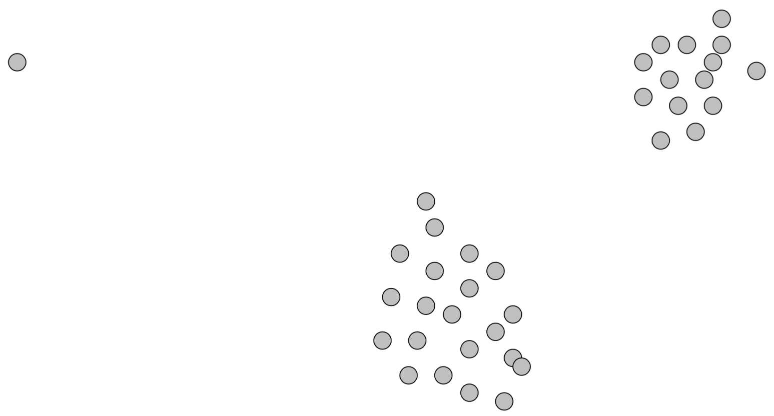
Exercise 1 :

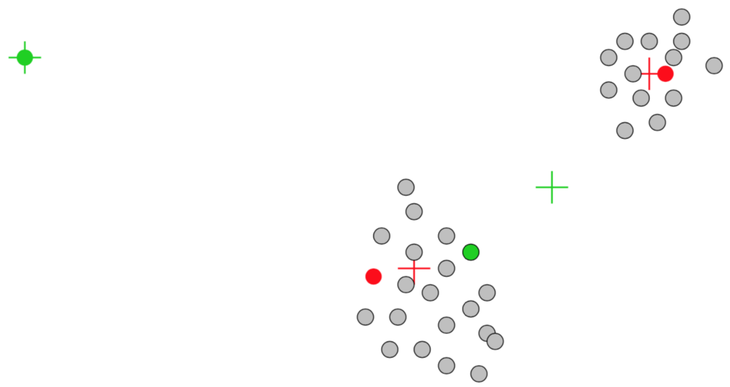
For the following data set:



- 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.

Solution:



- The green filled instances are possible initial cluster centers; the green crosses are the (approximate) final cluster centers.
- Same with red ...

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.

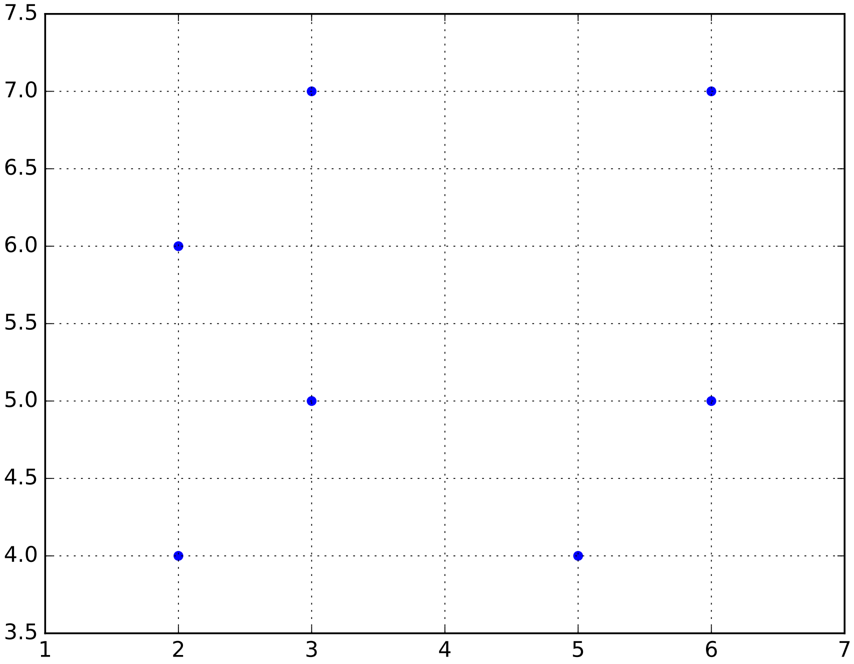


Figure 1: Data to be clustered in Exercise 2.

Solution:

The result of the first two iterations are illustrated in Figure 2. The cluster centers are located at:

- 1. iteration: $(2.5, 6.5)$ and $(4.4, 5)$.
- 2. iteration: $(2.33, 5.67)$ and $(5, 5.25)$.

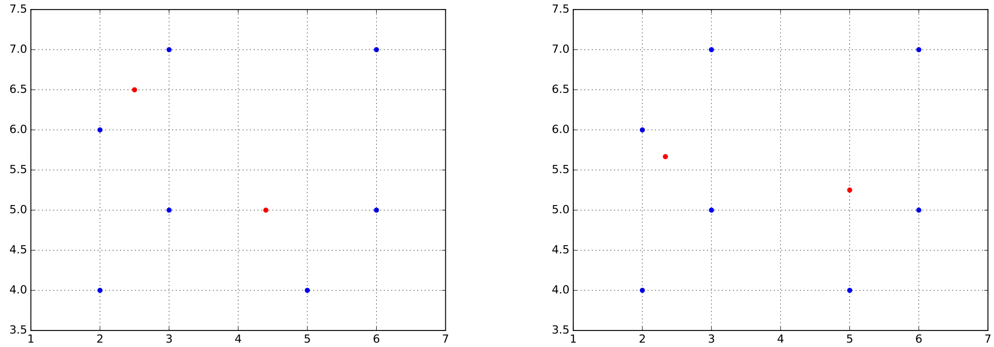


Figure 2: The cluster centers (red dots) after the first and second iterations, respectively.

The sum of squared errors are 31.0 and 12.91, respectively.

Exercise 3 :

Perform one more iteration of the EM algorithm for the example on Slide 10.33. Note that you will first need to complete the last two maximization calculations for the 2nd iteration, which is left unfinished on the slides.

Solution:

First we start by completing the second iteration of the EM-algorithm. This consists of updating the distributions for $P_2(F_2|C)$ and $P_2(F_3|C)$; note that the subscript 2 refers to the iteration number.

$$P_2(F_2|C) = \frac{\sum_{F_1,F_3} A(F_1,F_2,F_3,C)}{\sum_{F_1,F_2,F_3} A(F_1,F_2,F_3,C)}$$

$$= \frac{\begin{array}{c|cc} & \text{C=1} & \text{C= 2} \\ \hline F_2 = t & \text{0.88 + 0 + 0 + 0} & \text{0.12+0+0+0} \\ F_2 = f & \text{0+0.66+0.48+0.47} & \text{0+0.34+0.52+0.53} \end{array}}{(0.88+0.66+0.48+0.47, 0.12+0.34+0.52+0.53)}$$

$$= \frac{\begin{array}{c|cc} & \text{C=1} & \text{C= 2} \\ \hline F_2 = t & 0.88 & 0.12 \\ F_2 = f & 1.61 & 1.39 \end{array}}{(2.49, 1.51)}$$

$$= \begin{array}{c|cc} & \text{C=1} & \text{C= 2} \\ \hline F_2 = t & 0.35 & 0.08 \\ F_2 = f & 0.65 & 0.92 \end{array}$$

The tables above are structured in the same way as on the slides. Thus, the columns correspond to the two states of C and the rows correspond to $F_2 = t$ and $F_2 = f$, respectively. For $P_2(F_3|C)$ we end up with (the intermediate calculations follow the same steps as above):

$$P_2(F_3|C) = \frac{\begin{array}{c|cc} & C=1 & C=2 \\ \hline F_3=t & 2.01 & 0.99 \\ F_3=f & 0.48 & 0.52 \end{array}}{(2.49, 1.51)}$$
$$= \begin{array}{c|cc} & C=1 & C=2 \\ \hline F_3=t & 0.81 & 0.66 \\ F_3=f & 0.19 & 0.34 \end{array}$$

Continuing with the third iteration of the EM-algorithm, we start by calculating the expected counts that should go into our count table $A(F_1, F_2, F_3, C)$. This should be done using the conditional probability tables estimated during the last iteration (i.e., $P_2(C)$, $P_2(F_1|C)$, $P_2(F_2|C)$, and $P_2(F_3|C)$) and reduces to calculating $P_2(C|F_1, F_2, F_3)$ for the four different configurations of F_1, F_2, and F_3 observed in the data. In total we end up with:

F_1	F_2	F_3	$P(C\|F_1, F_2, F_3)$
t	t	t	(0.92,0.08)
t	f	t	(0.64,0.36)
t	f	f	(0.44,0.56)
f	f	t	(0.43,0.57)

Based on the updated count table, we recalculate the conditional probabilities of the model, i.e., $P_3(C)$, $P_3(F_1|C)$, $P_3(F_2|C)$, and $P_3(F_3|C)$. The calculation procedures are the same as for the previous two iterations and results in the following tables:

$$P_3(C) = (0.61, 0.39)$$

$$P_3(F_1|C) = \begin{array}{c|cc} & C=1 & C=2 \\ \hline F_1=t & 0.82 & 0.64 \\ F_1=f & 0.18 & 0.36 \end{array}$$

$$P_3(F_2|C) = \begin{array}{c|cc} & C=1 & C=2 \\ \hline F_1=t & 0.39 & 0.05 \\ F_1=f & 0.61 & 0.95 \end{array}$$

$$P_3(F_3|C) = \begin{array}{c|cc} & C=1 & C=2 \\ \hline F_3=t & 0.82 & 0.64 \\ F_3=f & 0.18 & 0.36 \end{array}$$

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?

Solution:

(i) The data point $d_3 = (70, 10)$ is the one closets to $d_2 = (52, 3)$ with a distance of 19.31.

```
d1  0.00         45.01   63.51   78.52   83.22
d2  45.01        0.00    19.31   33.96   38.33
d3  63.51        19.31   0.00    15.03   20.10
d4  78.52        33.96   15.03   0.00    5.83
d5  83.22        38.33   20.10   5.83    0.00
```

Distances

(ii) The normalized data set is given by:

$$(-1.79, -1.31), (-0.29, -1.03), (0.3, 0.87), (0.81, 1.15), (0.97, 0.33)$$

The data point closest to $(-0.29, -1.03)$ (corresponding to $d_2 = (52, 3)$) is $(-1.79, -1.31)$ (corresponding to $d_1 = (7, 2)$).

```
d1 0.00  1.53   3.03   3.58   3.22
d2 1.53  0.00   2.01   2.45   1.86
d3 3.03  2.01   0.00   0.57   0.86
d4 3.58  2.45   0.57   0.00   0.84
d5 3.22  1.86   0.86   0.84   0.00
```

Normalized distances

(iii) First we find the points that belong to the two clusters defined by the initially chosen cluster centers:

dist	d_2	d_4	d_5
d_1	45.01	78.51	83.22
d_3	19.31	15.03	20.10

Thus, the data points d_2, d_4, and d_5 all move to the cluster defined by d_3; d_1 now defines a cluster with a single element.

Since no points are assigned to the cluster defined by d_1 we keep that data point as a cluster center. The cluster center defined by the mean of the four remaining points becomes $(74.25, 8)$.

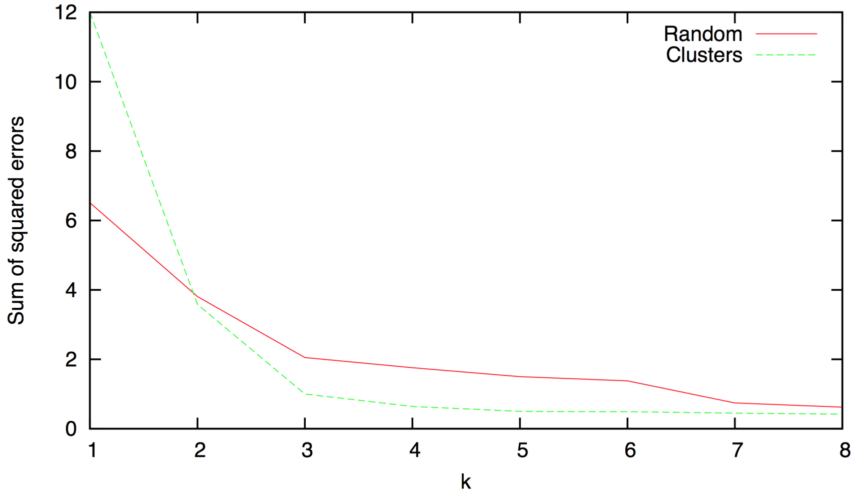
Exercise 5 :

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?

Solution:

1. A plot of the within cluster sum of squared errors:



The plot indicates that for clustering clusters.arff a particularly sharp drop in the error value (compared to the drop for a random dataset) up to $k = 3$, where the curve then levels off. The random data shows a more uniform decrease in error value over the whole k-range. “Knees” in the SSE error function are a (heuristic) indicator for the “right” number of clusters.

2. The comparison shows that for k = 3 the clusters in clustering clusters.arff are stable, i.e. the same clusters are returned independent of the random seed (which determines the random initial cluster centers of the algo- rithm). For clustering random.arff the final result is different for different settings of the seed, indicating that the computed clusters are not well-defined clusters in the data.