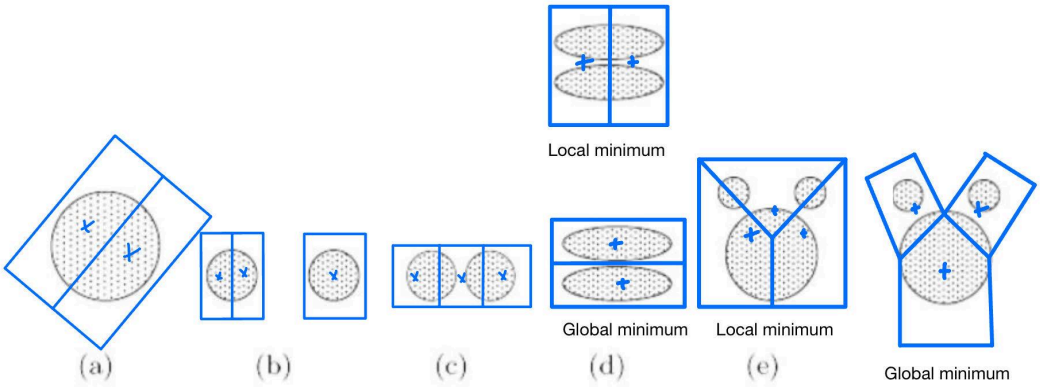
Homework 7

1 Exercises

1.1.2 Find all well-separated clusters in the set of points shown in Figure 7.35

1.1.6.1 For the following sets of two-dimensional points, provide a sketch of how they would be split into clusters by K-means for the given number of clusters.



1.1.6.2 For the following sets of two-dimensional points, indicate approximately where the resulting centroids would be. Assume that we are using the squared error objective function. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum.

a. K = 2. Assuming that the points are uniformly distributed in the circle, how many possible ways are there (in theory) to partition the points into two clusters? What can you say about the positions of the two centroids? (Again, you don't need to provide exact centroid locations, just a qualitative description.)

There are infinite ways to partition in theory. The circle can be split into two clusters with any line that bisects the circle. The centroids will lie on the perpendicular bisector of the line that splits the circle and will be symmetrical to each other.

b. K = 3. The distance between the edges of the circles is slightly greater than the radii of the circles.

One of the circles will be split and the other will be whole. The split circle will have centroids identical to the ones in a. The whole circle will have a centroid at its center.

c. K = 3. The distance between the edges of the circles is much less than the radii of the circles.
The centroids would be equidistant from each other along a line that splits the circles into upper and lower halves. This will create three clusters splitting the circles vertically.

d. K = 2.
For local minimum, the two clusters would separate the circles with centroids at the circle's center. For global minimum, the circles would be split vertically where one cluster would have half of the first and half of the second while the other cluster would have the rest. The centroids for this would not be in the circles but in between them at the center of the cluster.

e. K = 3. Hint: Use the symmetry of the situation and remember that we are looking for a rough sketch of what the result would be.
For local minimum, the centroids would all be along the edge of the big circle. For global minimum, the centroids would be split so each circle has one.

1.1.7 Suppose that for a data set

- there are m points and K clusters,
- half the points and clusters are in "more dense" regions,
- half the points and clusters are in "less dense" regions, and
- the two regions are well-separated from each other.

For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

~~**a.** Centroids should be equally distributed between more dense and less dense regions.~~
~~**b.** More centroids should be allocated to the less dense region.~~
c. More centroids should be allocated to the denser region.

1.1.11 Total SSE is the sum of the SSE for each separate attribute.
What does it mean if the SSE for one variable is low for all clusters?
The variable provides little value in distinguishing between clusters.
What does it mean if the SSE for one variable is low for just one cluster?
It indicates that this attribute plays a key role in defining that specific cluster.
What does it mean if the SSE for one variable is high for all clusters?
It suggests that the attribute may primarily consist of noise and does not contribute meaningful information for clustering.
What does it mean if the SSE for one variable is high for just one cluster?
It conflicts with the information provided by attributes with low SSE. This might indicate that the clusters formed by this attribute are inconsistent with those formed by other attributes, making it unhelpful.
How could you use the per variable SSE information to improve your clustering?

To improve clustering, it's important to eliminate attributes with consistently low or high SSE across all clusters, as they add little value to the clustering process. Attributes with high SSE for all clusters are especially problematic as they can introduce noise into the overall SSE computation.

1.1.12 The leader algorithm represents each cluster using a point, known as a leader, and assigns each point to the cluster corresponding to the closest leader, unless this distance is above a user-specified threshold. In that case, the point becomes the leader of a new cluster.

a. What are the advantages and disadvantages of the leader algorithm as compared to K-means?

The leader algorithm offers several advantages over K-means. It dynamically determines the number of clusters based on a distance threshold, making it ideal when the number of clusters is unknown. It is computationally efficient, processes data sequentially, and accommodates new data without requiring recalculations. However, the leader algorithm is sensitive to the order of data points, and selecting an appropriate threshold can be challenging. It does not refine clusters after formation, which can lead to suboptimal results. While the leader algorithm is suited for fast, dynamic clustering, K-means is preferable for iterative refinement and fixed cluster counts.

b. Suggest ways in which the leader algorithm might be improved.

The leader algorithm can be improved by addressing its limitations. To reduce order sensitivity, multiple random orderings of the data can be processed, and the clustering results could be averaged or combined to produce more stable clusters. Adaptive distance thresholds could be used instead of a fixed threshold, allowing the algorithm to adjust dynamically based on data density. Finally, incorporating a refinement step, similar to the iterative updates in K-means, could improve the quality of clusters after the initial assignment.

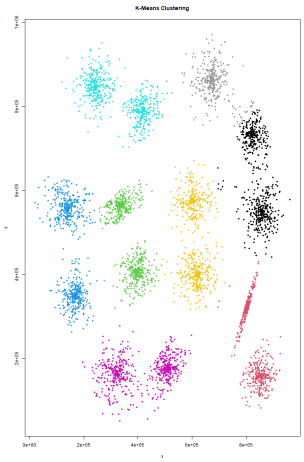
1.1.16 Use the similarity matrix in Table 7.13 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

2 Practicum problems

Q2.1(b)(iv) Of the three graphs, which one produces what you would consider the best clusters?

The clustering scheme that produces the best clusters is Complete linkage, as it groups countries that predominantly speak the same languages into the same clusters. Great Britain and Ireland are clustered together due to English, West Germany and Austria are together and the Scandinavian countries (Sweden, Norway, Denmark) are also grouped together.

Q2.2(e) Looking at the plot in 2.2(d), comment on the clusters created by k-means. Indicate whether the clustering is satisfactory or not by justifying your answer in 1-3 sentences



Based on the plot, the clustering is not satisfactory. We can see that multiple clusters are grouped together and they all do not form distinct groups. The current clustering does provide some insight and is accurate to an extent but it could definitely be better.

Q2.3(d) Comment, in 1-2 sentences, on the results of clustering after you have chosen MinPts and an optimal ε. Are all clusters well-separated? Are the noise points well identified?

After selecting a higher ε value than suggested by the elbow of the scree plot, the DBSCAN algorithm produced reasonably good clustering results. The clusters were well-separated, and the noise points were effectively identified, demonstrating the algorithm's robustness.