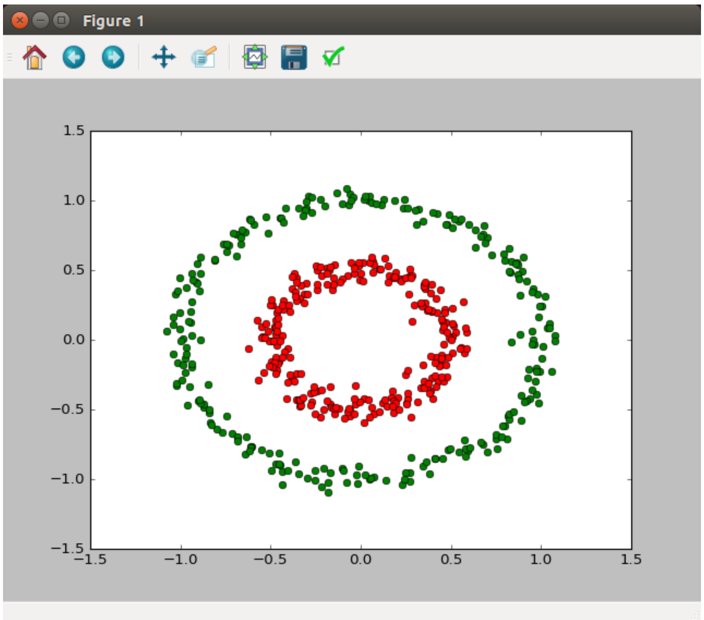
4.3 Kernel K-Means Clustering

(a)

Kernel Choice: Polynomial Kernel

$$K(x_1, x_2) = x_1^2 + x_2^2$$

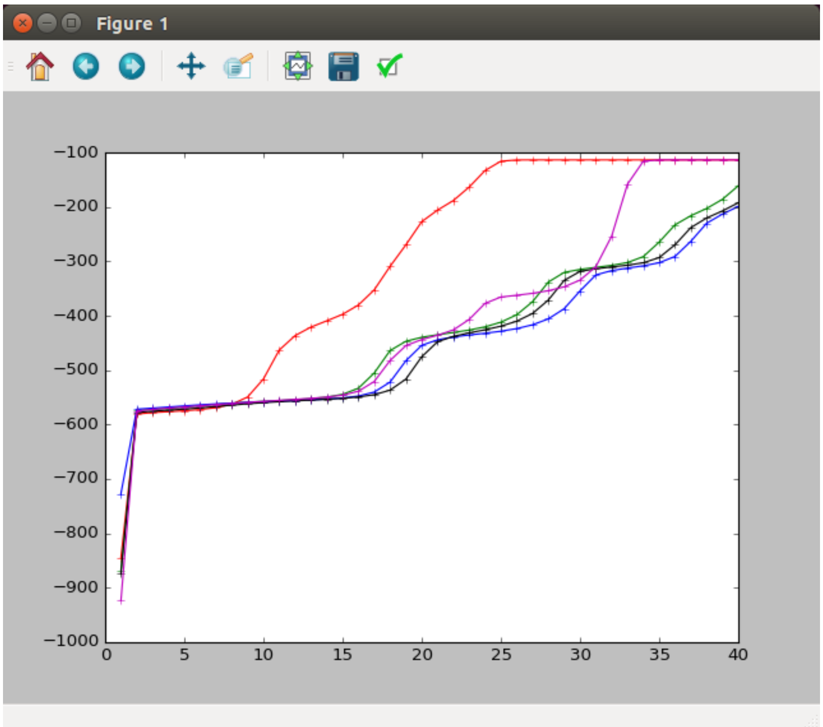
(b)



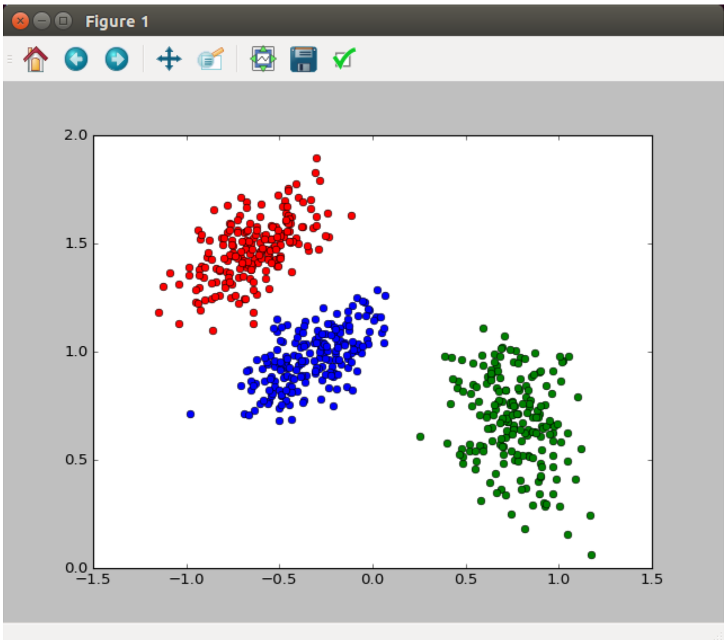
4.4 Gaussian Mixture Model

(a) Log Likelihood plots

The below graph shows the log likelihood plots for 40 iterations run 5 times to get a better convergence.



(b) Best run in terms of log likelihood



The mean and covariance values obtained for the best run are as below:

Mean Values for k = 1

[-0.63946289865377237, 1.4746064045257241]

Mean Values for k = 2

[0.75896032478310904, 0.67976982023018151]

Mean Values for k = 3

[-0.32592106449477271, 0.97133573846689225]

Covariance Values for k = 1

[[0.0359676   0.01549315]

 [0.01549315 0.01935168]]

Covariance Values for k = 2

[[0.02717056 -0.00840045]

 [-0.00840045 0.040442]]

Covariance Values for k = 3

[[0.03604954 0.01463887]

 [0.01463887 0.0162912]]

COLLABORATION

Brain stormed and collaborated with Adarsha Desai and Ravishankar Sivaraman for this assignment.