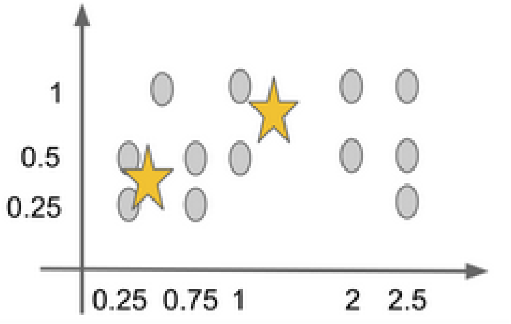
Data Science

Problem sheet 4

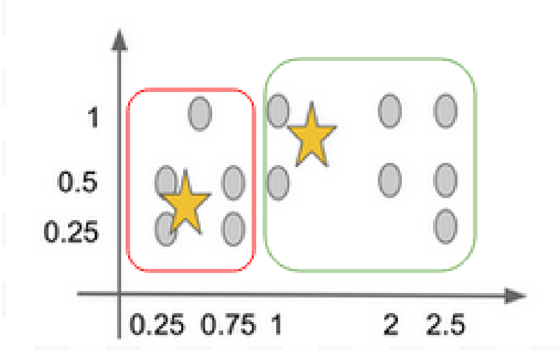
Ex 1

We initialize k-means clustering algorithm with the centroids marked by stars in the figure. Perform an iteration step! Calculate the positions of the new centroids!



Solution

We assign each data point to the nearest centroid. Then we calculate the mean of the coordinates of the data points in each cluster. The means of the coordinates are the new coordinates of the centroids.



Left (red) box:

$$x = \frac{0.25 + 0.25 + 0.5 + 0.75 + 0.75}{5} = 0.5$$
$$y = \frac{0.25 + 0.5 + 1 + 0.25 + 0.5}{5} = 0.5$$

The corresponding centroid: (0.5. 0.5)

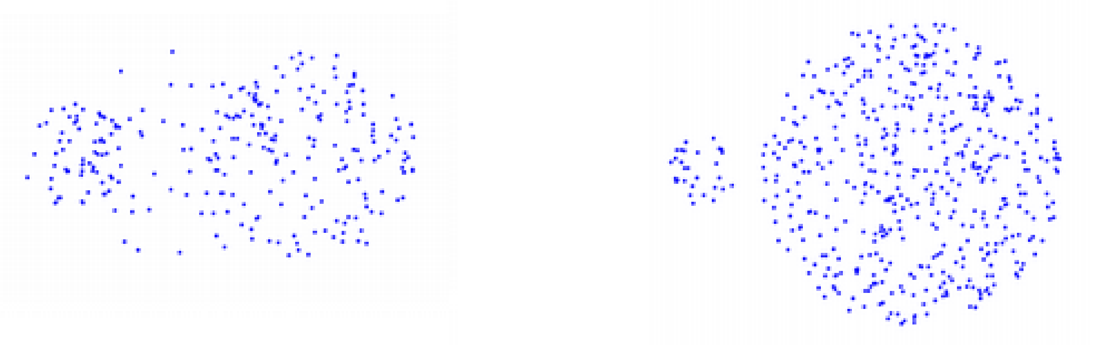
Right (green) box:

$$x = \frac{1 + 1 + 2 + 2 + 2.5 + 2.5 + 2.5}{7} = 1.9$$
$$y = \frac{0.25 + 0.5 + 0.5 + 0.5 + 1 + 1 + 1}{7} = 0.68$$

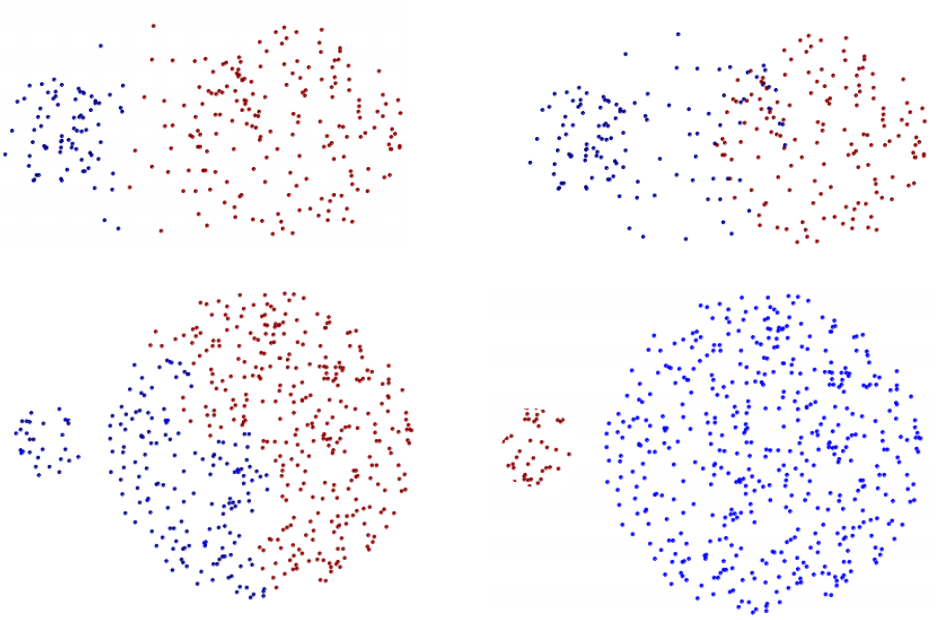
The corresponding centroid: (1.92, 0.68)

Ex 2

Consider the two-dimensional data sets below. How would the following clustering algorithms split the data into two clusters: k-means, single-linkage and complete-linkage hierarchical clustering?



Solution



For the first data set k-means and the complete-linkage would find similar clusters (top left figure) with sharp boundaries. Single-linkage is sensitive to noise and outliers (top right figure). For the second data set k-means and complete-linkage would also find similar clusters, namely they would break the large clusters (bottom left figure), while for this data set simple linkage would find the natural clusters (bottom right figure).

Ex 3

Use the following similarity matrix to perform single and complete linkage hierarchical clustering by drawing two dendrograms!

	1	2	3	4	5
1	1	0.15	0.6	0.15	0.95
2	0.15	1	0.5	0.2	0.2
3	0.6	0.5	1	0.05	0.7
4	0.15	0.2	0.05	1	0.85
5	0.95	0.2	0.7	0.85	1

Solution

Single linkage

	1, 5	2	3	4
1,5	1	0.2	0.7	$\boxed{0.85}$
2	0.2	1	0.5	0.2
3	0.7	0.5	1	0.05
4	$\boxed{0.85}$	0.2	0.05	1

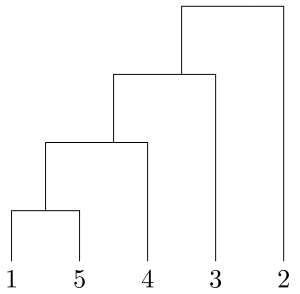
↓

	1, 5, 4	2	4
1, 5, 4	1	0.2	$\boxed{0.7}$
2	0.2	1	0.5
3	$\boxed{0.7}$	0.5	1

↓

	1, 5, 4, 3	2
1, 5, 4, 3	1	0.5
2	0.5	1

Dendrogram:



Complete linkage

	1, 5	2	3	4
1, 5	1	0.15	$\boxed{0.6}$	0.15
2	0.15	1	0.5	0.2
3	$\boxed{0.6}$	0.5	1	0.05
4	0.15	0.2	0.05	1

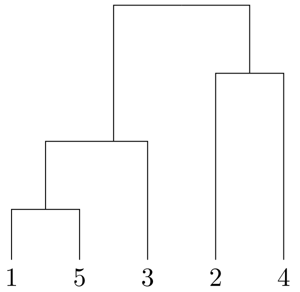
↓

	1, 5, 3	2	4
1, 5, 3	1	0.15	0.05
2	0.15	1	$\boxed{0.2}$
4	0.05	$\boxed{0.2}$	1

↓

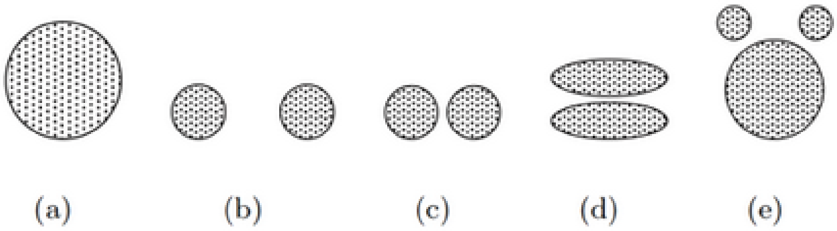
	1, 5, 3	2, 4
1, 5, 3	1	0.05
2, 4	0.05	1

Dendrogram:

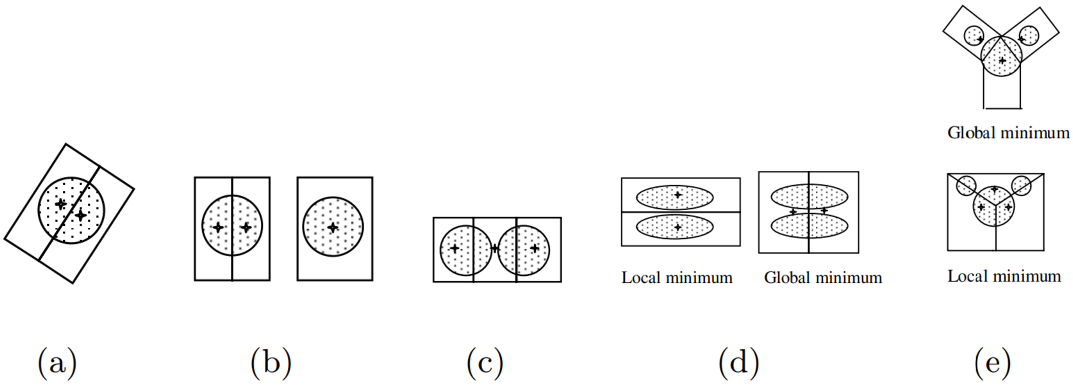


Ex 4

Consider the following sets of two-dimensional points. For the given number of clusters, provide a sketch of the resulting clusters found by k-means algorithm, also indicate the positions of centroids. Assume that Euclidean distance is used and the points are uniformly distributed. If you think that there are more than one possible solutions, then indicate whether a solution is a global or local minimum (i.e. in which case the SSE is smaller) (a) $k = 2$ (b) $k = 3$ (c) $k = 3$ (d) $k = 2$ (e) $k = 3$ (Note that the label of each diagram below matches the corresponding part of this question.)



Solution



a) k=2. In theory, there are an infinite number of ways to split the circle into two clusters - just take any line that bisects the circle. This line can make any angle $0° \leq \theta \leq 180°$ with the x axis. The centroids will lie on the perpendicular bisector of the line that splits the circle into two clusters and will be symmetrically positioned. All these solutions will have the same, globally minimal, error.

b) k=3. The distance between the edges of the circles is slightly greater than the radii of the circles. If you start with initial centroids that are close to real points, you will necessarily get this solution because of the restriction that the circles are more than one radius apart. Of course, the bisector could have any angle, as above, and it could be the other circle that is split. All these solutions have the same globally minimal error.

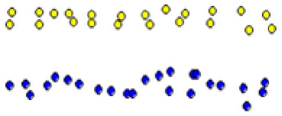
c) k=3. The distance between the edges of the circles is much less than the radii of the circles. The three boxes show the three clusters that will result in the realistic case that the initial centroids are close to actual data points.

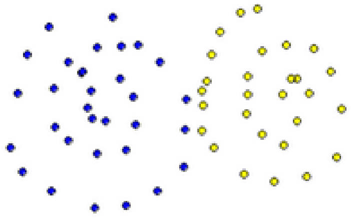
d) k=2. In both case, the rectangles show the clusters. In the first case, the two clusters are only a local minimum while in the second case the clusters represent a globally minimal solution.

e) k=3. For the solution shown in the top figure, the two top clusters are enclosed in two boxes, while the third cluster is enclosed by the regions defined by a triangle and a rectangle. (The two smaller clusters in the drawing are supposed to be symmetrical.) The second solution also possible, although it is a local minimum and might rarely be seen in practice for this configuration of points.

Ex 5

Which clustering algorithm would you use if the goal was to find the two natural clusters (marked by blue and yellow colors)? Consider the following algorithms: k-means, hierarchical clustering (both single and complete linkage) and DBSCAN.





Solution

In the first case, DBSCAN and single linkage algorithms would perform well. However, k-means would split the data vertically. Complete-linkage is also unable to find the natural clusters, since there exist yellow and blue points that are closer to each other than some points of the same color.

In the second case only k-means finds the natural clusters. The hierarchical clustering algorithms would merge the blue and yellow points in the middle as one of the first steps since they are very close to each other. DBSCAN would also fail to find the natural clusters.