clustering and EM

Anjali Krishna Prasad

November 2016

1 Clustering

1a) We know that

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

Lets assume X0 is in scalar space:

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

To minimize, we will euate it to 0 Lets assume X0 is in scalar space:

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

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

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

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

This is this sum of all the elements in the cluster k:

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

This is the number of elements in the cluster k

$$\sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}$$

So it is clearly mean of the elements in the kth cluster.It is the mean of all data points assigned to the cluster k, for any k, then the objective D is minimized.

1b) Suppose that the set x has n elements

$$x_1 < x_2 < .... < x_n$$

and

$$x < x_1$$

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

$$f(x) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} |x_n - x| = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} (x_n - x)$$

As x increases , each term in the above equation decreases until x reaches x1. So

$$f(x_1) < f(x) for all x < x_1$$

Now suppose

$$x_j <= x <= x + d <= x_{j+1}$$

then

$$f(x+d) = \sum_{n=1}^{j}\sum_{k=1}^{K} r_{nk}(x + d - x_n) + \sum_{n=j+1}^{N}\sum_{k=1}^{K} r_{nk}(x_n - (x+d))$$

$$= dj + \sum_{n=1}^{j}\sum_{k=1}^{K} r_{nk}(x - x_n) - d(N-j) + \sum_{n=j+1}^{N}\sum_{k=1}^{K} r_{nk}(x_n - x)$$

$$= d(2j - N) + \sum_{n=1}^{j}\sum_{k=1}^{K} r_{nk}(x - x_n) + \sum_{n=j+1}^{N}\sum_{k=1}^{K} r_{nk}(x_n - x)$$

so

$$f(x+d) - f(x) = d(2j - N)$$

This is negative if

$$2j < N$$

zero if

$$2j = N$$

and positive if

$$2j > N$$

Thus on the interval

$$[s_j, s_{j+1}]$$

f(x) is decreasing if 2j is less than N, is constant if 2j is equal to N and negative if 2j is greater than N

$$2j = N := j = N/2$$

Therefore in this case we will use median to minimize distortion function.

1 Rewrite the center as

$$\mu_k = \sum_{k=1}^{K} \gamma_k \phi(x_j)$$

Now look at the squared distance between the transforrmed data and the centers:

$$\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||\phi(x_n) - \mu_k||^2 = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||\phi(x_n) - \phi(x_k)||^2$$

where

$$\phi(x_l) = \sum_{nc_k}^{K} \phi(x_n)/N_k$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||\phi(x_n) - \phi(x_k)||^2 = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}[(\phi(x_n) - \sum_{lCc_k} \phi(x_l)] * [(\phi(x_n) - \sum_{lCc_k} \phi(x_l)]$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}[k(x_n, x_n) - 2\sum_{lCc_k} k(x_n, x_l) + \sum_{l',lCc_k} k(x_l, x_l')]$$

The equation of assigning a data point to its cluster.

$$argmin_k \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}[k(x_n,x_n) - 2\sum_{lCc_k} k(x_n,x_l) + \sum_{l',lCc_k} k(x_l,x_l')]$$

We find out the euclidean distance from the point to the mean , in this case the mean of the kernel space and we select the minimum distance from the mean of k clusters and assign this minimum k to the point.

Pseudo-code of the complete kernel K-means algorithm including initialization of cluster centers.

```
Initialize means to k random points where k is number of clusters
1: for all points x_n n=1,2,...,N do
2:      for all clusters c_i i=1,2,...,k do
3:          compute the following using equation in the previous step
                $||\phi(x_n) = m_i||^2$
4:      end for
5:      Find
                $c^*(x_n) = argmin_i(||\phi(x_n) - m_i||^2)$
6: end for
7:  for all clusters c_i i=1,2,...,k do
8:      Update cluster
                $C_i = (x_n|c^*(x_n) = i)$
9: end for
10:if converged then
11:     return final clusters
12:else
13:     Go to step 1
14:end if
```

2 Gaussian Mixture Model

Thus we can write the likelihood function

$$L = p(x_1|\alpha) = log(P(x|\theta_1)\alpha + P(x|\theta_2)(1-\alpha))$$

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

Determine the maximum likelihood estimate of alpha. We can write the likelihood as follows:

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

Thus, we see that the likelihood is simply a linear function of alpha where the sign of the slope is determined by which Gaussian produces the larger response. Since we know that

$$0 <= \alpha <= 1$$

, this tells us that if the slope is positive that we should choose = 1 and otherwise if the slope is negative we should choose alpha = 0. Using straightforward algebra one can show that the slope is positive whenever

$$x_2^1 >= log2$$

and we should set alpha = 1 otherwise set alpha = 0.

3 EM algorithm

3a)we have if xi=0

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

if

$$x_i > 0$$

then

$$p(x_i) = (1-\pi)\frac{\lambda^{x_i}e^{-\lambda}}{x_i!}$$

A proper hidden variable zi for the observations is that: When z = 1 ,

$$p(z_i) = \pi$$

when zi=0,

$$p(z_i) = (1-\pi)$$

For x greater than 0, we dont have a z as it doesnt depend on z.
E-Step: Compute

$$Q(\theta|\theta^{(t)}) = E_{p(z|x,\theta^{(t)})}[logp(x,z|\theta)]$$

$$Q(\theta|\theta^{(t)}) = E[\ log\prod_{i=1}^{n}(\sum_{i:x_i=0} 1.\pi)^{z_i} + (\sum_{i:x_i=0} e^{-\lambda}.(1-\pi))^{1-z_i} + \sum_{i:x_i>0}(1-\pi)\frac{\lambda^{x_i}e^{-\lambda}}{x_i!}]$$

$$= \sum_{i=1}^{n} E[z_i|x_i,\theta^{(t)}][\sum_{i:x_i=0}(log\pi)] + [1 - E[z_i|x_i,\theta^{(t)}][\sum_{i:x_i=0}[log(1-\pi)-\lambda]] + \sum_{i:x_i>0} log(1-\pi) + log\lambda^{x_i} - \lambda - log(x_i)!$$

Lets call the above equation 1.
Next we compute

$$E[z_i|x_i,\theta^{(t)}] = p(z_i = 1|x_i,\theta^{(t)}) = \frac{(x_i|z_i,\theta^{(t)})p(z_i = 1|\theta^{(t)})}{p(x_i|\theta^{(t)})}$$

$$E[z_i|x_i,\theta^{(t)}] = \frac{\pi}{\pi + (1-\pi)e^{-\lambda}}$$

M-step:

$$\theta^{t+1} = argmax_{\theta}E_{p(z|x,\theta^{(t)})}[logp(x,z|\theta)]$$

Maximizing

$$Q(\theta|\theta(t))w.r.t.\theta$$

yields the update equations. Differentiating equation 1 wrt pi.

$$\frac{\partial Q(\theta|\theta^{(t)})}{\partial\pi} = \sum_{i=1}^{n} E[z_i|x_i,\theta^{(t)}][\sum_{i:x_i=0}\frac{1}{\pi}] + [1 - E[z_i|x_i,\theta^{(t)}][\sum_{i:x_i=0}[-\frac{1}{1-\pi}] + \sum_{i:x_i>0}\frac{-1}{(1-\pi)} = 0$$

$$\pi = \sum_{x_i=0}\frac{E[z_i|x_i,\theta^{(t)}]}{n}$$

Differentiating equation 1 wrt lambda.

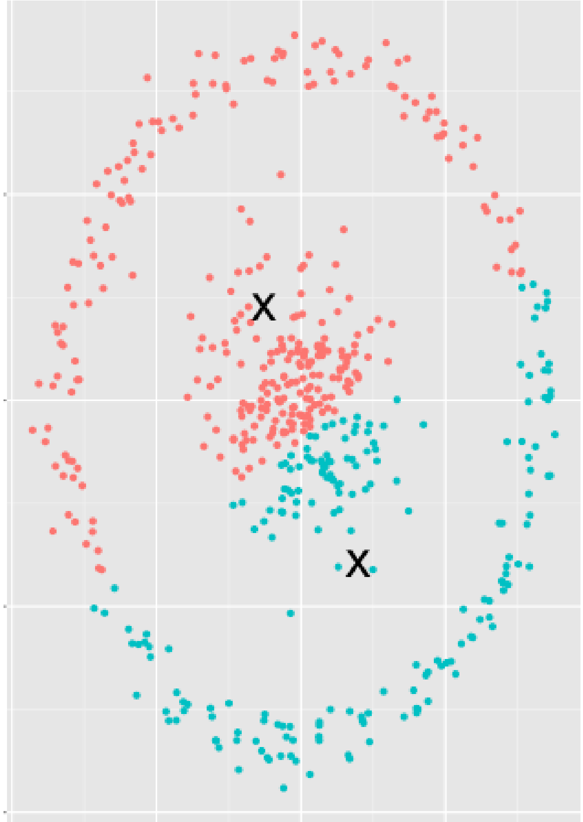
$$\frac{\partial Q(\theta|\theta^{(t)})}{\partial\lambda} = \sum_{i=1}^{n}[1 - E[z_i|x_i,\theta^{(t)}][\sum_{i:x_i=0}[-1] + \sum_{i:x_i>0}\frac{x_i}{\lambda} - 1 = 0$$

$$\lambda = \frac{\sum_{i:x_i>0} x_i}{(n - \sum_{i:x_i=0} E[z_i|x_i,\theta^{(t)}])}$$

4 Programming

4.2)

K-mean algorithm works on the concept of finding clusters with the minimum euclidean distance of each point in the cluster from its mean. As there are 2 circles, with more concentration in the circle, the means will be around the inner circle. So there will be points on the outer circle which will be closser to the other mean heace leading to incorrect clustering.

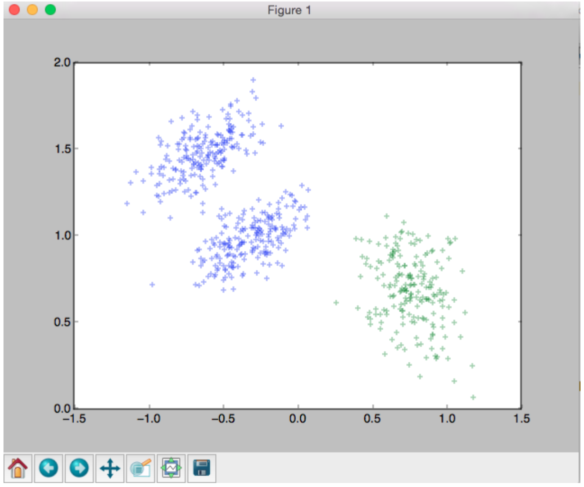


Attached image shows in details. k=2

```
4.3) a) I have chosen polynomial kernel
```

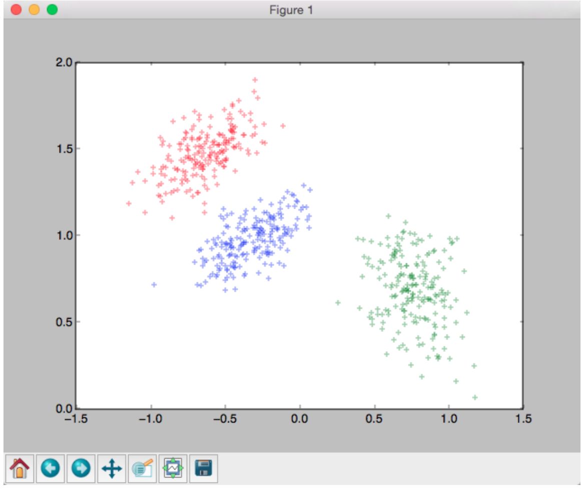
$$k(x_1, x_2) = x_1^2 + x_2^2 + x_1^2 x_2^2$$

```
4.4)Best iteration 5
covariance
[array([[ 0.03604954,  0.01463887],
    [ 0.01463887,  0.0162912 ]]),
array([[ 0.02717056, -0.00840045],
    [-0.00840045,  0.040442  ]]),
array([[ 0.0359676 ,  0.01549315],
    [ 0.01549315,  0.01935168]])]
mean
[[-0.32592106  0.97133574]
 [ 0.75896032  0.67976982]
 [-0.6394629   1.4746064 ]]
```



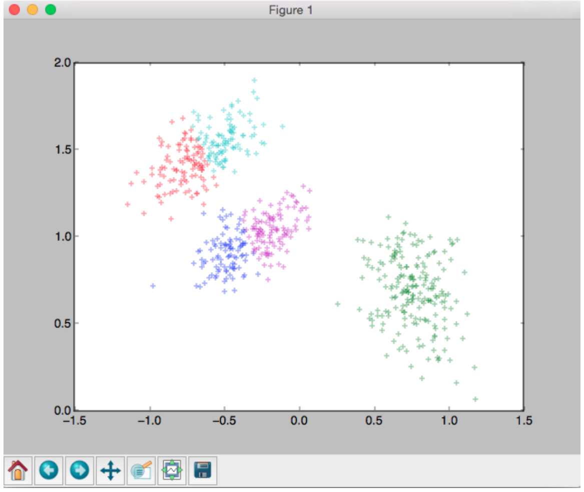
.5

Figure 1: k=2



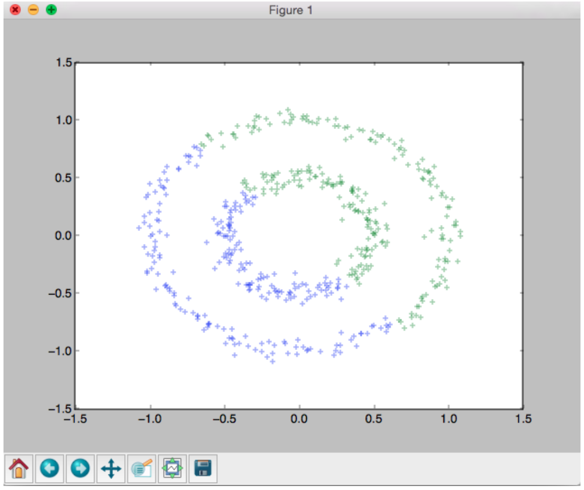
.5

Figure 2: k=3



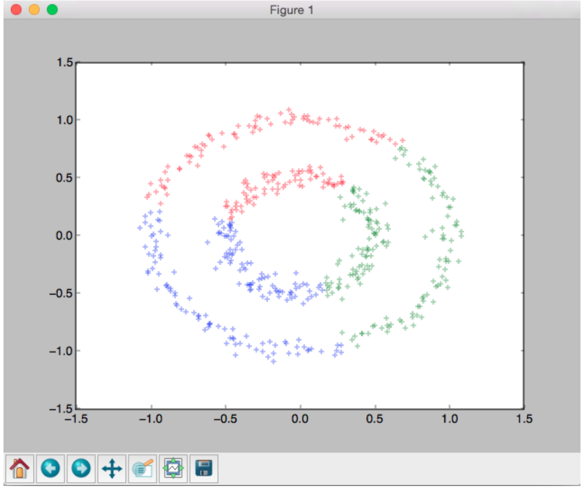
.5

Figure 3: k=5



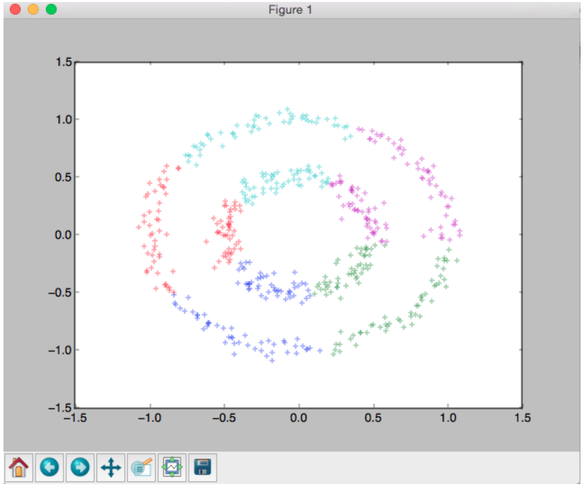
.5

Figure 5: k=2



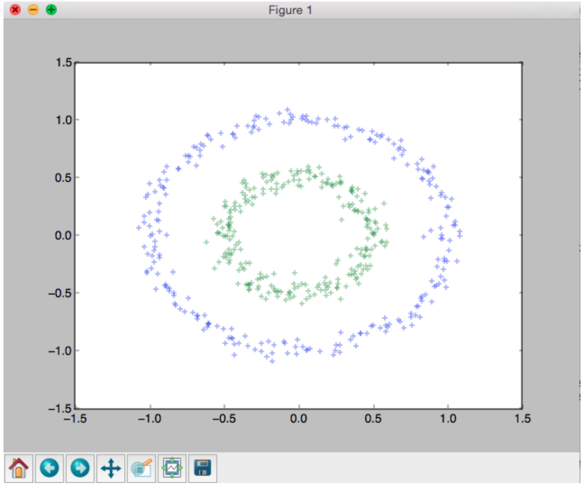
.5

Figure 6: k=3



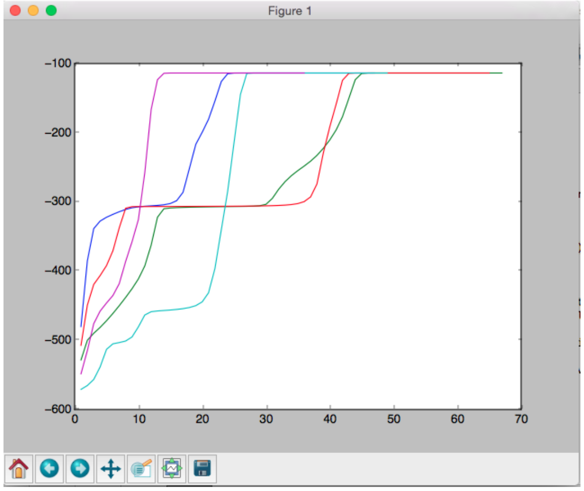
.5

Figure 7: k=5



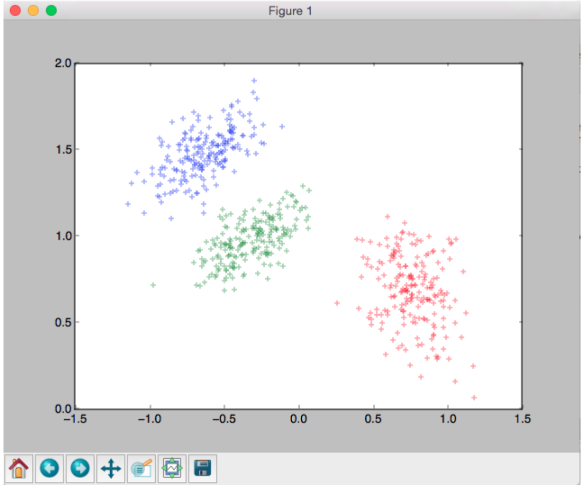
.5

Figure 9: kernel k mean with k = 2



.5

Figure 10: log likelihood after running 5 times



.5

Figure 12: EM for GMM