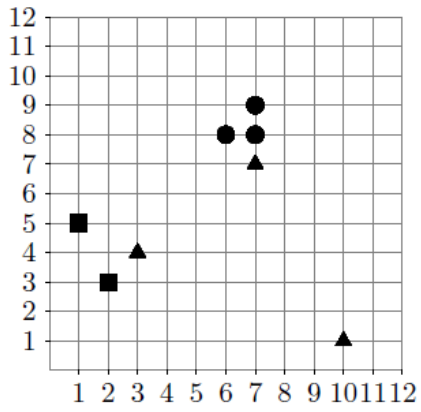
Uge 45

Ex 1

Given the following data set with 8 objects (in $\mathbb{R}^2$) as in the lecture:



Compute a complete partitioning of the data set into $k = 3$ clusters using the basic k-means algorithm (due to Forgy and Lloyd). The initial assignment of objects to clusters is given using the triangle, square, and circle markers.

Objects x are assigned to the cluster with the least increase in squared deviations $SSQ(x, c)$ where c is the cluster center.

$$SSQ(x, c) = \sum_{i=1}^{d} |x_i - c_i|^2$$

Start with computing the initial centroids, and draw the cluster assignments after each step and explain the step. Remember to use the least squares assignment!

You can use the data set sketches on the next page.

Give the final quality of the clustering (TD^2). How does it compare with the solutions for $k = 2$ discussed in the lecture? Can we conclude on $k = 3$ or $k = 2$ being the better parameter choice on this data set?

Also compute solutions with $k = 4$, $k = 5$, starting from some random initial assignments of objects to clusters. What do you observe in terms of the TD^2 measure?

We use the following algorithm:

- start with k (e.g., randomly selected) points as cluster representatives (or with a random partition into k "clusters")
- repeat:
  1. assign each point to the closest representative
  2. compute new representatives based on the given partitions (centroid of the assigned points)
- until there is no change in assignment

Below is some of the calculations made in the first iteration for $K = 3$. The full result is shown in the table afterwards (calculated using python script):

Computing centroids:

$$\mu_{C_i} = \frac{1}{|C_i|} \cdot \sum_{o \in C_i} o$$

Meaning the centroid is the mean vector of all points in cluster C.

Eg. for the square class (aka. 0):

$$\mu_0 = \frac{1}{2} \cdot \left(\binom{1}{5} + \binom{2}{3} \right) = \binom{1{,}5}{4}$$

Reassign points:

To reassign a point we need to find the centroid closesed, this is done by calculating the distance to all clusters (here using eucledian distance). We do this for the point $x = (1{,}5)$

$$dist(x, C_0) = \sqrt{(1 - 1{,}5)^2 + (5 - 4)^2} \approx 1{,}12$$

$$dist(x, C_1) = \sqrt{(1 - 6{,}667)^2 + (5 - 4)^2} \approx 5{,}75$$

$$dist(x, C_2) = \sqrt{(1 - 6{,}67)^2 + (5 - 8{,}34)^2} \approx 6{,}58$$

We find the point x belongs to the class 0 (the square) which we also finds true by inspection.

Measure for compactness

for a clustering we can compute the compactness:

for cluster:

$$TD^2(C) = \sum_{p \in C} dist(p, \mu_C)^2$$

aka Sum of Squared Error.

Note: $SSQ(\mu, p) = Euclidean(\mu, p)^2 = L_2^2(\mu, p)$.

for clustering:

$$TD^2(C_1, \dots, C_k) = \sum_{i=1}^{k} TD^2(C_i)$$

We calculate the compactness for $K = 3$:

$$SSQ(\mu_0, (1,5)) = |2-1|^2 + |4-5|^2 = 2$$

$$SSQ(\mu_0, (2,3)) = |2-2|^2 + |4-3|^2 = 1$$

$$SSQ(\mu_0, (3,4)) = |2-3|^2 + |4-4|^2 = 1$$

$$TD^2(C_0) = 2 + 1 + 1 = 4$$

$$SSQ(\mu_1, (10,1)) = |10-10|^2 + |1-1|^2 = 0$$

$$TD^2(C_1) = 0$$

$$SSQ(\mu_2, (6,8)) = |6{,}75-6|^2 + |8-8|^2 \approx 0{,}563$$

$$SSQ(\mu_2, (7,7)) = |6{,}75-7|^2 + |8-7|^2 \approx 1{,}06$$

$$SSQ(\mu_2, (7,8)) = |6{,}75-7|^2 + |8-8|^2 = 0{,}0625$$

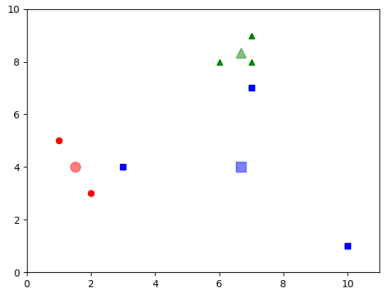
$$SSQ(\mu_2, (7,9)) = |6{,}75-7|^2 + |8-9|^2 \approx 1{,}06$$

$$TD^2(C_2) = 0{,}563 + 1{,}06 + 0{,}0625 + 1{,}06 \approx 2{,}75$$

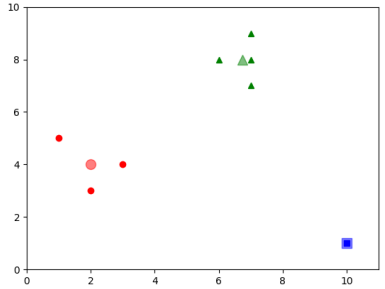
$$TD^2 = 4 + 0 + 2{,}75 = 6{,}75$$

The more clusters we have the smaller TD2, we cannot compare TD2 across different number of clusters (should use silhouette coefficient instead). Big difference going from 2 (61,5) to 3 (6,75) clusters, due to the point (10,1) now belonging to own cluster. (If we look at silhouette coefficient we find that the best is with 3 clusters (2 clustsers 0.6938))

K=3



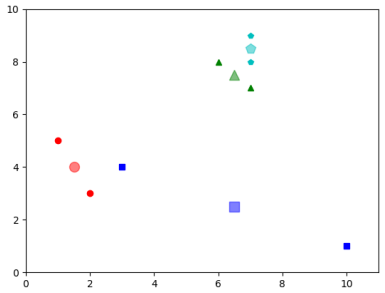
Centroids:	0	1	2
0 -> [1.5, 4]	[1, 5]	[3, 4]	[6, 8]
1 -> [6.67, 4]	[2, 3]	[10, 1]	[7, 8]
2 -> [6.67, 8.34]		[7, 7]	[7, 9]



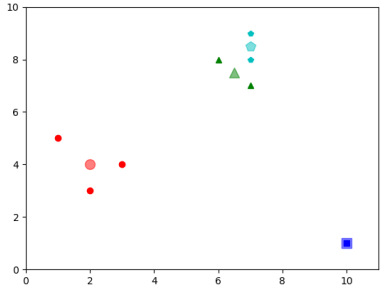
Centroids:	0	1	2
0 -> [2.0, 4.0]	[1, 5]	[10, 1]	[6, 8]
1 -> [10.0, 1.0]	[2, 3]		[7, 7]
2 -> [6.75, 8.0]	[3, 4]		[7, 8]
			[7, 9]

$TD2 = 6.75$

K=4



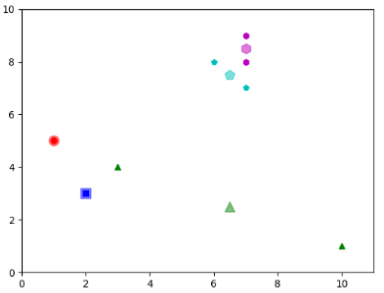
Centroids:	0	1	2	3
0 -> [2, 4]	[1, 5]	[3, 4]	[6, 8]	[7, 8]
1 -> [10, 1]	[2, 3]	[10, 1]	[7, 7]	[7, 9]
2 -> [6.5, 7.5]				
3 -> [7.0, 8.5]				



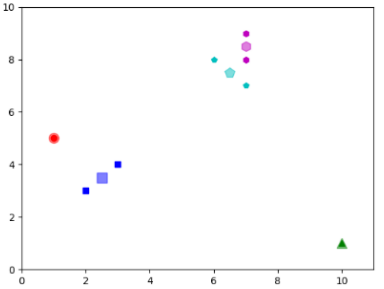
Centroids:	0	1	2	3
0 -> [2.0, 4.0]	[1, 5]	[10, 1]	[6, 8]	[7, 8]
1 -> [10.0, 1.0]	[2, 3]		[7, 7]	[7, 9]
2 -> [6.5, 7.5]	[3, 4]			
3 -> [7.0, 8.5]				

$TD2 = 5.5$

K=5



Centroids:	0	1	2	3	4
0 -> [1.0, 5.0]	[1, 5]	[2, 3]	[3, 4]	[6, 8]	[7, 8]
1 -> [2.0, 3.0]			[10, 1]	[7, 7]	[7, 9]
2 -> [6.5, 2.5]					
3 -> [6.5, 7.5]					
4 -> [7.0, 8.5]					



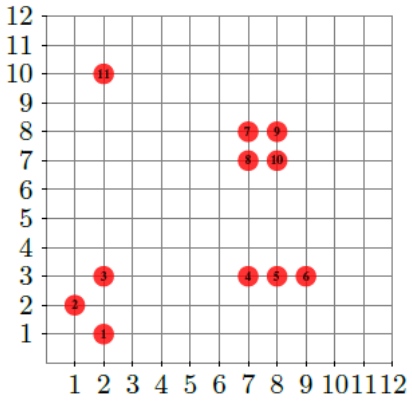
Centroids:	0	1	2	3	4
0 -> [1.0, 5.0]	[1, 5]	[2, 3]	[10, 1]	[6, 8]	[7, 8]
1 -> [2.5, 3.5]		[3, 4]		[7, 7]	[7, 9]
2 -> [10.0, 1.0]					
3 -> [6.5, 7.5]					
4 -> [7.0, 8.5]					

$TD2 = 2.5$

Ex 2

Exercise Clustering-2 Furthest First Initialization

Given the following data set with 11 objects (in $\mathbb{R}^2$):



Aim is now to perform a furthest-first initialization as seen in the lecture.

You should use the following distance measures in order to measure the distance between two points $p = (p_1, p_2)$ and $q = (q_1, q_2)$.

$$\begin{aligned} \text{dist}_2(p,q) &= \left(|p_1 - q_1|^2 + |p_2 - q_2|^2\right)^{\frac{1}{2}} \\ \text{dist}_1(p,q) &= |p_1 - q_1| + |p_2 - q_2| \\ \text{dist}_\infty(p,q) &= \max(|p_1 - q_1|, |p_2 - q_2|) \end{aligned}$$

It might help to fill out the similarity matrix noting all pair-wise distances between all points (note: only the upper triangle is required since the distance functions are symmetric). You find table sketches on the next page.

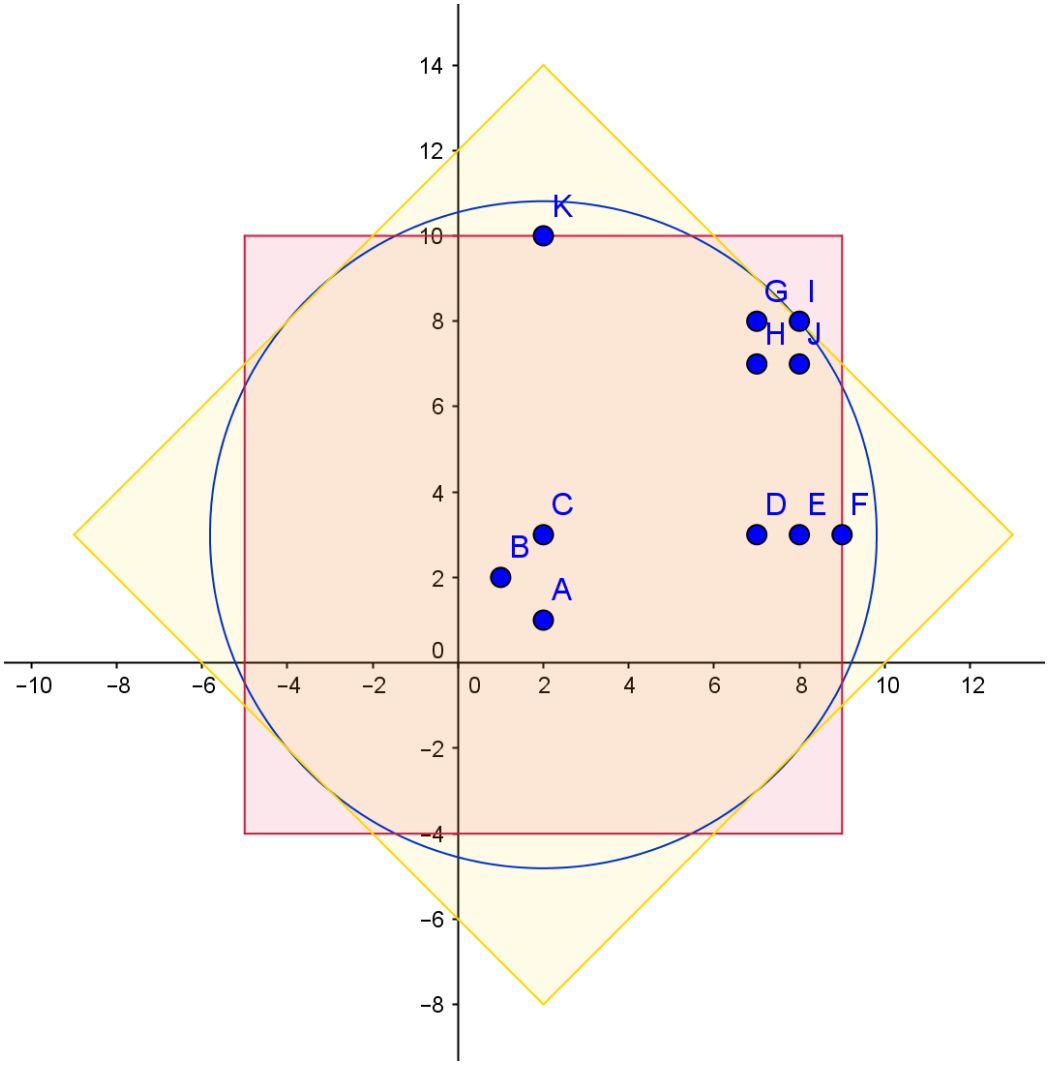
Let us choose point 3 as our first center. Define the next 3 centers according to the three different norms. (In case two or more points have the same distance, choose the point with the lower point number).

How do the norms differentiate from each other?

- choose random point as first centroid
- repeat:
  - Calculate distances for each point to center
  - Add new centroid as the one with greatest distance

We finding the first most distant point for each of the measures we get the following:

- blue = Euclidean
- yellow = Manhattan
- red = LMAX



L2	**L1**	**LMAX**
First centroid chosen as: [2, 3] distances for points [2.0, 1.4, 0.0, 5.0, 6.0, 7.0, 7.1, 6.4, 7.8, 7.2, 7.0]	First centroid chosen as: [2, 3] distances for points [2, 2, 0, 5, 6, 7, 10, 9, 11, 10, 7]	First centroid chosen as: [2, 3] distances for points [2, 1, 0, 5, 6, 7, 5, 5, 6, 6, 7]
Choosing new centroids, as most distant: [8, 8] Centroids [2, 3] [8, 8] distances for points [2.0, 1.4, 0.0, 5.0, 5.0, 5.1, 1.0, 1.4, 0.0, 1.0, 6.3]	Choosing new centroids, as most distant: [8, 8] Centroids [2, 3] [8, 8] distances for points [2, 2, 0, 5, 5, 6, 1, 2, 0, 1, 7]	Choosing new centroids, as most distant: [9, 3] Centroids [2, 3] [9, 3] distances for points [2, 1, 0, 2, 1, 0, 5, 4, 5, 4, 7]
Choosing new centroids, as most distant: [2, 10] Centroids [2, 3] [8, 8] [2, 10] distances for points [2.0, 1.4, 0.0, 5.0, 5.0, 5.1, 1.0, 1.4, 0.0, 1.0, 0.0]	Choosing new centroids, as most distant: [2, 10] Centroids [2, 3] [8, 8] [2, 10] distances for points [2, 2, 0, 5, 5, 6, 1, 2, 0, 1, 0]	Choosing new centroids, as most distant: [2, 10] Centroids [2, 3] [9, 3] [2, 10] distances for points [2, 1, 0, 2, 1, 0, 5, 4, 5, 4, 0]
Choosing new centroids, as most distant: [9, 3] Centroids [2, 3] [8, 8] [2, 10] [9, 3] distances for points [2.0, 1.4, 0.0, 2.0, 1.0, 0.0, 1.0, 1.4, 0.0, 1.0, 0.0]	Choosing new centroids, as most distant: [9, 3] Centroids [2, 3] [8, 8] [2, 10] [9, 3] distances for points [2, 2, 0, 2, 1, 0, 1, 2, 0, 1, 0]	Choosing new centroids, as most distant: [7, 8] Centroids [2, 3] [9, 3] [2, 10] [7, 8] distances for points [2, 1, 0, 2, 1, 0, 0, 1, 1, 1, 0]
Centroids [2, 3] -> p3 [8, 8] -> p9 [2, 10] -> p11 [9, 3] -> p6	Centroids [2, 3] -> p3 [8, 8] -> p9 [2, 10] -> p11 [9, 3] -> p6	Centroids [2, 3] -> p3 [9, 3] -> p6 [2, 10] -> p11 [7, 8] -> p7

Ex 3

As a warm-up on distance measures: For each of the following distance measures (Euclidean, Manhattan, maximum, weighted Euclidean, quadratic form)

$$\begin{aligned}
\text{dist}_2(p,q) &= \left(|p_1-q_1|^2+|p_2-q_2|^2+|p_3-q_3|^2\right)^{\frac{1}{2}} \\
\text{dist}_1(p,q) &= |p_1-q_1|+|p_2-q_2|+|p_3-q_3| \\
\text{dist}_\infty(p,q) &= \max(|p_1-q_1|,|p_2-q_2|,|p_3-q_3|) \\
\text{dist}_w(p,q) &= \left(w_1|p_1-q_1|^2+w_2|p_2-q_2|^2+w_3|p_3-q_3|^2\right)^{\frac{1}{2}} \\
\text{dist}_M(p,q) &= \left((p-q)M(p-q)^{\mathrm{T}}\right)^{\frac{1}{2}}
\end{aligned}$$

calculate the distance between $p=(2,3,5)$ and $q=(4,7,8)$. As w use $(1,1.5,2.5)$ and as M use both of the following:

$$M_1=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} \qquad M_2=\begin{pmatrix}1&0.9&0.7\\0.9&1&0.8\\0.7&0.8&1\end{pmatrix}$$

To calculate the distances just fill in the numbers in the given distance measure formula eg.:

$$dist_2(p,q)=\sqrt{|2-4|^2+|3-7|^2+|5-8|^2}=\sqrt{29}\approx 5{,}39$$

We can also just calculate it using a python script. If we are to implement other functions such as clustering theses different distance functions would come in very handy:

Euclidean distance: 5.38516480713

Manhattan distance: 9

Maximum distance: 4

Weighted Euclidean distance: 7.10633520178

Quadratic form M1: 5.38516480713

Quadratic form M2: 8.42614977318

Observe that the $M1$ matrix is the same as the Euclidean distance. This is due to the matrix being the identity matrix, which multiplies with weights of 1. Remember that matrix multiplication happens in the following way:

$$\begin{bmatrix} x & y \\ & \end{bmatrix}\cdot\begin{bmatrix} a & \\ b & \end{bmatrix}=\begin{bmatrix} xa+yb & \\ & \end{bmatrix}$$

dotting each row by each column.

Given 5 pictures as in Figure 1 with 36 pixels each.

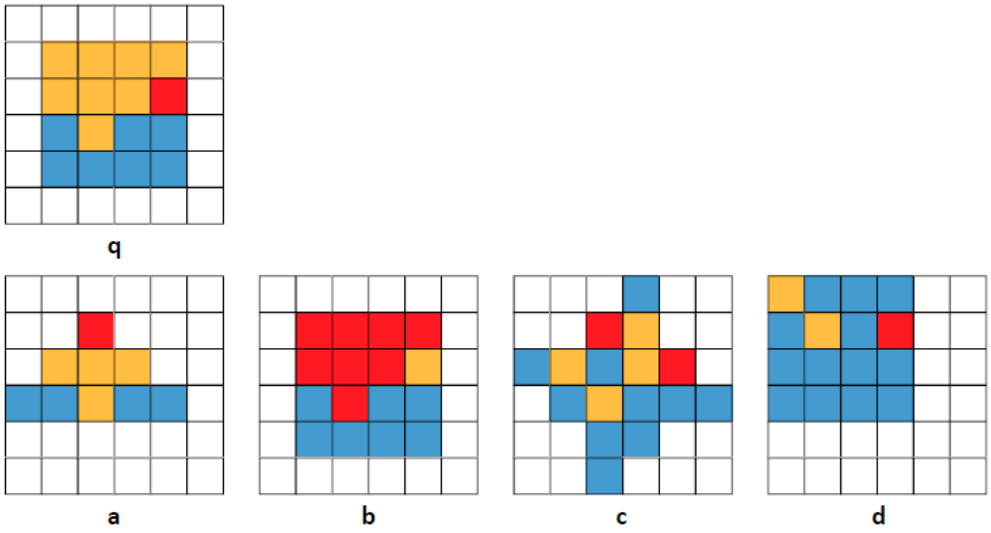


Figure 1: 6 × 6 pixel pictures

(a) Extract from each picture a color histogram with the bins *red*, *orange*, and *blue* (the white pixels are ignored).

(b) Which pictures are most similar to the query q, using Euclidean distance? Give a ranking according to similarity to q.

(c) The results are not entirely satisfactory. What could you change in the feature extraction or in the distance function to get better results? Report the improved feature extraction and features or the improved distance function.

a. Histograms (red, orange, blue):
   - $q(1,8,7)$
   - $a(1,4,4)$
   - $b(8,1,7)$
   - $c(2,4,10)$
   - $d(1,2,13)$

b. calculating distances using python script and the Euclidian distance:
   - $dist(q,a) = (a, 5.0)$
   - $dist(q,c) = (c, 5.1)$
   - $dist(q,d) = (d, 8.5)$
   - $dist(q,b) = (b, 9.9)$

c. improvements
   - Taking color similarity into account (similarity matrix)
   - Taking white pixels into account.
   - Taking shape into account
   - Taking position similarity into account
   - Taking neighboring pixels into account

Debatably, picture b is more similar to q than a or d are. The problem is that the Euclidean distance takes each color individually to compute the distance but does not take similarity between different colors (i.e., bins in the histogram) into account.

A solution would be to use the quadratic form (a.k.a. Mahalanobis-) distance. We need a similarity matrix to define the (subjective) similarity of bins with each other:

$$A = \begin{pmatrix} 1 & 0{,}9 & 0 \\ 0{,}9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Calculating the distances with our new measure we get:

- $dist(q, b) = (b, 3.1)$
- $dist(q, c) = (c, 4.3)$
- $dist(q, a) = (a, 5.0)$
- $dist(q, d) = (d, 8.5)$

Here we say that the first attribute (the color red) is very similar to the second attribute (the color orange).

We look at what happens when we do the multiplication:

$$dist(q, b) = \sqrt{(q - b) \cdot A \cdot (q - b)^T}$$

In this formula we assume vectors to be row vectors.

$$= \sqrt{\big((1\ \ 8\ \ 7) - (1\ \ 8\ \ 7)\big) \cdot A \cdot \big((1\ \ 8\ \ 7) - (1\ \ 8\ \ 7)\big)^T}$$

$$= \sqrt{(-7\ \ 7\ \ 0) \cdot A \cdot (-7\ \ 7\ \ 0)^T}$$

$$= \sqrt{(-7\ \ 7\ \ 0) \cdot \begin{pmatrix} 1 & 0{,}9 & 0 \\ 0{,}9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} -7 \\ 7 \\ 0 \end{pmatrix}}$$

We here observe that we put the weight 0,9 on the first and second attribute:

$$= \sqrt{(-7\ \ 7\ \ 0) \cdot \begin{pmatrix} -7 + 0{,}9 \cdot 7 \\ 0{,}9 \cdot -7 + 7 \\ 0 \end{pmatrix}}$$

$$= \sqrt{(-7\ \ 7\ \ 0) \cdot \begin{pmatrix} -0{,}7 \\ 0{,}7 \\ 0 \end{pmatrix}}$$

$$= \sqrt{(-7 \cdot -0{,}7 + 0{,}7 \cdot 7)} \approx 3{,}13$$