Exercise Sheet 3

(until 10.05.2019)

1. The following two-dimensional data set is given. Perform a *K-means* clustering with *K=3* using the Euclidean distance. Use the first three points as initial centroids. For each algorithm iteration, specify the distances between centroids and all points and calculate the changed centroids after each reassignment of the points.

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12
x	2.0	2.0	2.0	2.5	2.5	3.0	4.0	4.0	4.5	4.5	4.5	4.5
y	1.0	1.5	2.0	1.0	2.0	4.0	1.0	2.5	1.0	1.5	2.5	3.0

2. A school would like to group its pupils according to their performance at two intermediate examinations. It is assumed that there are at least 2 clusters of pupils. Load the file ***clustering-student-mat.csv***. The file contains for each of the two exams the number of points scored for a total of 395 students.
   Perform a *K-means* clustering for each $k \in \{2, 3,...,8\}$. Display the cluster assignments of the points in a scatter plot.
3. For the clustering in task 2, use the silhouette coefficient to find the optimal value for the number of clusters *K*. Evaluate the result for the representativeness of the centroids with respect to their cluster.
4. The following distance matrix is given. Perform agglomerative hierarchical clustering with single and complete linkage. Display the result in a dendrogram. The dendrogram should represent the order in which the points are joined.

	a	b	c	d	e
a	0.00	0.02	0.90	0.36	0.53
b	0.02	0.00	0.65	0.15	0.24
c	0.90	0.65	0.00	0.59	0.45
d	0.36	0.15	0.59	0.00	0.56
e	0.53	0.24	0.45	0.56	0.00

Dataset for task 2: http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/clustering-student-mat.csv