C A U

Kiel University
Christian-Albrechts-Universität zu Kiel

Faculty of Engineering

Intelligent Systems

Exercise 8 - Clustering

Simon Reichhuber

December 16, 2019

University of Kiel, Winter Term 2019

Table of Content

1. Single, Complete und Average Linkage
2. c-Means basics
3. Apply c-Means Clustering

Single, Complete und Average Linkage

1. A

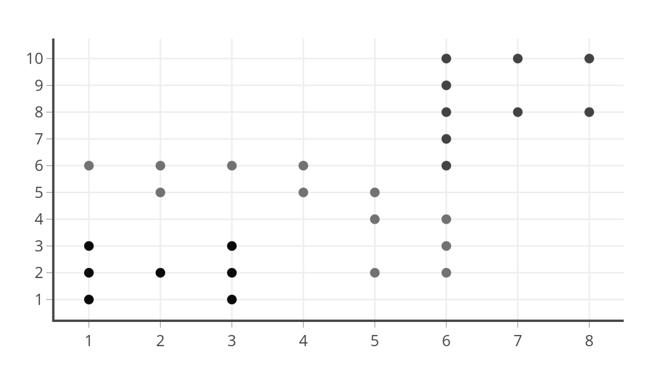
A. **How does Single Linkage Clustering work? Visualise the procedure using the data set of the figure with $C = 3$ and with $C = 2$. As distance measure choose the Manhattan distance.**

B. What are Pros and Cons of Single Linkage in comparison to complete Linkage regarding the treatment of outliers and the tendency of producing chains?

C. What is the difference between Single Linkage, Complete Linkage, and Average Linkage? How would Complete or Average Linkage cluster the data points in the figure with $C = 2$ and the Manhattan distance?

1. A Single Linkage Clustering

How does Single Linkage Clustering work? Visualise the procedure using the data set of the figure with $C = 3$ and with $C = 2$. As distance measure choose the Manhattan distance.

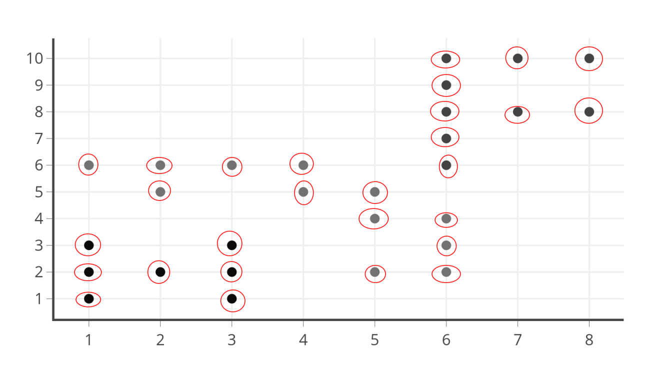


A) - Single Linkage

Given a set of samples x_i mit $i = 1, ..., N$ and additionally a number of cluster c that are required to be found. ($N \geq c$).

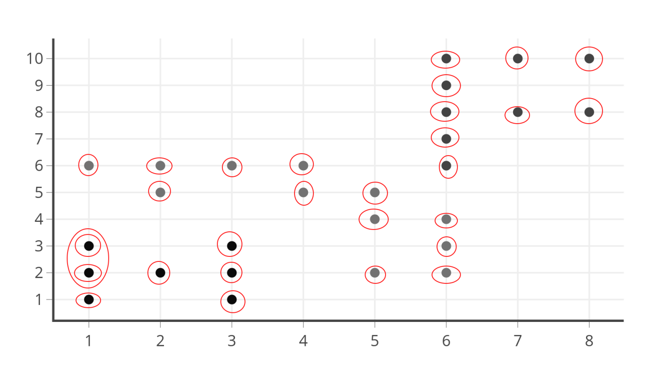
1. First, partition the samples such that every sample is assigned to a different cluster. Given N samples results in N clusters:
   - $C_i = \{x_i\}$
2. These N clusters are merged stepwise:
   - 2.1 Determine two clusters with minimum distance according to the criterion nearest neighbor and merge them.
   - 2.2 If the number of clusters is still above c, repeat step 2.1.

A) Single Linkage Example



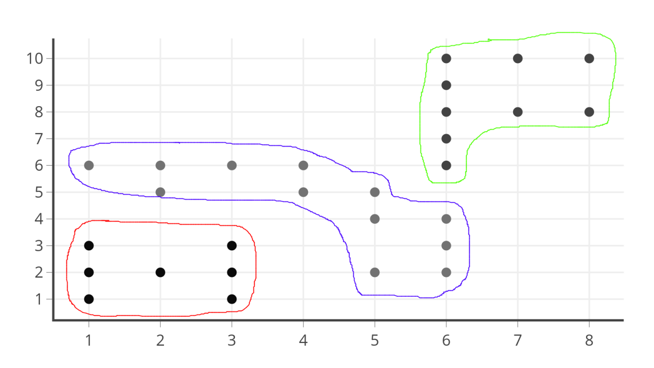
Every sample is assigned to a different clusterr.

A) Single Linkage example C=3



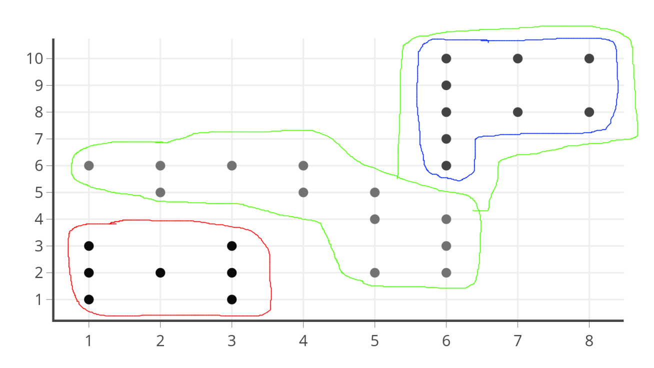
Merge two clusters, until the number of required clusters $c = 3$ is reached.

A) Single Linkage example C=3



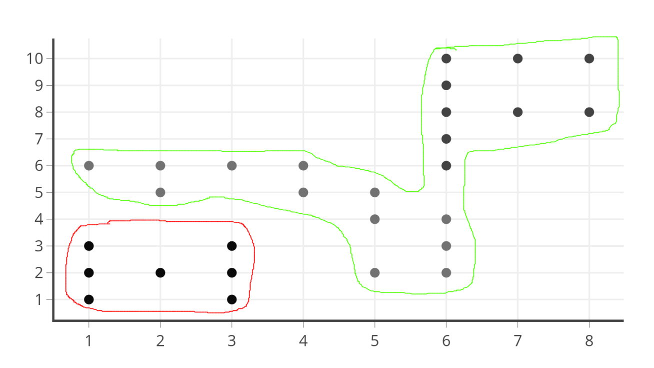
Resulting 3 clusters.

A) Single Linkage example C=2



Merge two clusters, until the number of required clusters $c = 2$ is reached.

A) Single Linkage Beispiel C=2



The possible result shows a poor clustering. The middle cluster should be merged to the bottom cluster → chaining problem

1. B

A. How does Single Linkage Clustering work? Visualise the procedure using the data set of the figure with $C = 3$ and with $C = 2$. As distance measure choose the Manhattan distance.

B. What are Pros and Cons of Single Linkage in comparison to complete Linkage regarding the treatment of outliers and the tendency of producing chains?

C. What is the difference between Single Linkage, Complete Linkage, and Average Linkage? How would Complete or Average Linkage cluster the data points in the figure with $C = 2$ and the Manhattan distance?

1. B Sinlge Linkage Clustering

What are Pros and Cons of Single Linkage in comparison to complete Linkage regarding the treatment of outliers and the tendency of producing chains?

Pros:

- Outliers are detected, because they will be added lastly

Cons:

- Only two samples matters for merging two clusters $\rightarrow$ What if these are noise or outliers?
- Chaining problem: long chains are able to form clusters with a large maximum point distance (Remark: sometimes chaining is required for clustering geographical data, i.e. rivers)

1. C

A. How does Single Linkage Clustering work? Visualise the procedure using the data set of the figure with $C = 3$ and with $C = 2$. As distance measure choose the Manhattan distance.

B. What are Pros and Cons of Single Linkage in comparison to complete Linkage regarding the treatment of outliers and the tendency of producing chains?

C. What is the difference between Single Linkage, Complete Linkage, and Average Linkage? How would Complete or Average Linkage cluster the data points in the figure with $C = 2$ and the Manhattan distance?

1. C Complete Linkage Clustering

- Similiarly as Single Linkage but maximum distant neighbor as distance measure

Pros:

- No chaining as in Single Linkage
  For example: In Exercise 1A with $c = 2$ the bottom and middle cluster will be merged definitely.

Cons:

- Outliers will be merged to clusters most likely and won't be detected as outliers

1. C Average Linkage Clustering

- Similiarly as Single Linkage but average distant neighbor as distance measure
- Average Linkage is a trade-off between Single Linkage and Complete Linkage
- In many cases this approach succeeds.
- For example:
  The middle clsuter in exercise 1A with $c = 2$ will be merged correctly to the bottom cluster.

c-Means basics

3. A

A. **Visualise and explain with the help of the figure how the *c*-Means algorithm works.**

B. What are Pros and Cons of the *c*-Means algorithm?

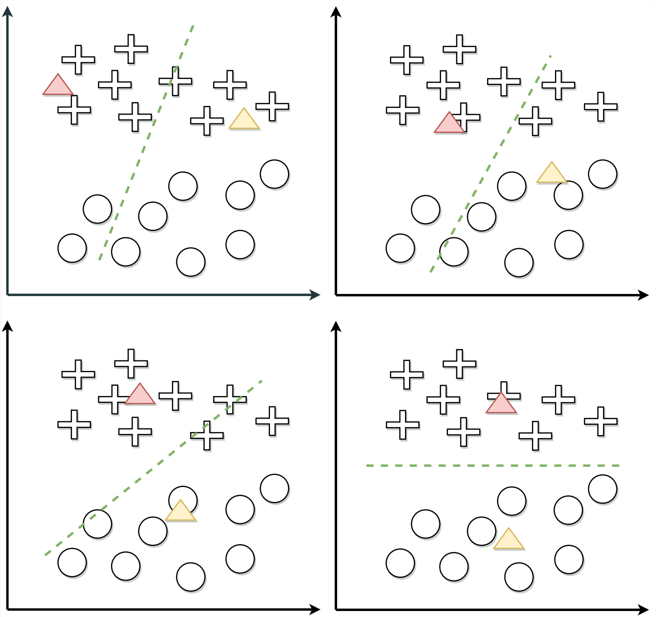
C. Which steps can be applied to optimise the clustering results?

3. A C-means procedure I

Given N d-dimensional samples $X = \{x_i\}_{i=1}^{N} \subset \mathbb{R}^d$ and the number of required clusters $c \in \mathbb{N}$

1. Distribute c cluster centres $c_j, j = 1, \ldots, c$ (arbitrary) in the space $\mathbb{R}^d$
2. Assign every sample to the nearest cluster centre $class(x_i) = argmin_{j=1,\ldots,k} ||x_i - c_j||_2$
3. Distribute new cluster centres $c_j = \frac{1}{|\{x \in X | class(x) = j\}|} \sum_{x \in \{x' \in X | class(x') = j\}} x$
4. Repeat steps 2)-3) and terminate after k iterations or if no new cluster assignement took place

3. A C-means procedure II



3. B

A. Visualise and explain with the help of the figure how the *c*-Means algorithm works.

B. What are Pros and Cons of the *c*-Means algorithm?

C. Which steps can be applied to optimise the clustering results?

What are Pros and Cons of the c-Means algorithm?

Pros:

- Easy, understandable
- Every sample is assigned to a cluster
- Algorithm terminates

Cons:

- Greedy procedure $\rightarrow$ convergence at local minimum possible
- Number of clusters has to be known or estimated
- Also outliers will be assigned to clusters
- Sensible according to poor initial cluster centres placement

3. C

A. Visualise and explain with the help of the figure how the *c*-Means algorithm works.

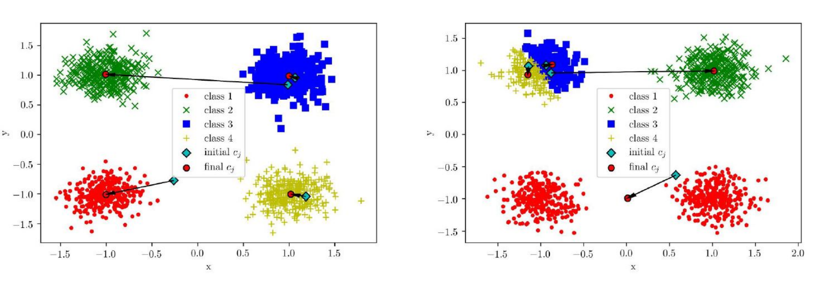
B. What are Pros and Cons of the *c*-Means algorithm?

C. Which steps can be applied to optimise the clustering results?

3. C

Which steps can be applied to optimise the clustering results?

In the figure: How can we prevent the clustering in the right image?



a) Successful clustering

b) Failed clustering

3. C C-MEANS TUNING

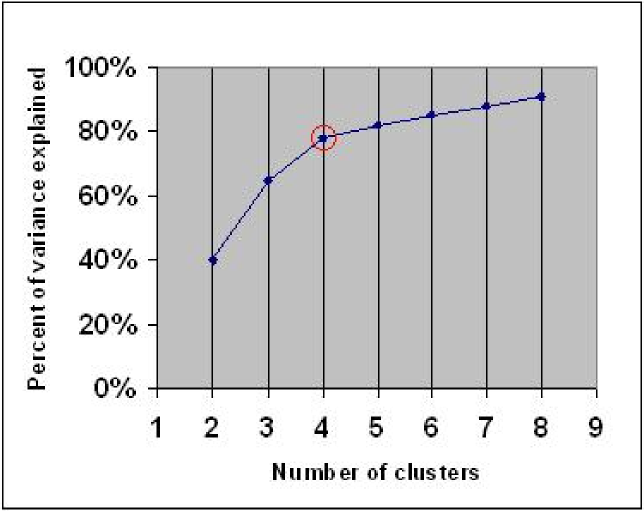
Unlucky placement of cluster centres:

- Repeat procedure with different randomised initial cluster centres placements
- Intelligent cluster centres placement

Not suitable number of clusters:

- Begin with small *c* and evaluate the minimum cluster distance with ascending *c*
- Decide according to elbow method

3. C Elbow-criterion



(via Wikimedia Commons, CC BY-SA 3.0)

3. C CLEVER CLUSTER CENTRE PLACEMENT

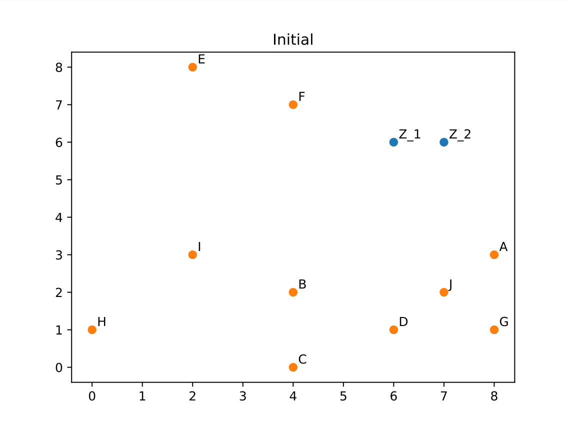
- **Solution 1:**
  Increase c and merge similar clusters
- **Solution2:**
  Apply clever cluster centre distribution:
  - Approach 1: Random distribution
    a) Select c samples randomly as cluster centre
    b) Choose random points within minimum sphere containing all samples.
  - Approach 2: Sort samples $x_1, \ldots, x_N$ according distance to global centre $c_{global} = \frac{1}{N}\sum_{i=1}^{N} x_i$
    and select nearest $(1 + (j-1) \cdot \lfloor \frac{N}{c} \rfloor)$-th sample as cluster centre.

Apply c-Means Clustering

2. A

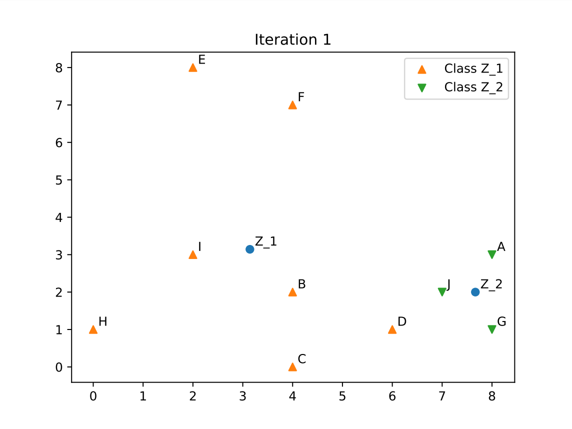
A. Proceed 4 iterations of the c-Means clustering on the points given in the figure. Note the Euclidean distances into the table. Remark: You can use a ruler to measure the Euclidean distances.

2. A Initial



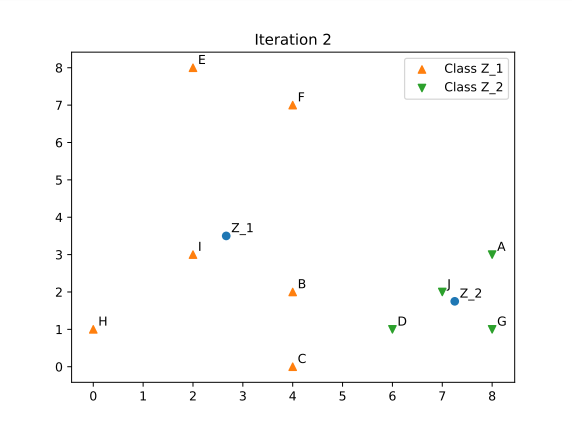
Iteration		A	B	C	D	E	F	G	H	I	J
0	$Z_1 = (6.0, 6.0)$	3.6	4.5	6.3	5.0	4.5	2.2	5.4	7.8	5.0	4.1
0	$Z_1 = (7.0, 6.0)$	3.2	5.0	6.7	5.1	5.4	3.2	5.1	8.6	5.8	4.0

2. A Iteration 1



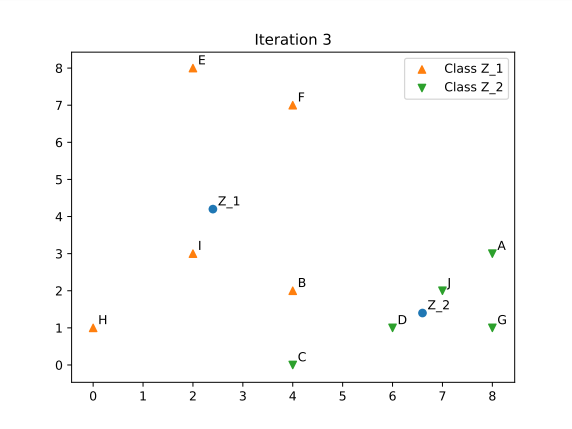
Iteration		A	B	C	D	E	F	G	H	I	J
1	$Z_1 = (3.1, 3.1)$	4.9	1.4	3.3	3.6	5.0	4.0	5.3	3.8	1.2	4.0
1	$Z_1 = (7.7, 2.0)$	1.1	3.7	4.2	1.9	8.3	6.2	1.1	7.7	5.8	0.7

2. A Iteration 2



Iteration		*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
2	$Z_1 = (2.7, 3.5)$	5.4	2.0	3.7	4.2	4.5	3.7	5.9	3.7	0.8	4.6
2	$Z_1 = (7.2, 1.8)$	1.5	3.3	3.7	1.5	8.2	6.2	1.1	7.3	5.4	0.4

2. A Iteration 3



Iteration		*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
3	$Z_1 = (2.4, 4.2)$	5.7	2.7	4.5	4.8	3.8	3.2	6.4	4.0	1.3	5.1
3	$Z_1 = (6.6, 1.4)$	2.1	2.7	3.0	0.7	8.0	6.2	1.5	6.6	4.9	0.7