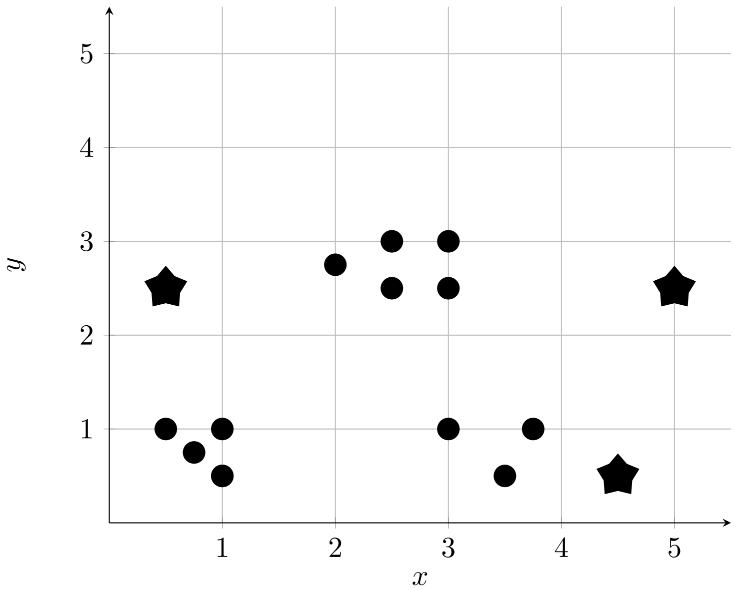
Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Tutorial sheet 8: Clustering

Prof. Dr. Martin Bichler, Prof. Dr. Jalal Etesami
Julius Durmann, Markus Ewert, Johannes Knörr, Yutong Chao

Exercise T8.1 *k-means*

Group the following data into three clusters applying the k-Means algorithm and the Euclidean distance function.



i	x_i	y_i
1	2.5	3
2	3	3
3	2	2.75
4	2.5	2.5
5	3	2.5
6	0.5	1
7	1	1
8	3	1
9	3.75	1
10	0.75	0.75
11	1	0.5
12	3.5	0.5

Table 1 Dataset

i	x_i	y_i
A	0.5	2.5
B	5	2.5
C	4.5	0.5

Table 2 Initial Centroids

Exercise T8.2 *Problems with k-means*

You are given the following small dataset in Table 3:

i	x_i	y_i
0	1	1
1	1	2
2	7	1
3	7	2

Table 3 Small Dataset

(a) Perform 2-means clustering using the L2 norm with initial centroids $A = (4, 1)$ and $B = (4, 2)$ on the data in Table 3. Based on your result, discuss one problem with k-means and name one possible remedy.

(b) Now a fifth point $(3, 20)$ is added to the dataset in Table 3. Perform 2-means clustering using the L2 norm with initial centroids $A = (1, 1)$ and $B = (7, 1)$. Based on your result, discuss another problem with k-means and name one possible remedy.

Exercise T8.3 *Hierarchical clustering*

You are given the following dataset:

i	x_i	y_i
0	0	0
1	-1	1
2	3	0
3	-3	5
4	1	2
5	-2	-3
6	0	2
7	4	1
8	-3	-1
9	-2	-2

Table 4 Dataset

You are further given the distance matrix D_1 based on the L1 norm.

$$
D_1 = \begin{bmatrix}
0 & 2 & 3 & 8 & 3 & 5 & 2 & 5 & 4 & 4 \\
 & 0 & 5 & 6 & 3 & 5 & 2 & 5 & 4 & 4 \\
 & & 0 & 11 & 4 & 8 & 5 & 2 & 7 & 7 \\
 & & & 0 & 7 & 9 & 6 & 11 & 6 & 8 \\
 & & & & 0 & 8 & 1 & 4 & 7 & 7 \\
 & & & & & 0 & 7 & 10 & 3 & 1 \\
 & & & & & & 0 & 5 & 6 & 6 \\
 & & & & & & & 0 & 9 & 9 \\
 & & & & & & & & 0 & 2 \\
 & & & & & & & & & 0
\end{bmatrix}
$$

You want to perform bottom-up hierarchical clustering using the L1 norm. Use single-linkage clustering, i.e., the distance between two sets of observations A, B is defined as $\min_{a \in A, b \in B} d(a, b)$. The first couple of steps have already been conducted and there are currently four clusters:

- Cluster 1: points 0, 1, 4, 6
- Cluster 2: points 2, 7
- Cluster 3: points 5, 8, 9
- Cluster 4: point 3

Complete the remaining steps of the hierarchical clustering.

Exercise T8.4 *k-means for image compression*

The goal of this exercise is to use k-means clustering for image compression in Python.

a) Load an image of the famous painting "American Gothic" by Grant Wood and refactor it to an RGB-Image. You can access the painting here. Use the following code:

```
# Load the image
url = #INSERT LINK HERE
img = io.imread(url)
io.imsave("original.png", img)

# notice, that the image has 3 channels (red, green, blue)
print("shape:", img.shape)

# split the image into the channels (red, green, blue)
r = img[:, :, 0]
g = img[:, :, 1]
b = img[:, :, 2]
)
```

The first two columns in img describe the position of the pixel in the painting. The variables r, g, and b together encode the color of each pixel.

b) How many unique colors does the painting contain?

c) Apply k-means clustering to the pixel colors. Choose $k = 5$ as the number of clusters. Plot the resulting compressed image. The following code snippet may be helpful:

```
km = KMeans(n_clusters=5, init="random", max_iter=300)
km.fit(img)
new_colors = km.cluster_centers_[km.predict(img)]
```

d) Apply k-means clustering for $k = \{1, 2, 3, 5, 10, 20, 50\}$. Plot and save the compressed image in each iteration. Be aware of increased runtimes on personal computers. Observe the size of the image files. Looking at the images, at what point do you notice only minor differences?

e) Determine a reasonable number of clusters using the "elbow criterion". For this purpose, plot the total within-cluster sum of squares (attribute `inertia_` of the `KMeans` object) against the number of clusters, e.g., for $k \in [1, 10]$. Does the elbow point correspond to your visual impression in part d)?