Machine Learning Homework 5 Solutions

Name : Achyut Kulkarni

1.

a)

We know that,

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

Hence, once we differentiate the distortion function we get ;

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

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

$$2\mu_k \sum_{n=1}^{N} r_{nk} = 2\sum_{n=1}^{N}(r_{nk}x_n)$$

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

b) The distortion measure is given by:

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

Taking derivative of D w.r.t to μ_k

$$\frac{\partial D}{\partial \mu_k} = \frac{\partial}{\partial \mu_k}\{D = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|_1\} = 0$$

Now, as per definition $\frac{\partial \|z\|}{\partial z} = sign(z)$

such that

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

$$sign(z) = 1 \quad if \, z > 0$$

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

Hence, for each cluster k ,

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

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

Where $n_1 = number\ of\ data\ points\ such\ that\ (x_n - \mu_k) > 0$ for cluster k

$n_2 = number\ of\ data\ points\ such\ that\ (x_n - \mu_k) < 0$ for cluster k

Hence, $n_1 = n_2$

$$=> n\,(x_n > \mu_k) = n\,(x_n < \mu_k)$$

The above equation can only hold true if μ_k is an element-wise median of all data points assigned to the k-th cluster, such that objective D is minimized.

c)

i) Now we have the new distortion measure D as :

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

We know that ;

$$\mu'_k = \frac{\sum_{i=1}^{N} r_{ik}\varphi(x_i)}{\sum_{i=1}^{N} r_{ik}}$$ hence expanding the above term as we get ;

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}[\ \varphi(x_n)^2 + \mu'^2_k - 2\varphi(x_n)\mu'_k]$$

This can be put in vector form as

$\varphi(x_n)^2 = \Phi(x_n)^T\Phi(x_n)$ and substituting μ'_k we get :

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

But we know that the kernel function is expressed as ;

$$\Phi(x_i)^T\Phi(x_j) = K(x_i, x_j)$$

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

ii) For every point we calculate the distance using the above function to each cluster and assign the point to the cluster(C) which has the shortest (min) distance from the point.

Hence;

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

iii)

Algorithm for Kernel K-means clustering:

a)Let X = $\{x_1, x_2, x_3, ..., x_n\}$ be the set of data points and *'K'* be the number of clusters.

b)Compute the distance of each data point and the cluster center in the transformed space using:

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

where,

$$K(x_i,\ x_j) = \Phi(x_i)^T\Phi(x_j)$$

c) Assign data point to that cluster center whose distance is minimum.

d) Until data points are re-assigned repeat from step b).

2) Gaussian Mixture Model

Gaussian Distributions:

$f(x_1|\theta_1) = 1/\sqrt{2\pi}\ .\ exp(-x_1^2/2)$ with $\mu_1 = 0$ and $\sigma_1^2 = 1$

$f(x_1|\theta_2) = 1/\sqrt{\pi}\ \ .\ exp(-x_1^2)$ with $\mu_2 = 0$ and $\sigma_2^2 = 0.5$

For the given sample x_1 : $P(x_1) = P(\theta_1|x_1) + P(\theta_2|x_1)$

$f(\theta_1) = \alpha$ and $f(\theta_2) = 1 - \alpha$

$$P(x_1) = P(\theta_1\,|\,x_1) + P(\theta_2\,|\,x_1) = \alpha/\sqrt{2\pi}\ .\ exp(-x_1^2/2) + (1-\alpha)/\sqrt{\pi}\ .\ exp(-x_1^2)$$
$$= \alpha\,(\ 1/\sqrt{2\pi}\ .\ exp(-x_1^2/2) - 1/\sqrt{\pi}\ .\ exp(-x_1^2)\ ) + 1/\sqrt{\pi}\ .\ exp(-x_1^2)$$

$P(x_1)$ is a linear function of α of form $a\ .\ \alpha + b$ where $0 \le \alpha \le 1$ and the sign of the slope determines whether $f(x_1|\theta_1) > f(x_1|\theta_2)$ or not.

Where slope (a) is $1/\sqrt{2\pi}\ .\ exp(-x_1^2/2) - 1/\sqrt{\pi}\ .\ exp(-x_1^2)$

If slope is non-negative $\alpha = 1$,

$1/\sqrt{2\pi}\ .\ exp(-x_1^2/2) \ \ge\ 1/\sqrt{\pi}\ .\ exp(-x_1^2)$

$exp(x_1^2/2) \ \ge\ \sqrt{2}$

$x_1^2 \ \ge\ log(2)$

Similarly If slope is negative $x_1^2 \ < log(2)$, $\alpha = 0$

$$\alpha^* = \left\{ x_1^2 \ < log(2),\ \alpha = 0 \ \ and \ x_1^2 \ \geq log(2),\ \alpha = 1 \right\}$$

3) EM algorithm :

Given Zero-inflated Poisson Distribution:

$$p(x_i) = \left\{ \pi + (1-\pi)e^{-\lambda} \ if \ x_i = 0 \ ; \ (1-\pi)\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \ if \ \ x_i > 0 \right\}$$

a) The hidden random variable here are the values associated with the zero counts in the data . That is we can describe the Random Variable z_i as having either ON (Zero state) or OFF(Poisson state)

Thus we can say :

$$z_i = \left\{ 1 \ if \ x_i = 0; \quad 0 \ \ if \ x_i > 0 \right\}$$

$$log(p(x_i)) = log \left\{ \pi + (1-\pi)e^{-\lambda} \ if \ x_i = 0 \ ; \ (1-\pi)\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \ if \ \ x_i > 0 \right\}$$

$$l(\theta) = \sum_{xi=0} log(\pi + (1-\pi)e^{-\lambda}) + \sum_{xi>0} log(1-\pi) + \sum_{xi>0} x_i log\lambda - \sum_{xi>0} \lambda - \sum_{xi>0} log(x_i!)$$

This is incomplete as difficult to calculate MLE hence,

Now with hidden variable z_i we have ; the joint distribution as

$$(p(x_i, z_i)) = \left\{ z_i\pi + (1-z_i)(1-\pi)e^{-\lambda} \ if \ x_i = 0 \ ; \ (1-z_i)(1-\pi)\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \ if \ \ x_i > 0 \right\}$$

Now L1 = $\prod_{xi=0} (z_i\pi + (1-z_i)(1-\pi)e^{-\lambda})$ and L2 = $\prod_{xi>0} (1-z_i)(1-\pi)\frac{\lambda^{x_i} e^{-\lambda}}{x_i!})$

Applying log on both equations and adding we get ;

$$log(p(\ ,z_i)) = log \left\{ z_i\pi + (1-z_i)(1-\pi)e^{-\lambda} \ if \ x_i = 0 \ ; \ (1-z_i)(1-\pi)\frac{\lambda^{x_i}e^{-\lambda}}{x_i!} \ if \ x_i > 0 \right\}$$

$$log(p(x_i ,z_i)) = \sum_{i=0}^{n} z_i log(\pi) + (1-z_i)log(1-\pi) - \lambda + \sum_{i>0}^{n} log(1-\pi) + (x_i)log\lambda i - \lambda - log(x_i!))$$

This is the complete log likelihood !!

Now let calculate EM steps :

E Step :

$$Q(\theta,\theta_0) = \sum_{i=0}^{n} E_{p(z|x)}z_i log(\pi) + (1-E_{p(z|x)}z_i)log(1-\pi)-\lambda + \sum_{i>0}^{n} log(1-\pi) + (x_i)log\lambda i - \lambda - log(x_i!))$$

$$E_{p(z|x)}z_i = 0 * P(Z_i = 1 \,|\, x) + 1 * P(Z_i = 0 \,|\, x_i = 0)$$

$$= \frac{P(X_i = 0 \,|\, Z_i = 1) * P(Z = 1_i)}{P(X_i = 0 \,|\, Z_i = 1) * P(Z = 1_i) + P(X_i = 0 \,|\, Z_i = 0) * P(Z_i = 0)}$$

$$= \frac{\pi_0}{\pi_0 + (1-\pi_0)e^{-\lambda 0}}$$

Hence, the above expectation becomes :

$$Q(\theta,\theta_0) = \sum_{i=0}^{n} \frac{\pi_0}{\pi_0 + (1-\pi_0)e^{-\lambda 0}} log(\pi) + (1-\frac{\pi_0}{\pi_0 + (1-\pi_0)e^{-\lambda 0}})log(1-\pi)-\lambda +$$

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

M Step :

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

$$=> \sum_{i=0}^{n}(1-E_{p(z|x)}z_i)(-1) + \sum_{i>0}^{n}(\frac{x_i}{\lambda} - 1) = 0$$

$$=> \lambda' = \frac{\sum_{i>0}^{n}(x_i)}{n - \sum_{i=0}^{n}(E_{p(z|x)}z_i)}$$

$$=> \lambda' = \frac{\sum_{i>0}^{n}(x_i)}{n - \sum_{i=0}^{n}(Z_i')} where \ Z_i' = \frac{\pi_0}{\pi_0 + (1-\pi_0)e^{-\lambda 0}}$$

ii)

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

$$\sum_{i=0}^{n}\left(\frac{E[Z_i]}{\pi} - \frac{1-E[Z_i]}{1-\pi}\right) - \sum_{i>0}^{n}\left(\frac{1}{1-\pi}\right) = 0$$

$$\sum_{i=0}^{n}\left(\frac{E[Z_i]}{\pi} + \frac{E[Z_i]}{1-\pi}\right) - \left(\frac{n}{1-\pi}\right) = 0$$

$$=> \pi' = \sum_{i=0}^{n}\left(\frac{Z'_i}{n}\right)$$

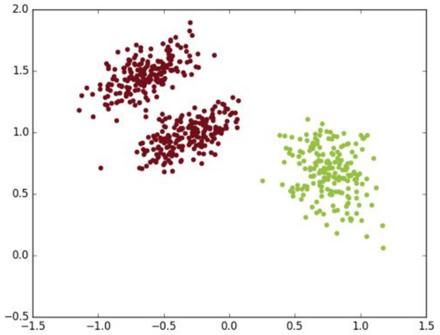
And we know that :

$$z'_i = \frac{\pi_0}{\pi_0 + (1-\pi_0)e^{-\lambda 0}}$$

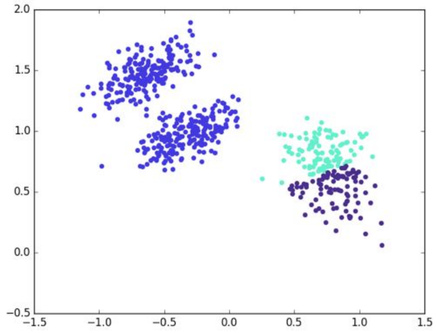
Programming :

4.2 Implement k-means :

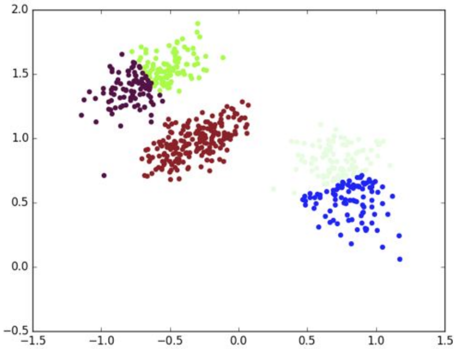
a)



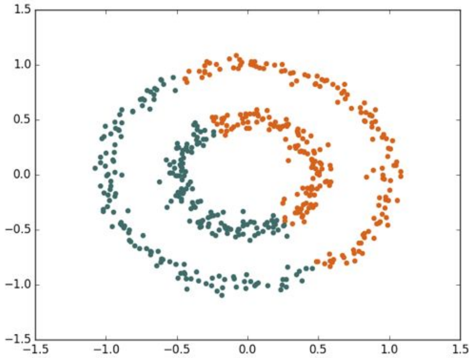
K = 2



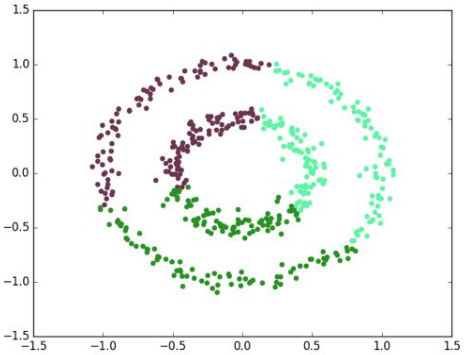
K = 3



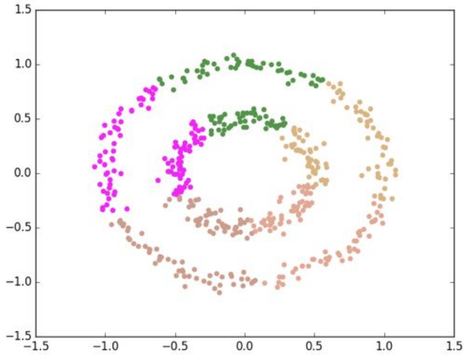
K = 5



K = 2



K = 3



K = 5

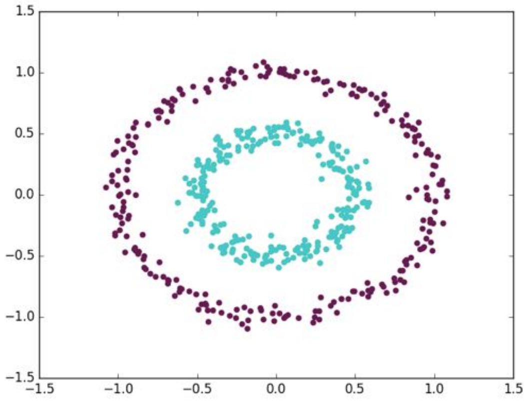
b) The K-means algorithm fails to cluster the circle data points as K-means cluster is basically a linear separator that is it tries to cluster data points which are linearly separable and does not if the data points are not linearly separable

4.3 Implement kernel k-means

A) The polynomial kernel used is :

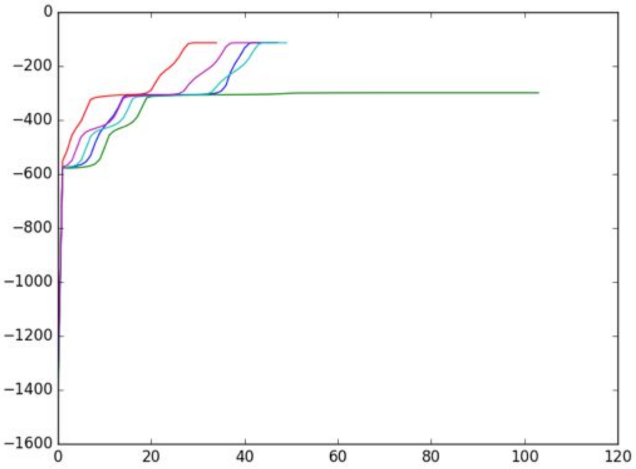
Kernel: $K(x_{i,} x_j) = \| x_i . x_j \| + 15 . \| x_i \|_2^2 . \| x_j \|_2^2$

b) The transformed space is separated by K-means ;
Hence, Kernelized K-means successfully clusters the set of points correctly .



4.4 Implement Gaussian Mixture Model

a)



b)

Best Parameters:

$\mu_0 = (0.75896,\ 0.67977)$

$\Sigma_0 = [[\ 0.02717056,\ -0.00840045],\ [-0.00840045,\ \ 0.04044199]]$

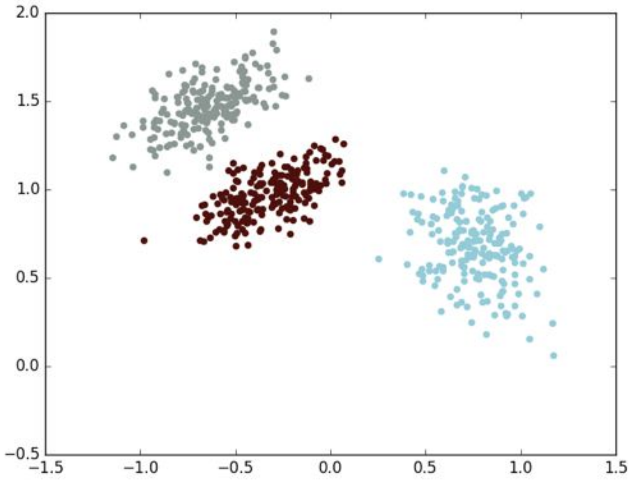
$\pi_0 = 0.33333517664398987$

$\mu_1 = (-0.325922,\ 0.971336)$

$\pi_1 = 0.3355079325540168$
$\Sigma_1 = [[\ 0.03604972,\ \ 0.01463863],\ [\ 0.01463863,\ \ 0.01629125]]$

$\mu_2 = (-0.639463,\ 1.474608)$
$\pi_2 = 0.33115689080199284$
$\Sigma_2 = [[\ 0.03596772,\ \ 0.01549327],\ [\ 0.01549327,\ \ 0.01935129]]$

The most likely cluster assignment


2.0
1.5
1.0
0.5
0.0
−0.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5