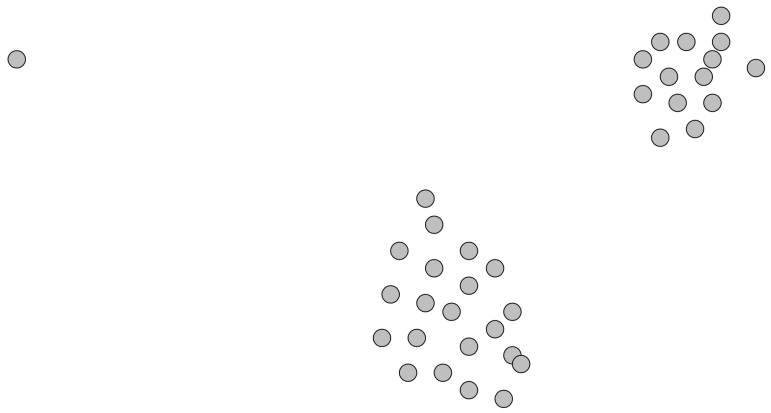
Exercise 1 :

For the following data set:



- identify two initial cluster centers such that 2-means clustering initialized with these cluster centers will terminate with the single outlier forming one cluster, and all other points the second cluster
- identify two initial cluster centers such that 2-means clustering initialized with these cluster centers will terminate with the two big groups of points forming different clusters (and the outlier belonging to one of those)

In both cases indicate the (approximate) position of the final cluster centers.

Exercise 2 :

Consider the data points plotted in Figure 1.

- Perform two iterations of the k-means algorithm using
  - the data points $(2, 6)$ and $(3, 5)$ as the initial cluster centers.
  - the Euclidean distance as distance metric
- Calculate the sum of squared errors using the initial cluster centers and the cluster centers that you found above.

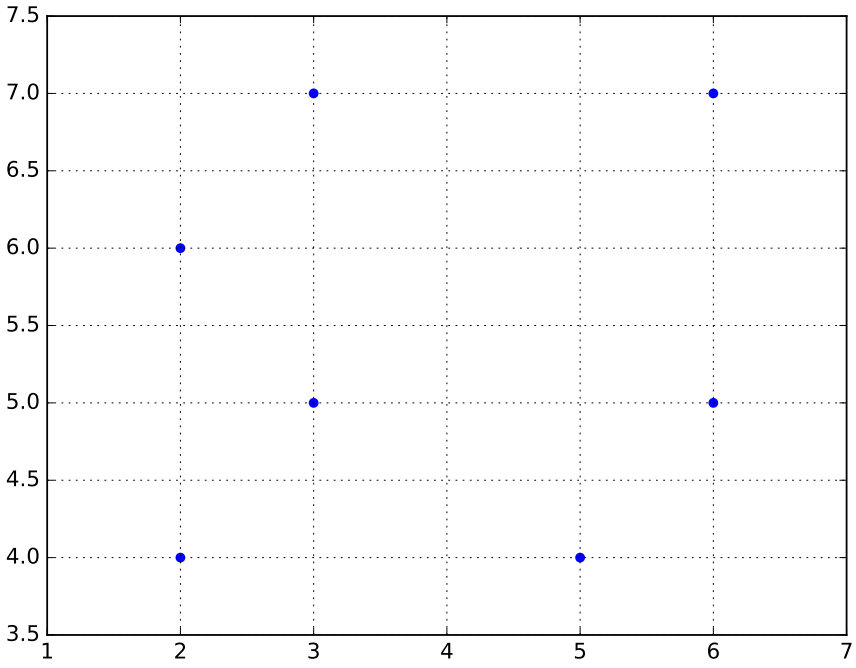


Figure 1: Data to be clustered in Exercise 2.

Exercise 3 :

Perform one more iteration of the EM algorithm for the example on Slide 10.20. Note that you will first need to complete the last two maximization calculations for the 2nd iteration, which is left unfinished on the slides.

Exercise 4 :

Consider the following five data points living in $\mathbb{R}^2$:

d_1	$(7, 2)$
d_2	$(52, 3)$
d_3	$(70, 10)$
d_4	$(85, 11)$
d_5	$(90, 8)$

Using the Euclidean distance to measure distances:

(i) Find the data point closest to d_2 in the data set, i.e., $(52, 3)$.

(ii) Normalize the data using Z-score normalization. Find the data point closest to d_2 using the transformed data set.

(iii) Let the data points $(7, 2)$ and $(70, 10)$ be initial cluster centers for the k-means algorithm. Perform one more k-means iteration by updating these cluster centers using the non-normalized data set above.

(iv) Do you see any potential problems in directly using k-means clustering with the Euclidean distance based the data set above? What could you do to mitigate such problems?

Exercise 5 :

Use WEKA to perform clustering experiments on the datasets clustering clusters.arff and clustering random.arff.

1. Perform k-means clustering for $k = 1, 2, 3, 4, 5, 6, 7, 8$ on the two data sets. For each clustering, WEKA outputs the "Within cluster sum of squared errors" (which corresponds to the sum of squared errors). Make a plot of this error as a function of k for both datasets. How can plots like these be used to determine the "right" number of clusters?
2. For $k = 3$ perform 4 runs each of k-means clustering using different setting of the random seed (left click in the 'Clusterer' text field to set the random seed). Compare the results obtained in the different runs using the "Visualize cluster assignment" function (accessible via the Result list panel). How does this help you to decide which of the two datasets has 3 "real" clusters?
