Clustering on c2 Dataset

Zitong Lian z.lian@student.tue.nl 0979901
Boshen Lyu b.lyu@student.tue.nl 0976857
We claim that we contribute equally and cooperate happily

Exercise 1:

Usage:

By running "dataloader_1b.py", users can choose from initializing methods and clustering. Number of clusters is also able to be set. Then some clustering/initializing algorithm including
1. the first k points of P;
2. k points of P picked uniformly at random;
3. K-Means++;
4. Gonzales algorithm
is implemented on c2.txt data.Results are given in a figure, where each cluster has its color, and whose title tells cost per run.

Performance:

Experimental results for different values of $k = \{3, 4, 5\}$ are given below:
The result of first k points of P is shown as figure 1:
The result of the k points of P picked uniformly at random is shown as figure2:

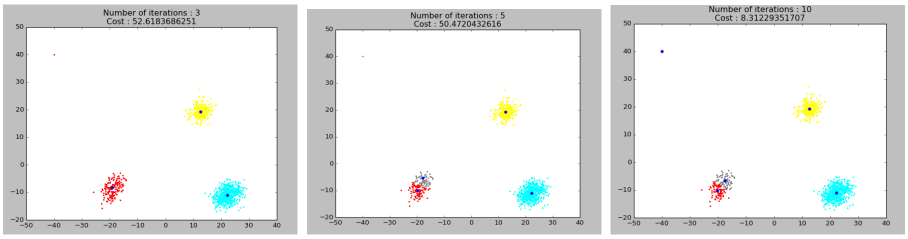

Figure 1: k =3, k = 4, k = 5

The result of K-Means++ is shown as figure3:
The result of Gonzales algorithm is shown as figure4:

For each fixed k, development of the k-means cost can be seen in the log:
As for "the first k points of P " and "Gonzales algorithm" initial method, the result are same when we run several times:
As to K-Means++ and random k points of P picked uniformly at random, samples of five

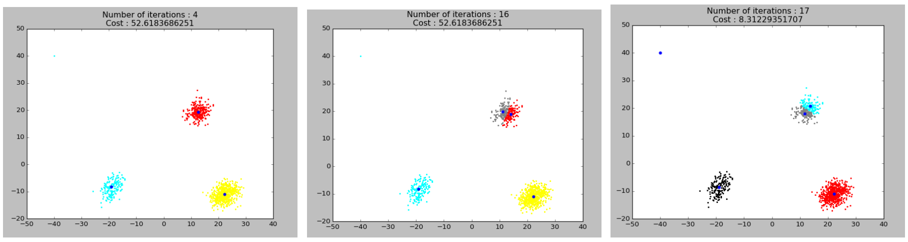

Figure 2: k =3, k = 4, k = 5

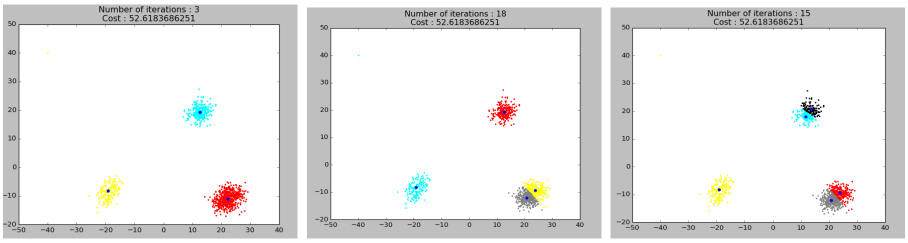

Figure 3: k =3, k = 4, k = 5

runs are provided, with standard deviation and average given below:

The thing we put pay most attention on is the cost. As we can see, for initial method 1 to 4, the cost decrease as the count of center increase. For method 1 and 4,when k = 3 and 4, the cost of the former is greater than the later. However when we pick 5 centers, method 1 descent significantly and become lower than method 4. For the method 2 and 3, we run several times to get their average and standard deviation. Both result illustrate the same trend, when the centers increase, the average decrease and the standard deviation decrease. We notice that

method k	3	4	5
method 1	52.6184	50.4720	8.3123
method 4	39.6823	14.9354	13.1609

Table 1: cost for method 1 and 4

k	run1	run2	run3	run4	run5	average	standard deviation
3	52.6183	51.0507	52.6183	52.6183	52.6183	52.3048	0.7011
4	50.472	52.6183	52.6183	51.9271	52.6183	52.0508	0.9319
5	52.6183	50.5425	8.3122	52.6183	52.6183	43.3419	19.6028

Table 2: cost for method 2

k	run1	run2	run3	run4	run5	average	standard deviation
3	52.6183	52.6183	52.6183	52.6183	52.6183	52.6183	0
4	52.6183	52.6183	52.6183	52.6183	8.3122	45.2334	18.0879
5	50.4090	50.4090	52.6183	8.3122	50.5425	42.4582	19.1113

Table 3: cost for method 3

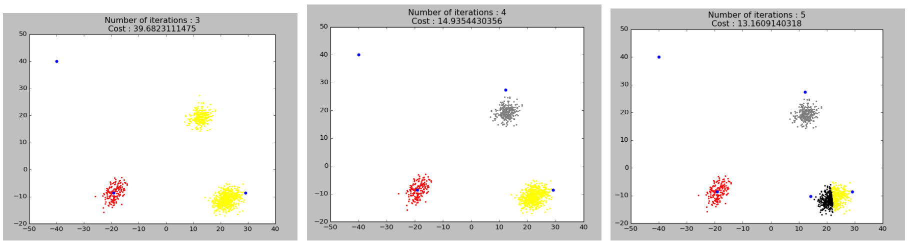

Figure 4: k =3, k = 4, k = 5

for method 3, if we pick 3 clusters, the cost keep absolutely steady. In conclusion, in the interval from 3 to 5, the more clusters we pick the lower cost we can get.

Exercise 2:

Assume we are in the case d = 1. we can get the partical derivative of $\phi(U,b)$ by

$$\frac{\partial\phi(U,b)}{\partial b} = \sum_{p_i \in U} -2 * (p_i - b)$$

let $\frac{\partial\phi(U,b)}{\partial b} = 0$, we can get the minimum of $\phi(U,b)$, then

$$b = \frac{1}{n}\sum_{p_i \in U} p_i$$

So, we have proved in $d = 1$

when $d > 1$, Let x is the center, p_{ij} means the jth dimension of vector p_i and b_j means the center in jth dimension, then we have :

$$\phi(U,b) = \sum_{i=1}^{n}(p_{ij} - b_j)^2$$

$$\frac{\partial\phi(U,b)}{\partial b} = \sum_{p_i \in U} -2 * (p_{ij} - b_j)$$

To maximize $\phi(U,b)$, $\partial\phi(U,b)$ shall be zero,we can get solution:

$$b_j = \frac{1}{n}\sum_{p_i \in U} p_i j$$

Because, in each dimension j, we have :

$$b = \frac{1}{n}\sum_{p_i \in U} p_i$$

Thus the barycenter of all points minimizes ϕ.

Exercise 3:

No it doesn't. We disapprove it by giving counterexample as followed: Suppose p,q,r are all unit vectors. Angle between p and q is $\alpha = \pi/4$, angle between q and r is $\beta = \pi/4$, and angle between p and r is $\gamma = \pi/2$. Thus $d_{cos}(p,q) = 1 - cos(\alpha) = 1 - \sqrt{2}/2$. Similarly, $d_{cos}(q,r) = 1 - \sqrt{2}/2$, and $d_{cos}(p,r) = 1$. Thus $d_{cos}(p,q) + d_{cos}(q,r) = 2 - \sqrt{2} < 1 = d_{cos}(p,r)$, which contradict to triangle inequality, and it makes the disapprove.

Exercise 4:

(a) We would expect that the K-Means method is more sensible to noise.

Suppose our outlier is p, and q represents one of the majority that stay close to each other while far away from p. When choosing center point for all points, K-Means takes the maximum distance between any point and the center, resulting that p influences much while the center is chosen far to it,since the influence inceases exponentially when the distance acsends. However, K-Median average all distances, and influence of p is not that significant, which means this method is more robust to noise.

(b) Instead of we compute the average to update the center of cluster in K-means method, we compute the median.

For this question, I also run those four initial method:

k = 4

As we can see, the points are in high dimension, we cannot observe intuitively, but we can still compare the cost. We draw a table as follow to make it easy to compare. We also run the second and third method multiple(5) times to get the average.

When k =4, we find when we use the first k points of P as initial centers, we get the minimal cost. It looks like that the first k elements get the appropriate positions by coincidence.

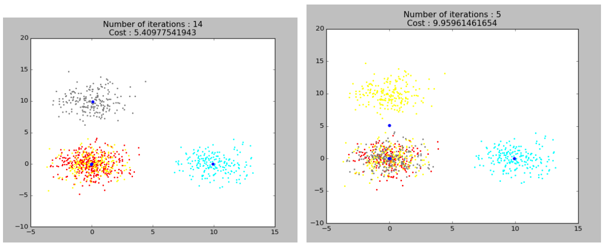

Figure 5: method 1,2

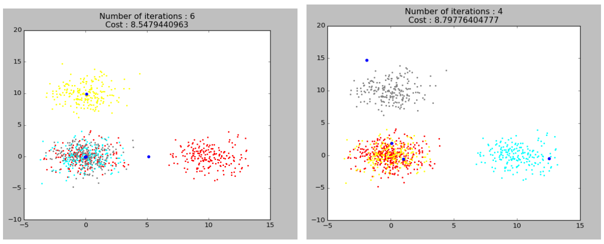

Figure 6: method 3,4

method	average	standard deviation
method 1	5.4098	
method 2	8.9321	2.3282
method 3	9.8494	2.6804
method 4	8.7978	

Table 4: cost for method 1 and 4