Exercise 4

Task 1

a) 1st, we randomly choose A, B as C_1, C_2.

2nd, calculate the Euclidean distance between C, D to ~~each~~ A, B respectively

$d_{CA} = \sqrt{4+0.36} \approx 2.088$, $d_{CB} = \sqrt{0.04+2.25} \approx 1.513$

$d_{DA} = \sqrt{1+0.25} \approx 1.118$ $d_{DB} = \sqrt{0.64+2.56} \approx 1.789$

C is closer to B, D is closer to A. We assign C to cluster K_2, D to cluster K_1.

3rd. calculate the new cluster centroids

$$C_1 = \frac{(1.2+0.2,\ 0.8+0.3)}{2} = (0.7,\ 0.55),$$

$$C_2 = \frac{(-0.6-0.8,\ -1.3+0.2)}{2} = (-0.7,\ -0.55)$$

4th, we calculate the Euclidean distance between A, B, C, D and C_1, C_2:

$d_{AC_1} = \sqrt{0.25+0.0625} \approx 0.559$, $d_{AC_2} = \sqrt{1.9^2+0.0625} \approx 1.916$

$d_{BC_1} = \sqrt{1.3^2+1.85^2} \approx 2.261$, $d_{BC_2} = \sqrt{0.01+0.75^2} \approx 0.757$

$d_{CC_1} = \sqrt{1.5^2+0.35^2} \approx 1.54$, $d_{CC_2} = \sqrt{0.01+0.75^2} \approx 0.757$

$d_{DC_1} = \sqrt{0.25+0.25^2} \approx 0.559$. $d_{DC_2} = \sqrt{0.81+0.7225} \approx 1.238$

Assign A, D to K_1. B, C to K_2, as in 2nd step. Converge.

b) see the code.

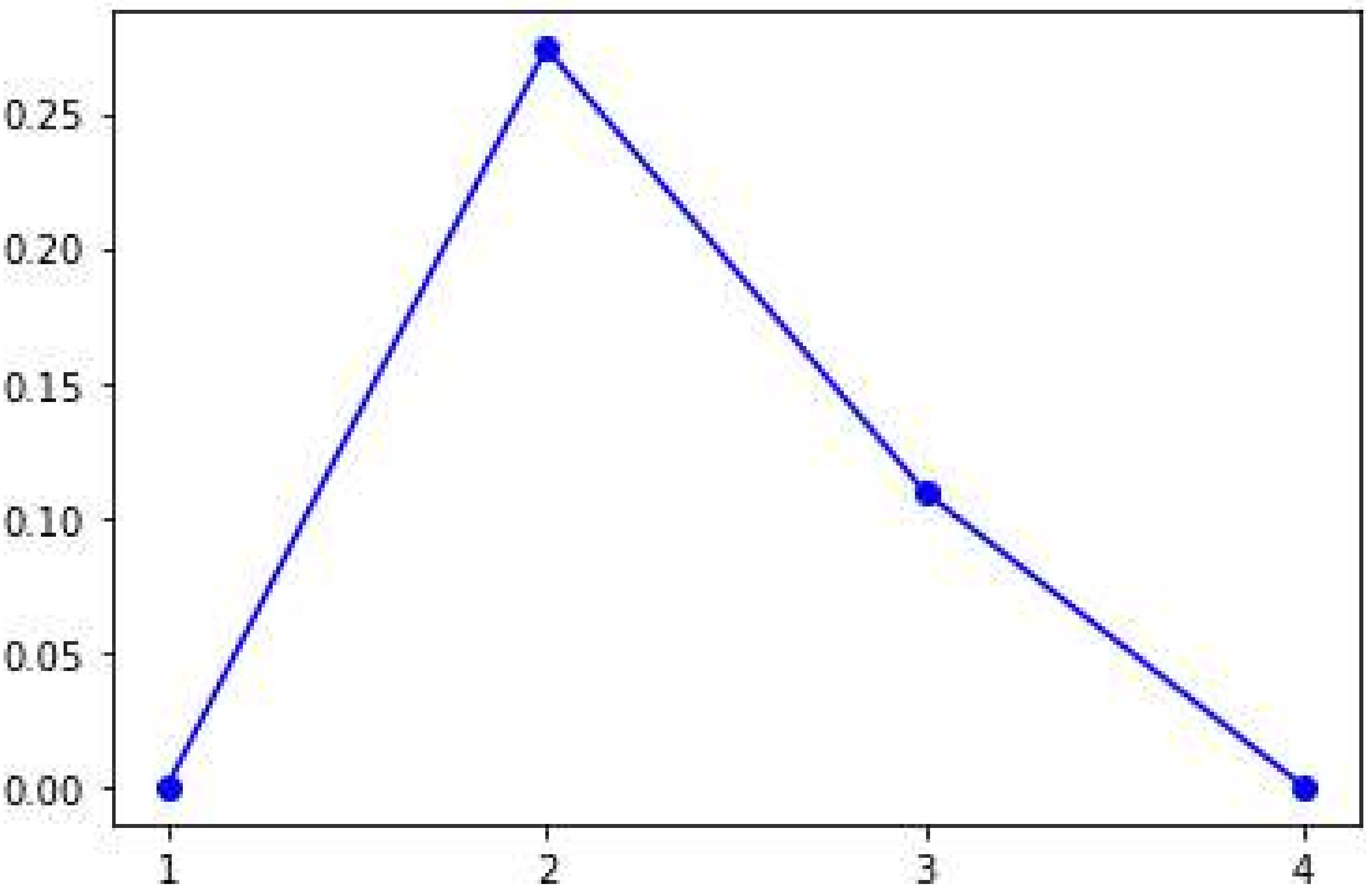
c) Silhouettes are for evaluating the result of clustering. $d(A,o)$ shows the distances between target node and other nodes in the same cluster. the smaller $d(A,o)$ is, the more similar o is with other nodes of the same cluster. ~~In contrast,~~ And $d(B,o)$ shows the distance to nodes in other cluster.

$$s(o) = \frac{d(B,o) - d(A,o)}{\max\{d(A,o), d(B,o)\}} = \begin{cases} 1 - \frac{d(A,o)}{d(B,o)} & \text{if } d(B,o) > d(A,o) \\ -1 + \frac{d(B,o)}{d(A,o)} & \text{if } d(A,o) > d(B,o) \\ 0 & \text{if } d(A,o) = d(B,o) \end{cases}$$

if $s(o) \to 1$, it says $d(B,o) \gg d(A,o)$, the result of clustering is quite good.

if $s(o) \to -1$, it says $d(A,o) \gg d(B,o)$, o supposed to be classified to other cluster.

if $s(o) = 0$, it says there is no much difference if o in current cluster or other cluster. o might be a boundary point.


0.25
0.20
0.15
0.10
0.05
0.00
1
2
3
4

Task 2

a) see the picture.

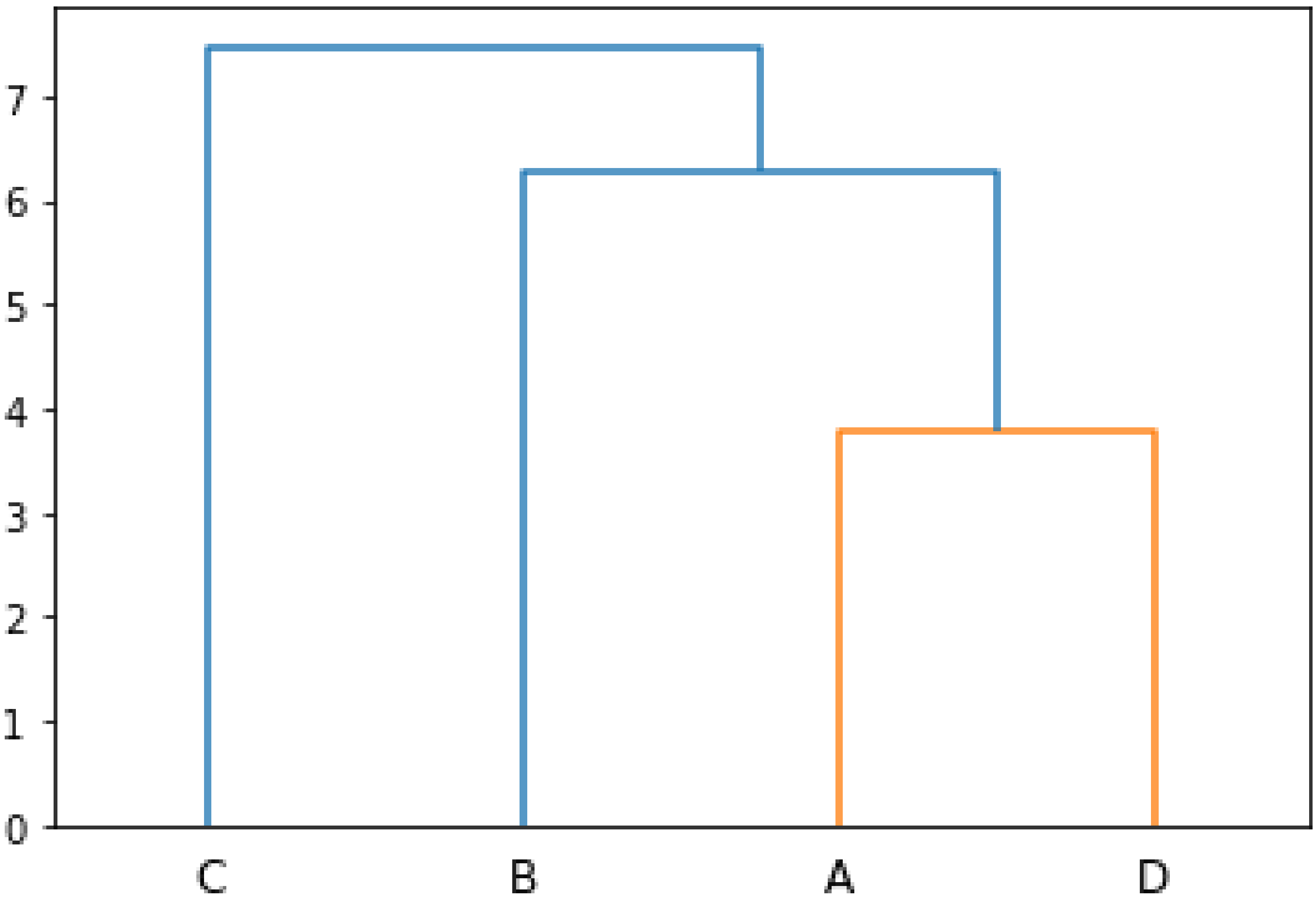
b) For patitioning methods, the clusters are not overlapping. However, hierarchy methods are consist of nested clusters, the higher level clusters based on clusters from previous levels.

c) For hard clustering, the class of an object is certain, it can only assigned to one cluster. Meanwhile, soft clustering assign an obeject the probabilities belong to different clucters. The sum ~~of~~ probabities of an object in different clusters is 1.

d) In the case thet the shapes of different clusters are complicated, not regular, or there are not much space of gap between different clusters, we could consider density-based methods, as it is used despite the shapes and sizes of ~~ech~~ each clusters.

Advantages: it could detect different clusters despite shapes and size of clusters.

Disadvantage: If the distribution of density inside a cluster is not even, then it might misclassify.


7
6
5
4
3
2
1
0
C
B
A
D