Chapter 7 - Clustering

Students may discuss questions together but must generate final solutions alone. I.e., you may help each other learn but not by providing answers.

Undergraduate students - do all **except** yellow highlighted. Graduate students. Do all **including** yellow highlighted. Ignore numbers in square brackets.

k-means questions

1. #1

1. 220 data vectors → 216

32 components

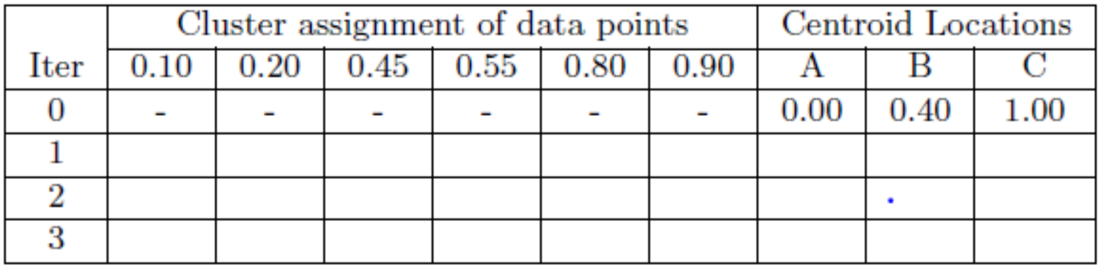
4 byte

Before : $220 \times 32 \times 4 = 28160$

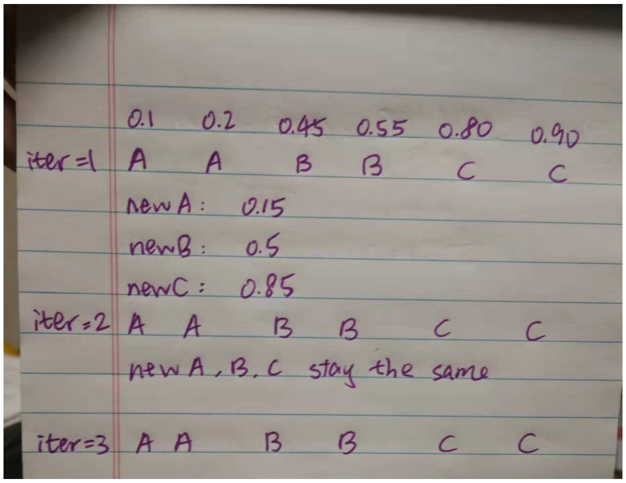
After : $216 \times 32 \times 4 = 27648$

2. [5] Consider the following set of one-dimensional points: f0.1, 0.2, 0.45, 0.55, 0.8, 0.9g. All the points are located in the range between [0,1].

   (a) Suppose we apply kmeans clustering to obtain three clusters, A, B, and C. If the initial centroids are located at {0, 0.4, 1}, respectively, show the cluster assignments and locations of the centroids after the first three iterations by lling out the following table.



	Cluster assignment of data points						Centroid Locations		
Iter	0.10	0.20	0.45	0.55	0.80	0.90	A	B	C
0	-	-	-	-	-	-	0.00	0.40	1.00
1									
2									
3									



	0.1	0.2	0.45	0.55	0.80	0.90
iter=1	A	A	B	B	C	C

new A : 0.15

new B : 0.5

new C : 0.85

iter=2	A	A	B	B	C	C

new A, B, C stay the same

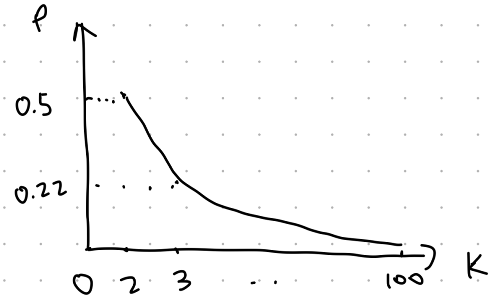
iter=3	A	A	B	B	C	C

(b) For the dataset given in part (a), is it possible to obtain empty clusters? If possible, what are the values of the initial centroids? If not, state why.

Yes, it is possible to obtain empty clusters. It happens when all the values of centroids are greater than the max number or smaller than the min number in our dataset.

3. #4a

3. 4(a) k = 2 to K=100



$P = \frac{K!}{K^k}$

when $k=2$, $p = \frac{2!}{2^2} = \frac{2}{4} = \frac{1}{2}$

when $k=3$, $p = \frac{3!}{3^3} = \frac{6}{27} = 0.22$

4. #13. There are two questions in #13. Consider them (a) and (b) when you answer.
   (a) If the number of clusters is the same as the number of observations, the K-means result would be the same as Voronoi algorithm.
   (b) The boundary of Voronoi diagram is not flexible as the boundary of K-means.

5. [4] IGNORE THIS QUESTION. For each situation shown in Figure 7.2, explain whether it is a feasible solution of k-means clustering (for the given value of k). Assume Euclidean distance is used as the metric to determine the closest centroid to a point. Explain your reasoning. We consider a solution to be feasible if k-means converges to the centroids shown in the diagram (with the proper choice of initial centroid). The locations of the centroids are marked as X.

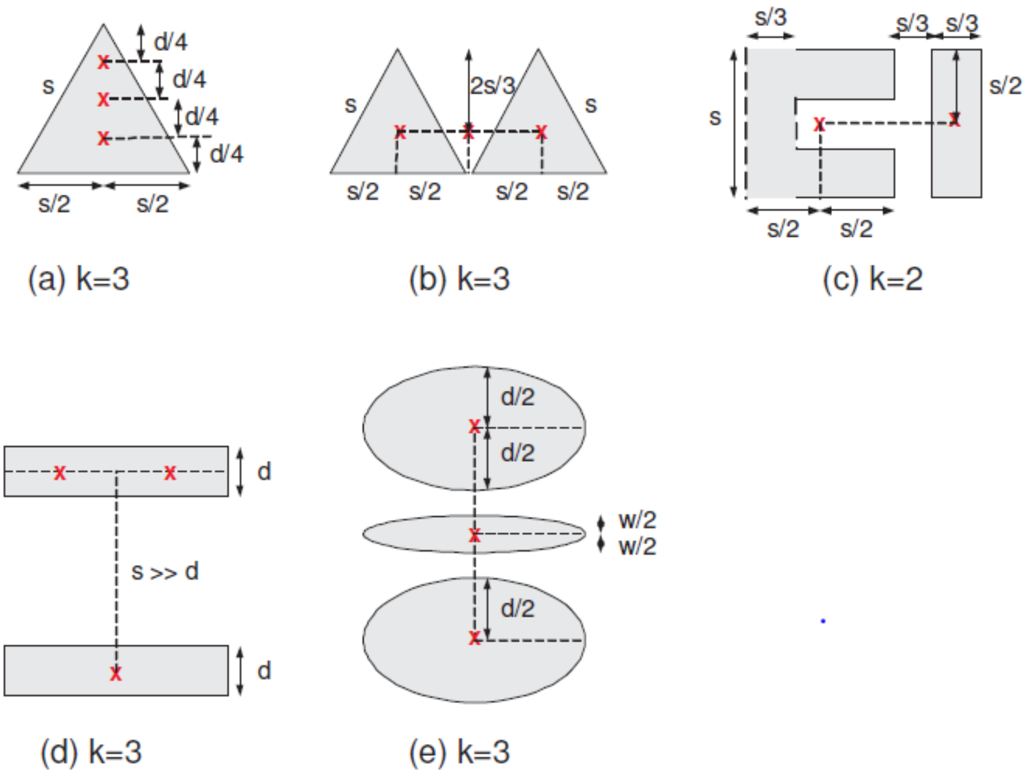
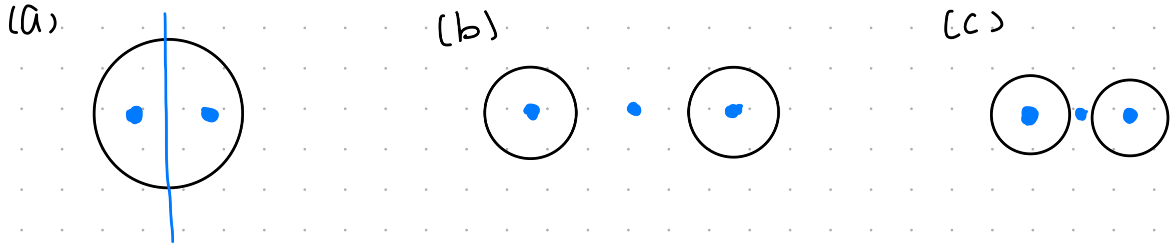


Figure 7.2. K-means clustering.

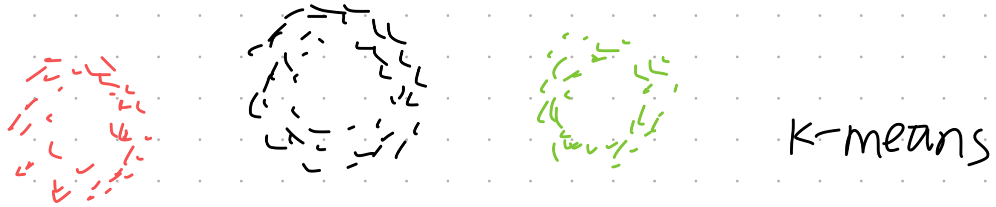
6. #6 Undergraduates do a, b, and c. Do not be concerned with global versus local minima. Graduate students do a, b, c, and d. Discuss, where appropriate, if there are both local and global solutions (in particular do this for c).
   (a) K=2; There are infinite ways to partition the dataset since all diameters could be the boundary. The centroids would locate in the half circle and our two centroids are symmetric.
   (b) K=3; There is only one solution and see picture below.
   (c) K=3; I think the answer would be the same as part be.

6.



7. [Tan #9 adapted] Consider a data set that consists of three circular clusters, that are identical in terms of the number and distribution of points, and whose centers lie on a line and are located such that the center of the middle cluster is equally distant from the other two. Draw a diagram of this situation and discuss how k-means clustering and bisecting k-means clustering might (or might not) reach different solutions.

7.



They will get the same results.

8. Suppose k=2 for k-means clustering and the distance metric being used is the Euclidean distance (2-norm). There are two features, x and y. You have two centroids at the following points: (a, b) and (c, d). Describe the boundary shape between these two clusters and write the equation for the boundary (you can simplify by combining constants into one constant term). Note: the equation for a circle centered at (a, b) is:

$$(x - a)^2 + (y - b)^2 = r^2$$

The boundary shape would be a circle.

9. Delete exercise as this was a duplicate with 4. Just continue with 10.

Hierarchical clustering questions

10. #16
    See another PDF under the same folder.

11. Make a table that explains hierarchical clustering as follows: Each row represents a different inter-cluster metric. Column 1 is the metric name; column 2 are advantages; column 3 are disadvantages.

	Advantages	Disadvantages
Single linkage	Can handle non-elliptical shapes	Sensitive to noise
Complete linkage	Less susceptible to noise	Tends to break large cluster
Average linkage	Less susceptible to noise	Biased toward globular clusters

12. You are given two different clustering situations. For (a) and (b) explain how DBSCAN might behave (produce clusters) for different eps values.

    a. A square where the points are randomly placed such that every point is equally likely to be selected as an observation.
       It works fine

    b. A square in which observations are uniformly spaced.
       It works fine