CSCI 567 Assignment 5
Fall 2016

Snehal Adsule
2080872073
adsule@usc.edu

November 9, 2016

1 Problem 1

1.1 1 (a)

Consider the given distortion function as follows:
$D = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}||x_n - \mu_k||_2^2$
Differentiating with respect to μ_k

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

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

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

The above equation shows that μ_k is nothing but mean of the the points in a particular cluster

1.2 1 (b)

Consider the L1 norm for the distortion as follows:
$D = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}||x_n - \mu_k||_1$
differentiating with respect to μ_k

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

Now,

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

$$\begin{aligned} sign(x_n - \mu_k) &= +1 \quad \text{if} \ \ x_n - \mu_k > 0 \\ &= -1 \quad \text{if} \ \ x_n - \mu_k < 0 \end{aligned}$$

Therefore, if we sort all the points we will have the optimum right at the centre , which is nothing but the median of all the points.

1.3 1 (c) 1

Kernal K means

$$\tilde{D} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\phi(x_n) - \tilde{\mu_k}||^2, \text{ where } , \tilde{\mu_k} = \frac{\sum_{i=1}^{N} r_{ik}\phi(x_i)}{\sum_{i=1}^{N} r_{ik}}$$

Consider, $||\phi(x_n) - \tilde{\mu_k}||^2$

$$\begin{aligned} ||\phi(x_n) - \tilde{\mu_k}||^2 &= (\phi(x_n) - \tilde{\mu_k})^T (\phi(x_n) - \tilde{\mu_k}) \\ &= \phi(x_n)^T \phi(x_n) - 2\tilde{\mu}^T \phi(x_n) + \tilde{\mu}^T \tilde{\mu} \\ &= \phi(x_n)^T \phi(x_n) - 2\frac{\sum_{i=1}^{N} r_{ik}\phi(x_i)^T \phi(x_n)}{\sum_{i=1}^{N} r_{ik}} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik} r_{jk}\phi(x_i)^T \phi(x_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik} r_{jk}} \end{aligned}$$

Lets assume that $n_k = \sum_{i=1}^{N} r_{ik}$, so that it simplifies to:

$$\begin{aligned} ||\phi(x_n) - \tilde{\mu_k}||^2 &= \phi(x_n)^T \phi(x_n) - 2\frac{\sum_{i=1}^{N} r_{ik}\phi(x_i)^T \phi(x_n)}{n_k} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik} r_{jk}\phi(x_i)^T \phi(x_j)}{n_k^2} \\ &= K(x_n, x_n) - 2\frac{\sum_{i=1}^{N} r_{ik} K(x_i, x_n)}{n_k} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik} r_{jk} K(x_i, x_j)}{n_k^2} \end{aligned}$$

We can express the Distortion function just in terms of kernel matrix as follows,

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

1.4 1 (c) 2

We compute the distance for all points x_n for each cluster and choose the minimum using above equation for $\tilde{D}$, where $n_k = \sum_{i=1}^{N} r_{ik}$, therefore membership assignment will be

$$r_{nk} = \begin{cases} 1 & k = \arg\min_k ||\phi(x_n) - \tilde{\mu_k}||_2^2 \\ 0 & \text{otherwise} \end{cases}$$

1.5 1 (c) 3

1) Randomly choose k points of N as cluster centroids[1..k]
2) Choose a kernel function (RBF,polynomial, sigmoid etc), and compute the kernel matrix K(i...N,j..N)
3) Now conmpute the distance $\tilde{D}$ as for each point x_n, with respect to k cluster $K(x_n, x_n) - 2\frac{\sum_{i=1}^{N} r_{ik}K(x_i,x_n)}{n_k} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ik}r_{jk}K(x_i,x_j)}{n_k^2}$
4) For each data point determine the membership ,compute matrix r_{nk}
5) update μ_k for new cluster centroid
6) Check for convergence , repeat from step 3)

2 Problem 2

2.1 2 (a) 1

Given

$$f(x|\theta_1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}} x^2 \text{ and, } f(x|\theta_2) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$

We can express max likelihood as follows:

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

differentiating with respect to α, for maximum likelihood

$$\frac{\partial L(x)}{\partial \alpha} = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}} x^2 - \frac{1}{\sqrt{\pi}} e^{-x^2}$$

We observe that the maximum likehood is independent of alpha and it dependant on the value of L. If $\frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}} x^2 > \frac{1}{\sqrt{\pi}} e^{-x^2}$, α will take part in increasing the likelihood , if both are equal then there is no impact of α. If $\frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}} x^2 < \frac{1}{\sqrt{\pi}} e^{-x^2}$, α will tend to zero.

3 Problem 3

3.1 3 (a)

Let z_i be a latent variable such that $z_i = 1$ if x_i is from the zero state (zero inflated state), and $z_i = 0$ if x_i is from the Poisson state (for zero truncated state). Let $z_i = 1$ with probability π, and $z_i = 0$ with probability $(1-\pi)\lambda$.

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

$$Z_i = \begin{cases} 1 & X_i \text{ is zero with } \pi_i \\ 0 & \text{if} X_i > 0 \text{ , } (1-\pi)e^{-\lambda} \end{cases}$$

Therefore,

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

Assuming I as indicator function of membership,

$$\begin{aligned} L((X,Z)|\theta) &= \prod_{x_i=0} \pi^{z_i} \times ((1-\pi)e^{-\lambda})^{1-z_i} \times \prod_{x_i>0} (1-\pi)e^{\frac{\lambda_i^x e^{-\lambda}}{x_i!}} \\ LL = \log L &= \sum_{I(x_i=0)} z_i \log(\pi) + (1-z_i)(\log(1-\pi) - \lambda) \\ &+ \sum_{I(x_i>0)} (\log(1-\pi) + (\lambda_i^{x_i}) - \lambda - \log(x_i!)) \end{aligned}$$

3.2 3 (b)

Say, $\theta = (\pi, \lambda)$, and θ_0 for the old parameter from previous iteration of the EM algorithm.
Consider E step

$$\begin{aligned} Q(\theta, \theta_0) &= \sum_z [P(Z|X,\theta) \log P((X,Z),\theta)] \\ &= \sum_{I(x_i=0)} E_{P(Z|X)}[z_i] \log(\pi) + (1 - E_{P(Z|X)}[z_i])\big(\log(1-\pi) - \lambda\big) \\ &+ \sum_{I(x_i>0)} \big(\log(1-\pi) + (\lambda_i^{x_i}) - \lambda - \log(x_i!)\big) \end{aligned}$$

Solving for $E_{P(Z|X_i)}[z_i]$

$$\begin{aligned} E_{P(Z|X_i)}[z_i] &= 0 \times p(Z_i = 0|X) + 1 \times p(Z_i = 1|X_i = 0) \\ &= \frac{p(X_i = 0|Z_i = 1)p(Z_i = 1)}{p(X_i = 0|Z_i = 0)p(Z_i = 0) + p(X_i = 0|Z_i = 1)p(Z_i = 1)} \\ &= \frac{\pi_0}{\pi_0 + (1-\pi_0)e^{-\lambda_0}} \end{aligned}$$

Now, we can re-write $Q(\theta, \theta_0)$

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

In M step, we will maximize Q to compute update for all parameters as follows:
Differentiate wrt λ

$$
\begin{aligned}
\frac{\partial Q}{\partial \lambda} &= 0 \\
&= \sum_{I(x_i=0)} (1 - E[z_i])(-1) + \sum_{I(x_i>0)} (\frac{x_i}{\lambda} - 1) = 0 \\
\implies \hat{\lambda} &= \frac{\sum_{I(x_i>0)} x_i}{n - \sum_{I(x_i=0)} E[z_i]} \\
\hat{\lambda} &= \frac{\sum_{I(x_i>0)} x_i}{n - \sum_{I(x_i=0)} \hat{z}_i} \\
\text{where } \hat{z} &= \frac{\pi_0}{\pi_0 + (1-\pi_0)e^{-\lambda_0}}
\end{aligned}
$$

Differentiate wrt π

$$
\begin{aligned}
\frac{\partial Q}{\partial \pi} &= 0 \\
&= \sum_{I(x_i=0)} (\frac{E[z_i]}{\pi} - \frac{1 - E[z_i]}{1-\pi}) - \sum_{I(x_i>0)} \frac{1}{1-\pi} = 0 \\
&= \sum_{I(x_i=0)} (\frac{E[z_i]}{\pi} + \frac{E[z_i]}{1-\pi}) - \frac{n}{1-\pi} = 0 \\
\implies \hat{\pi} &= \sum_{I(x_i=0)} \frac{\hat{z}_i}{n}
\end{aligned}
$$

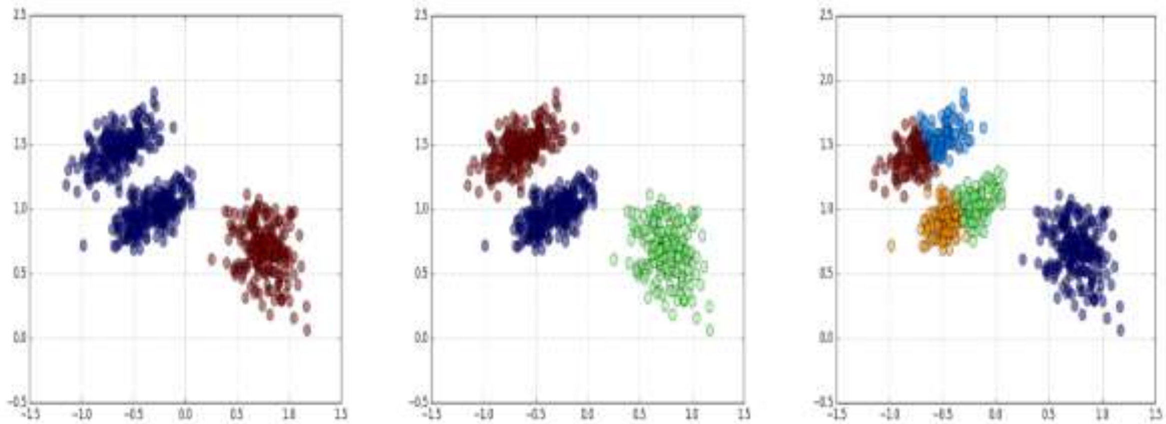
Therefore, the updates rules are :
$\hat{z}_1 = \frac{\pi_0}{\pi_0+(1-\pi_0)e^{-\lambda_0}}$, $\hat{\lambda}_1 = \frac{\sum_{I(x_i>0)} x_i}{n-\sum_{I(x_i=0)} \hat{z}_1}$, $\hat{\pi} = \sum_{I(x_i=0)} \frac{\hat{z}_i}{n}$

4. Programming

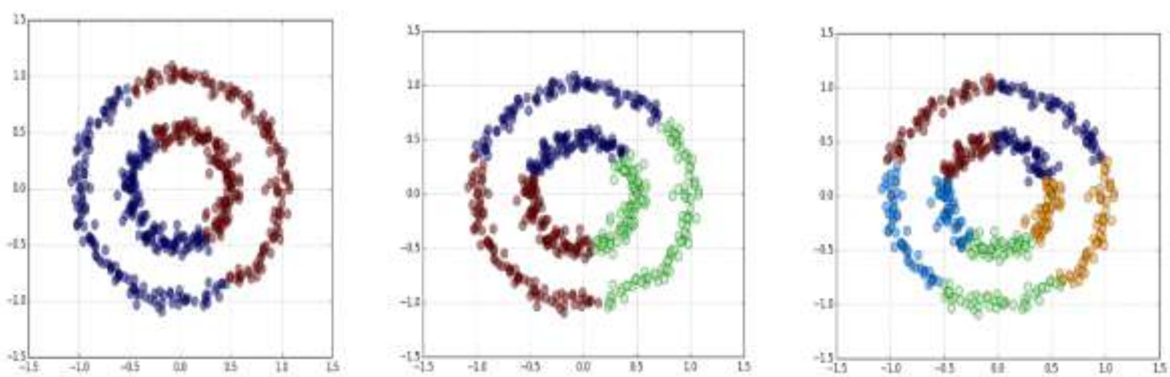
4.2 (a) Implemented k means till no change observed in the clusters assigned.

4.2.a.1

Blob plots for K=2, K=3 and K=5



Circle plots for K=2, K=3 and K=5



4.2. (b) The two circle as shown above are not linearly separable in the original space, and that's why it is divide into 2 half circles. K- means work on the linear separation of the data points. However, we can transform this into higher dimensional feature space where they might be separable and compute k-means in new feature space.

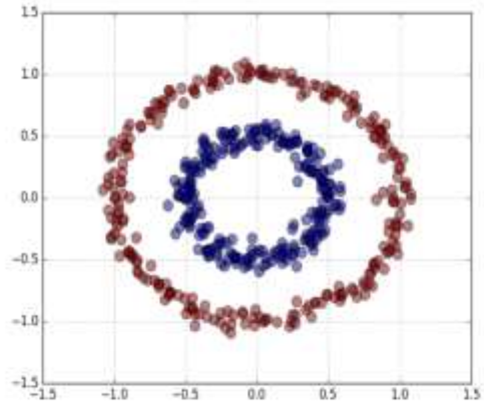
4.3 (a) Experimented with various kernel, as it takes time to converge.

RBF :- $K(x_i, x_j) = e^{(-\gamma||x_i - x_j||^2)}$ where γ = 50

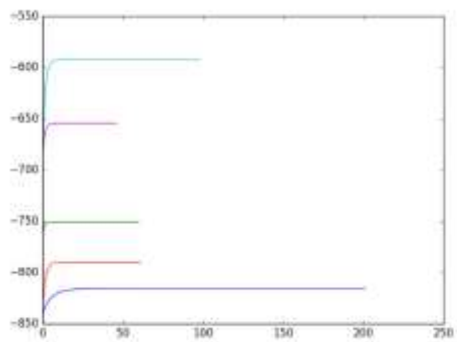
Polynomial :- $K(x_i, x_j) = (1 + x_i * x_j)^4$ where c=1 and d=4

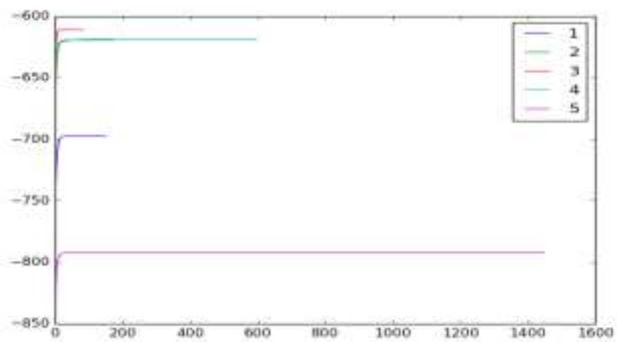
For other combination the output was observed to get stuck in the local minimum.

4.3. (b) Following plot was observed for the kernel k means with polynomial kernel , for k=2, c=1 and d=4

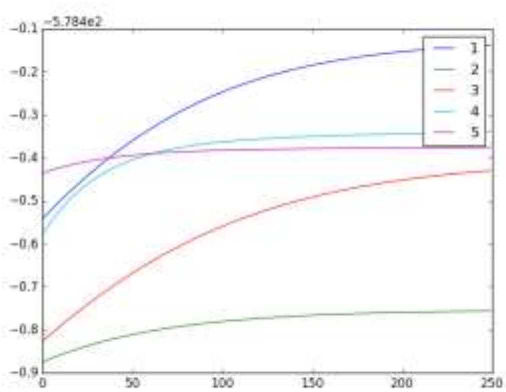


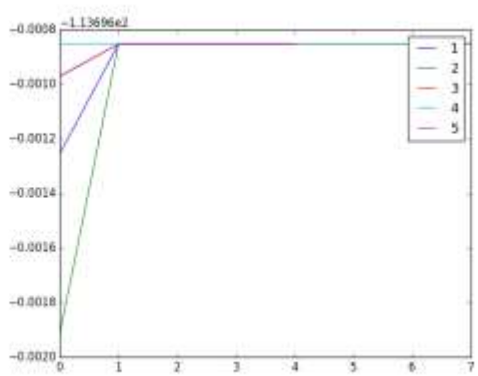
4.4 (a) When randomly initialized the clusters different graph were observed, and takes long time to converge.



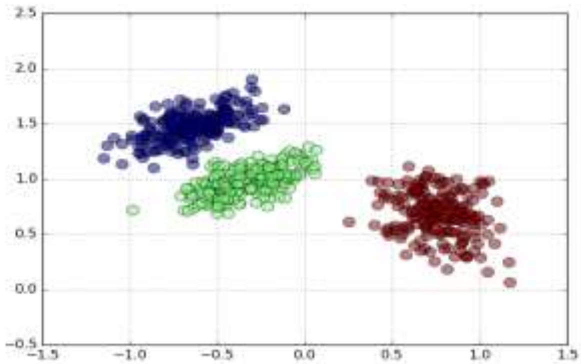


However, when initialized with output of k-mean , it converges very fast, as shown below.





4.4 (b) Best plot cluster assignments



Best Mean and covariance for the best log likelihood as shown below:

Mean 1= ([-0.63945121, 1.4745009]), Covariance 1= [[0.03595823, 0.01548446],

[0.01548446, 0.01938158]]

Mean 2= ([0.75895991, 0.6797701]), Covariance 2= [[0.02717078, -0.0084006],

[-0.0084006 , 0.04044207]]

Mean 3= ([-0.32583659, 0.97128509])], Covariance 3= [[0.03603558, 0.01465724],

[0.01465724, 0.0162877]]