University of Augsburg, Institute of Computer Science — SS 2019
Prof. Dr. P. Fischer — 24. June 2019
J. Kastner, L. Rudenko — Solution 7

Analyzing Massive Data Sets

Exercise 1: Inverted Index (homework)

Documents $D_1 - D_6$ and a query Q:

- D_1: ”the cat lies on this bench”
- D_2: ”the black cat on your desk”
- D_3: ”the red cat on my table”
- D_4: ”the red cat lies on my table”
- D_5: ”the black rabbit sits under this desk”
- D_6: ”this white rabbit lies behind your bench”
- Q: ”your rabbit sits behind this bench under our desk”

a) **Inverted index**:

- *the*: D_1, D_2, D_3, D_4, D_5
- *cat*: D_1, D_2, D_3, D_4
- *lies*: D_1, D_4, D_6
- *on*: D_1, D_2, D_3, D_4
- *this*: D_1, D_5, D_6
- *bench*: D_1, D_6
- *black*: D_2, D_5
- *your*: D_2, D_6
- *desk*: D_2, D_5
- *red*: D_3, D_4
- *my*: D_3, D_4
- *table*: D_3, D_4
- *rabbit*: D_5, D_6
- *sits*: D_5
- *under*: D_5
- *white*: D_6
- *behind*: D_6

b) **Candidates**:

Q: ”your rabbit sits behind this bench under our desk”

- *your*: D_2, D_6
- *rabbit*: D_5, D_6

- *sits*: D_5
- *behind*: D_6
- *this*: D_1, D_5, D_6
- *bench*: D_1, D_6
- *under*: D_5
- *desk*: D_2, D_5

The candidates are the following documents: D_1, D_2, D_5, D_6

c) Documents with the similarity **at least 0.4**:

- D_1, Q: $2/13 < 0.4$
- D_2, Q: $2/13 < 0.4$
- D_5, Q: $5/11 \approx 0.45 > 0.4$
- D_6, Q: $5/11 \approx 0.45 > 0.4$

The similarity between query Q and the documents D_5 and D_6 is more than 0.4.

Exercise 2: Hierarchical Clustering (homework)

First we compute the distances between the words and write them into a matrix. We measure the distance between two words with **Levenshtein** distance (chapter 4, slide 18). The number of edit operation has to be determine. **Levenshtein** allows *removing*, *adding* and *replacing* as edit operations.

For example the **Levenshtein** distance between **able** and **notable** is 3, because we have to remove the first three characters n, o and t to turn **notable** into **able** (or we have to add n, o and t at the beginning of **able** to turn it into **notable**).

	able	notable	tabloid	taboo	tabby	disable	lovably	stab	ability	labor
able	0	3	4	3	3	3	4	4	4	3
notable	3	0	5	4	4	3	4	4	6	8
tabloid	4	5	0	3	4	6	6	5	5	4
taboo	3	4	3	0	2	5	5	3	6	2
tabby	3	4	4	2	0	5	4	3	5	3
disable	3	3	6	5	5	0	4	5	7	5
lovably	4	4	6	5	4	4	0	5	6	4
stab	4	4	5	3	3	5	5	0	7	4
ability	4	6	5	6	5	7	6	7	0	6
labor	3	8	4	2	3	5	4	4	6	0

- **single linkage**: minimum distance of all element pairs from the two clusters.

  1) $d(tabby, taboo) = d(taboo, labor) = 2$. We decide to combine **tabby** and **taboo** to the cluster {tabby, taboo}.

  2) $d(taboo, labor) = 2 = D(labor, \{tabby, taboo\})$, so we combine **labor** to the cluster {**tabby**, **taboo**} and have new cluster {labor, tabby, taboo}.

3) $d(able, notable) = d(able, taboo) = d(able, tabby) = d(able, disable) = d(able, labor) = d(notable, disable) = d(tabloid, taboo) = d(taboo, stab) = d(tabby, stab) = d(tabby, labor) = 3$. We decide arbitrarily to combine **able** and **notable** to the new cluster {able, notable}.

4) Because of *single linkage* approach, the distance $D(disable, \{able, notable\}) = = d(notable, disable) = d(able, disable) = 3$. We combine **disable** to the cluster {**able**, **notable**} and have new cluster {disable, able, notable}. (The pair (notable, disable) with distance 3 was chosen arbitrarily).

5) The next random candidate with distance 3 is the pair (tabby, stab): $D(stab, \{tabby, taboo, labor\}) = d(tabby, stab) = 3$. We combine **stab** to the cluster {**tabby**, **taboo**, **labor**} and have new cluster {stab, tabby, taboo, labor}.

6) We still have the candidates with distance 3. We can even combine our two big clusters: {**notable, able, disable**} and {**stab, tabby, taboo, labor**} because $D(\{$**notable, able, disable**$\}, \{$**stab, tabby, taboo, labor**$\}) = d(able, taboo) = d(able, tabby) = d(able, labor) = 3$. But we decide to combine **tabloid** to the cluster {**stab, tabby, taboo, labor**}: $D(tabloid, \{stab, tabby, taboo, labor\}) = d(tabloid, taboo) = 3$. And we have the new cluster: {stab, tabby, taboo, labor, tabloid}

7) In this step we have to combine **two big clusters**, because $D(lovably, \{disable, able, notable\}) = d(lovably, able) = d(lovably, notable) = d(lovably, disable) = 4$, $D(lovably, \{stab, tabby, taboo, labor, tabloid\}) = d(lovably, tabby) = d(lovably, labor) = 4$, $D(ability, \{disable, able, notable\}) = d(ability, able) = 4$. All other candidates ($D(lovably, ability)$ and $D(ability, \{stab, tabby, taboo, labor, tabloid\})$ have larger distance to each other). But $D(\{notable, able, disable\}, \{stab, tabby, taboo, labor, tabloid\}) = d(able, taboo) = d(able, tabby) = d(able, labor) = 3$, so we combine these two clusters to {notable, able, disable, stab, tabby, taboo, labor, tabloid}

8) We a looking for **three clusters**, each word is in a cluster by itself. So we are ready. We have one big cluster and two very small clusters containing only one word, see Figure 1.

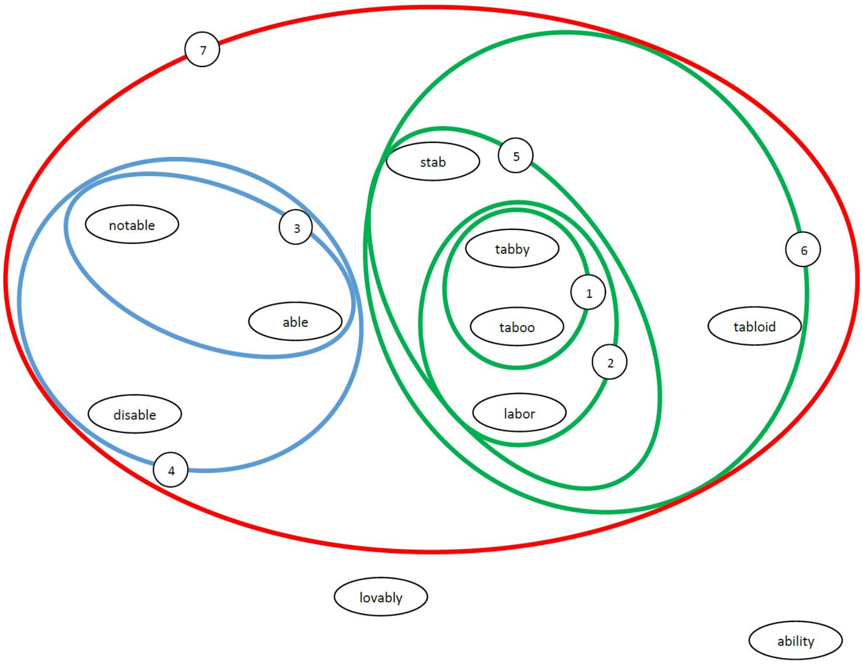

Abbildung 1: Single Linkage

Note, that the result can be different depending on the decision which two clusters are combined in each step.

- **complete linkage**: maximum distance of all element pairs from the two clusters.

  1) $d(tabby, taboo) = d(taboo, labor) = 2$. We decide to combine **tabby** and **taboo** to the cluster {tabby, taboo}.

  2) The minimum distance we have now is 3: $D(labor, \{tabby, taboo\}) = d(labor, tabby) = d(able, notable) = d(able, taboo) = d(able, tabby) = d(able, disable) = d(able, labor) = d(notable, disable) = d(tabloid, taboo) = d(taboo, stab) = d(tabby, stab) = d(tabby, labor) = 3$. We decide to combine **labor** to the cluster {**tabby, taboo**} and have new cluster {tabby, taboo, labor}.

  3) Now we randomly take one of the candidates with distance 3, e.g. **able** and **notable** and combine them to the new cluster {able, notable}.

  4) The distance $D(disable, \{able, notable\}) = d(notable, disable) = d(able, disable) = 3$. We combine **disable** to the cluster {**able**, **notable**} and have the new cluster {disable, able, notable}. (The decision to combine **disable** to the cluster {**able, notable**} is coincidental).

  5) Because of *complete linkage* the distance between two big clusters is 8!
  $D(\{disable, able, notable\}, \{tabby, taboo, labor\}) = d(notable, labor) = 8$.
  So we are looking other candidates with the smaller distance:

  - $D(stab, \{tabby, taboo, labor\}) = d(stab, labor) = 4$,
    $D(stab, \{disable, able, notable\}) = d(stab, disable) = 5$.
  - $D(tabloid, \{tabby, taboo, labor\}) = d(tabloid, labor) = d(tabloid, tabby) = 4$,
    $D(tabloid, \{disable, able, notable\}) = d(tabloid, disable) = 5$.
  - $D(lovably, \{tabby, taboo, labor\}) = d(lovably, taboo) = 5$,
    $D(lovably, \{disable, able, notable\}) = d(lovably, able) = d(lovably, notable) = d(lovably, disable) = 4$.
  - $D(ability, \{tabby, taboo, labor\}) = d(ability, taboo) = d(ability, labor) = 6$,
    $D(ability, \{disable, able, notable\}) = d(ability, disable) = 7$.
  - $d(ability, tabloid) = 5$

  And we decide to combine **stab** to the cluster {**tabby, taboo, labor**} and now we have a new cluster {stab, tabby, taboo, labor}.

  6) In this step we can combine the other candidate with distance 4: **lovably** and cluster {**disable, able, notable**}: $D(lovably, \{disable, able, notable\}) = d(lovably, able) = d(lovably, notable) = d(lovably, disable) = 4$. The new cluster is {lovably, disable, able, notable}.

  7) We have:

  - $D(\{lovably, disable, able, notable\}, \{stab, tabby, taboo, labor\}) = d(notable, labor) = 8$.
  - $D(tabloid, \{stab, tabby, taboo, labor\}) = d(tabloid, stab) = 5$,
    $D(tabloid, \{lovably, disable, able, notable\}) = d(tabloid, disable) = d(tabloid, lovably) = 6$.
  - $D(ability, \{stab, tabby, taboo, labor\}) = d(ability, stab) = 7$,
    $D(ability, \{lovably, disable, able, notable\}) = d(ability, disable) = 7$.
  - $d(ability, tabloid) = 5$.

  We combine **ability** and **tabloid** to the cluster {ability, tabloid}.

  8) We a looking for **three clusters**, each word is in a cluster by itself. So we are ready, see Figure 2.

  Note, that the result can be different depending on the decision which two clusters are combined in each step.

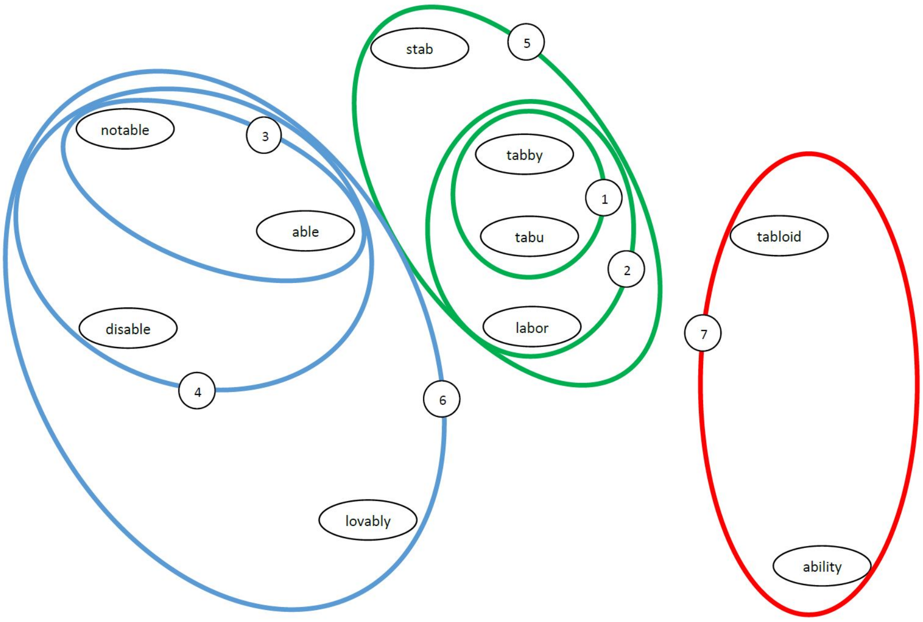

Abbildung 2: Complete Linkage

Exercise 3: k-means Clustering (live)

The solution was discussed in the exercise.

Exercise 4: BFR (live)

At the beginning we have two clusters: $\{AD\}$ with $A(1,9)$ and $D(2,9)$ as well as $\{KL\}$ with $K(11,3)$ and $L(10.2)$. First we calculate N, SUM, $SUMSQ$ plus variance and standard deviation σ for each dimension in each cluster:

- **cluster** $\{AD\}$:
  - $N = 2$
  - $SUM = \sum_{p \in C} p = \binom{1}{9} + \binom{2}{9} = \binom{3}{18}$
  - $SUMSQ = \sum_{p \in C} p^2 = \binom{1}{9}^2 + \binom{2}{9}^2 = \binom{5}{162}$
  - $centroid_{\{AD\}} = SUM/N = \binom{1.5}{9}$

  $variance\ \sigma_i^2 = (SUMSQ_i/N) - (SUM_i/N)^2$

  - σ^2 in the first dimension $(x) = 5/2 - (3/2)^2 = 0.25$
  - $standard\ deviation\ (\sigma)$ in the first dimension $(x) = \sqrt{0.25} = 0.5$
  - $\sigma_x^2 = 162/2 - (18/2)^2 = 0$
  - $\sigma_y = 0$
- **cluster** $\{KL\}$:
  - $N = 2$
  - $SUM = \binom{10}{2} + \binom{11}{3} = \binom{21}{5}$
  - $SUMSQ = \binom{10}{2}^2 + \binom{11}{3}^2 = \binom{221}{13}$
  - $centroid_{\{KL\}} = SUM/N = \binom{10.5}{2.5}$

  - $\sigma_x^2 = 221/2 - (21/2)^2 = 0.25$
  - $\sigma_x = \sqrt{0.25} = 0.5$
  - $\sigma_y^2 = 13/2 - (5/2)^2 = 0.5$
  - $\sigma_y = \sqrt{0.25} = 0.5$

Now we can read the points from the first chunk and see whether they belong to the clusters (discard set, DS), compressed set (CS, minicluster) or retained set (RS). Closeness to the cluster is determined with **Mahalanobis distance**.

Mahalanobis distance: $d(p, c) = \sqrt{\sum_{i=1}^{d} (\frac{p_i - c_i}{\sigma_i})^2}$, where d is the number of dimensions, σ_i - standard deviation of points (p) in the cluster (c) in the ith dimension.

- $B(2, 8)$:
  - B in cluster $\{AD\}$?
    $d(B, centroid_{\{AD\}}) = \sqrt{(\frac{2-1.5}{0.5})^2 + (\frac{8-9}{0})^2} = \sqrt{(\frac{0.5}{0.5})^2} = 1 < 2.5$ (threshold)
    $\Rightarrow B$ can be part of this cluster.
  - B in cluster $\{KL\}$?
    $d(B, centroid_{\{KL\}}) = \sqrt{(\frac{2-10.5}{0.5})^2 + (\frac{8-2.5}{0.5})^2} = \sqrt{(-17)^2 + (11)^2} = \sqrt{289 + 121} = \sqrt{410} \approx 20.25 > 2.5 \Rightarrow B$ is not part of this cluster.

  $B(2, 8)$ is part of the cluster $\{AD\}$ and we have new cluster $\{ABD\}$ with $A(1, 9)$, $B(2, 8)$ and $D(2, 9)$. Update N, SUM, $SUMSQ$ as well as variance σ^2 and standard deviation σ for this cluster:

  - $N = 3$
  - $SUM = \binom{3}{18} + \binom{2}{8} = \binom{5}{26}$
  - $SUMSQ = \binom{5}{162} + \binom{2}{8}^2 = \binom{9}{226}$
  - $centroid_{\{ABD\}} = SUM/N = \binom{1.67}{8.67}$
  - $\sigma_x^2 = 9/3 - (5/3)^2 = 3 - 2.79 = 0.21$
  - $\sigma_x = \sqrt{0.21} \approx 0.46$
  - $\sigma_y^2 = 226/3 - (26/3)^2 \approx 75.33 - 75.11 = 0.22$
  - $\sigma_y = \sqrt{0.22} \approx 0.47$
- $F(3, 3)$
  - F in cluster $\{ABD\}$?
    $d(F, centroid_{\{ABD\}}) = \sqrt{(\frac{3-1.67}{0.46})^2 + (\frac{3-8.67}{0.47})^2} = \sqrt{(\frac{1.33}{0.46})^2 + (\frac{-5.67}{0.47})^2} \approx \sqrt{(2.89)^2 + (-12.06)^2}$
    $\approx \sqrt{8.36 + 145.44} = \sqrt{153.8} \approx 12.4 > 2.5 \Rightarrow F$ is not part of this cluster.
  - F in cluster $\{KL\}$?
    $d(F, centroid_{\{KL\}}) = \sqrt{(\frac{3-10.5}{0.5})^2 + (\frac{3-2.5}{0.5})^2} = \sqrt{(-15)^2 + (1)^2} = \sqrt{225 + 1} = \sqrt{226} \approx 15.03 > 2.5 \Rightarrow F$ is not part of this cluster.

  $F \notin$ any DS (cluster).

- $I(8,9)$
  - I in cluster $\{ABD\}$?
    $d(I, centroid_{\{ABD\}}) = \sqrt{(\frac{8-1.67}{0.46})^2 + (\frac{9-8.67}{0.47})^2} \approx \sqrt{(13.76)^2 + (0.7)^2} \approx \sqrt{189.34 + 0.49} = \sqrt{189.83} \approx 13.78 > 2.5 \Rightarrow I$ is not part of this cluster.
  - I in cluster $\{KL\}$?
    $d(I, centroid_{\{KL\}}) = \sqrt{(\frac{8-10.5}{0.5})^2 + (\frac{9-2.5}{0.5})^2} = \sqrt{(-5)^2 + (13)^2} = \sqrt{25 + 169} = \sqrt{194} \approx 13.9 > 2.5 \Rightarrow I$ is not part of this cluster.

  $I \notin$ any DS (cluster).
- $N(11,2)$
  - N in cluster $\{ABD\}$?
    $d(N, centroid_{\{ABD\}}) = \sqrt{(\frac{11-1.67}{0.46})^2 + (\frac{2-8.67}{0.47})^2} \approx \sqrt{(20.28)^2 + (-14.19)^2}$
    $\approx \sqrt{411.28 + 201.36} = \sqrt{612.64.34} \approx 24.75 > 2.5 \Rightarrow N$ is not part of this cluster.
  - N in cluster $\{KL\}$?
    $d(N, centroid_{\{KL\}}) = \sqrt{(\frac{11-10.5}{0.5})^2 + (\frac{2-2.5}{0.5})^2} = \sqrt{(1)^2 + (-1)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.4 < 2.5 \Rightarrow N$ is part of this cluster.

  We have new cluster $\{KLN\}$ with $K(11,3)$, $L(10,2)$ and $N(11,2)$. Update N, SUM, $SUMSQ$ as well as variance and standard deviation σ for this cluster:

  - $N = 3$
  - $SUM = \binom{21}{5} + \binom{11}{2} = \binom{32}{7}$
  - $SUMSQ = \binom{221}{13} + \binom{11}{2}^2 = \binom{342}{17}$
  - $centroid_{\{KLN\}} = SUM/N = \binom{10.67}{2.33}$
  - $\sigma_x^2 = 342/3 - (32/3)^2 \approx 114 - 113.78 = 0.22$
  - $\sigma_x = \sqrt{0.22} \approx 0.47$
  - $\sigma_y^2 = 17/3 - (7/3)^2 \approx 5.67 - 5.44 \approx 0.23$
  - $\sigma_y = \sqrt{0.23} \approx 0.48$
- $M(10.5, 1.5)$
  - M in cluster $\{ABD\}$?
    $d(M, centroid_{\{ABD\}}) = \sqrt{(\frac{10.5-1.67}{0.46})^2 + (\frac{1.5-8.67}{0.47})^2} \approx \sqrt{(19.2)^2 + (-15.26)^2}$
    $\approx \sqrt{368.64 + 232.87} = \sqrt{601.51} \approx 24.53 > 2.5 \Rightarrow M$ is not part of this cluster.
  - M in cluster $\{KLN\}$?
    $d(M, centroid_{\{KLN\}}) = \sqrt{(\frac{10.5-10.67}{0.47})^2 + (\frac{1.5-2.33}{0.48})^2} \approx \sqrt{(-0.36)^2 + (-1.73)^2}$
    $\approx \sqrt{0.13 + 2.99} = \sqrt{3.12} \approx 1.77 < 2.5 \Rightarrow M$ is part of this cluster.

  We have new cluster $\{KLMN\}$ with $K(11,3)$, $L(10,2)$, $M(10.5,1.5)$ and $N(11,2)$, update N, SUM, $SUMSQ$ as well as variance and standard deviation σ for this cluster:

  - $N = 4$
  - $SUM = \binom{42.5}{8.5}$
  - $SUMSQ = \binom{452.25}{19.25}$

  - $centroid_{\{KLMN\}} = SUM/N = \binom{10.625}{2.125}$
  - $\sigma_x^2 = 452.25/4 - (42.5/4)^2 \approx 113.06 - 112.89 = 0.17$
  - $\sigma_x = \sqrt{0.17} \approx 0.41$
  - $\sigma_y^2 = 19.25/4 - (8.5/4)^2 \approx 4.8 - 4.52 = 0.28$
  - $\sigma_y = \sqrt{0.28} \approx 0.53$

We have read all points from the first chunk: B is a part of the $\{ABD\}$ cluster, N and M are in the $\{KLMN\}$ cluster. Points that are assigned to a cluster or a minicluster, i.e., those that are not in the retained set, are written out, with their assignment, to secondary memory. The points $F(3,3)$ and $I(8,9)$ can be clustered with **hierarchical clustering** approach (complete linkage), if they are close enough:

Note, that diameter of the DS clusters cannot be calculated without reading from the disc (what we want to avoid). We do not preserve points in the main memory, but the summary: $N, SUM, SUMSQ$. $centroid, \sigma^2$ and σ can be derived from the summary. Possible solution would be to use standard deviation for diameter calculation.

- $diameter_{\{ABD\}} = d(A,B) = \sqrt{2}$
- $diameter_{\{KLMN\}} = d(K,M) = \sqrt{2.5}$

$\Rightarrow average\ diameter \approx 1.5$

If we combine I and F to the cluster $\{IF\}$ (because they have the minimal distance), the diameter of this cluster will be $7.8 > 2 * 1.5 \Rightarrow$ these two points are not close enough and are in the RS.

The processing result of the first chunk you can the in the Figure 3.

Now we can read the points from the **second chunk**:

- $C(1.5, 9.5)$:
  - C in cluster $\{ABD\}$?
    $d(C, centroid_{\{ABD\}}) = \sqrt{(\frac{1.5-1.67}{0.46})^2 + (\frac{9.5-8.67}{0.47})^2} \approx \sqrt{(-0.37)^2 + (1.77)^2} \approx \sqrt{0.14 + 3.13} = \sqrt{3.27} \approx 1.81 < 2.5 \Rightarrow C$ can be part of this cluster.
  - C in cluster $\{KLMN\}$?
    $d(C, centroid_{\{KLMN\}}) = \sqrt{(\frac{1.5-10.625}{0.41})^2 + (\frac{9.5-2.125}{0.53})^2} \approx \sqrt{(-22.26)^2 + (13.92)^2} \approx \sqrt{495.51 + 193.77} = \sqrt{689.28} \approx 26.25 > 2.5 \Rightarrow C$ is not part of this cluster.

  We have new cluster $\{ABCD\}$ with $A(1,9)$, $B(2,8)$, $C(1.5, 9.5)$ and $D(2,9)$. Update N, SUM, $SUMSQ$ as well as variance σ^2 and standard deviation σ for this cluster:

  - $N = 4$
  - $SUM = \binom{6.5}{35.5}$
  - $SUMSQ = \binom{11.25}{316.25}$
  - $centroid_{\{ABCD\}} = \binom{1.625}{8.875}$
  - $\sigma_x^2 = 11.25/4 - (6.5/4)^2 \approx 2.8 - 2.64 = 0.16$
  - $\sigma_x = \sqrt{0.16} = 0.4$

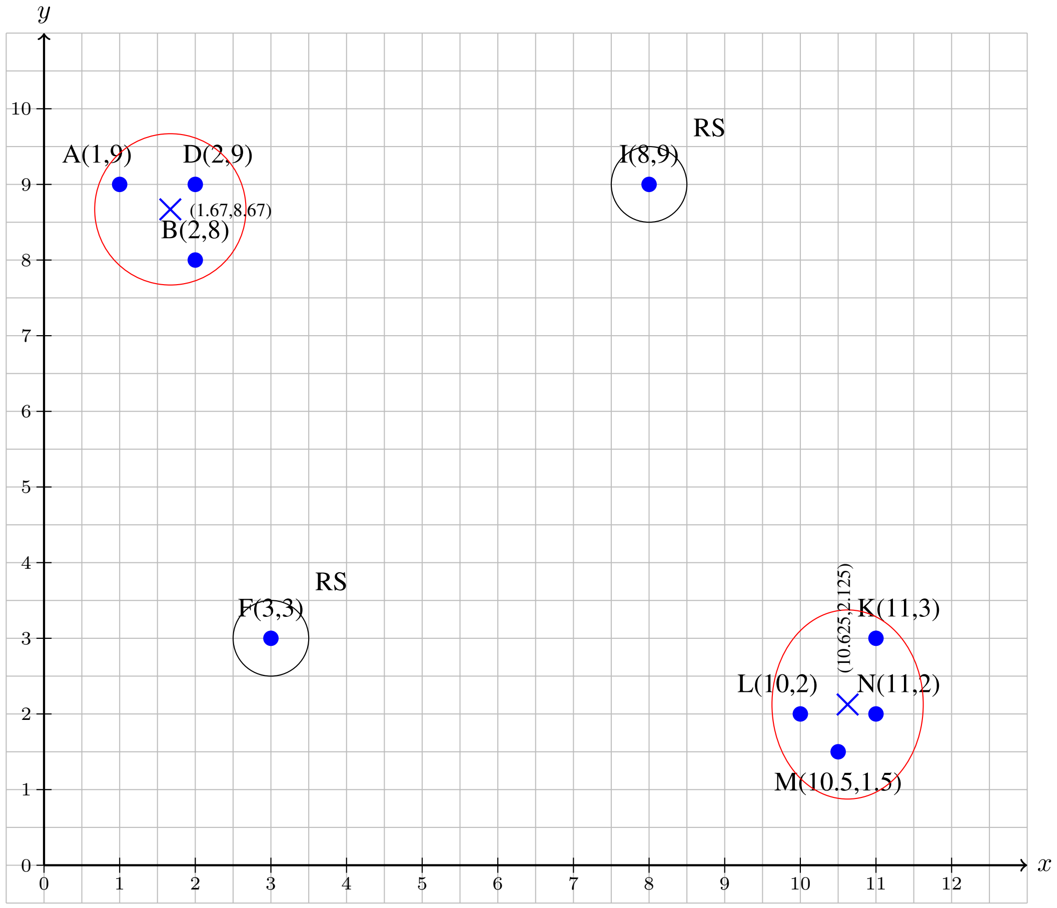

Abbildung 3: Result after chunk 1.

  - $\sigma_y^2 = 316.25/4 - (35.5/4)^2 \approx 79.1 - 78.77 = 0.33$
  - $\sigma_y = \sqrt{0.33} \approx 0.57$

- $G(4,5)$:
  - G in cluster $\{ABCD\}$?
    $d(G, centroid_{\{ABCD\}}) = \sqrt{(\frac{4-1.625}{0.4})^2 + (\frac{5-8.875}{0.57})^2} \approx \sqrt{(5.94)^2 + (-6.8)^2} \approx \sqrt{35.28 + 46.24} = \sqrt{81.52} \approx 9.03 > 2.5 \Rightarrow G$ is not part of this cluster.
  - G in cluster $\{KLMN\}$?
    $d(G, centroid_{\{KLMN\}}) = \sqrt{(\frac{4-10.625}{0.41})^2 + (\frac{5-2.125}{0.53})^2} \approx \sqrt{(-16.16)^2 + (5.42)^2}$
    $\approx \sqrt{261.15 + 29.38} = \sqrt{290.53} \approx 17.04 > 2.5 \Rightarrow G$ is not part of this cluster.

  $G \notin$ any DS (cluster).

- $H(5,5)$:
  - H in cluster $\{ABCD\}$?
    $d(G, centroid_{\{ABCD\}}) = \sqrt{(\frac{5-1.625}{0.4})^2 + (\frac{5-8.875}{0.57})^2} = \sqrt{(8.44)^2 + (-6.8)^2} \approx \sqrt{71.23 + 46.24} = \sqrt{117.47} \approx 10.84 > 2.5 \Rightarrow H$ is not part of this cluster.

  - H in cluster $\{KLMN\}$?
    $d(G, centroid_{\{KLMN\}}) = \sqrt{(\frac{5-10.625}{0.41})^2 + (\frac{5-2.125}{0.53})^2} \approx \sqrt{(-13.72)^2 + (5.42)^2}$
    $\approx \sqrt{188.24 + 29.38} = \sqrt{217.62} \approx 14.75 > 2.5 \Rightarrow H$ is not part of this cluster.

  $H \notin$ any DS (cluster).

- $J(10, 3)$:
  - J in cluster $\{ABCD\}$?
    $d(J, centroid_{\{ABCD\}}) = \sqrt{(\frac{10-1.625}{0.4})^2 + (\frac{3-8.875}{0.57})^2} \approx \sqrt{(20.94)^2 + (-10.31)^2}$
    $\approx \sqrt{438.48 + 106.3} = \sqrt{544.78} \approx 23.34 > 2.5 \Rightarrow J$ is not part of this cluster.
  - J in cluster $\{KLMN\}$?
    $d(J, centroid_{\{KLMN\}}) = \sqrt{(\frac{10-10.625}{0.41})^2 + (\frac{3-2.125}{0.53})^2} \approx \sqrt{(-1.52)^2 + (1.65)^2} \approx \sqrt{2.3 + 2.7} =$
    $\sqrt{5} \approx 2.24 < 2.5 \Rightarrow J$ is part of this cluster.

  We have new cluster $\{JKLMN\}$ with $J(10, 3)$, $K(11.3)$, $L(10.2)$, $M(10.5, 1.5)$ and $N(11.2)$. Update N, SUM, $SUMSQ$ as well as variance σ^2 and standard deviation σ for this cluster:

  - $N = 5$
  - $SUM = \binom{52.5}{11.5}$
  - $SUMQ = \binom{552.25}{28.25}$
  - $centroid_{\{JKLMN\}} = \binom{10.5}{2.3}$
  - $\sigma_x^2 = 552.25/5 - (52.5/5)^2 = 110.45 - 110.25 = 0.2$
  - $\sigma_x = \sqrt{0.2} \approx 0.45$
  - $\sigma_y^2 = 28.25/5 - (11.5/5)^2 = 5.65 - 5.29 = 0.36$
  - $\sigma_y = \sqrt{0.36} = 0.6$

- $E(3, 2)$:
  - E in cluster $\{ABCD\}$?
    $d(E, centroid_{\{ABCD\}}) = \sqrt{(\frac{3-1.625}{0.4})^2 + (\frac{2-8.875}{0.57})^2} \approx \sqrt{(3.44)^2 + (-12.06)^2}$
    $\approx \sqrt{11.83 + 145.44} = \sqrt{157.27} \approx 12.54 > 2.5 \Rightarrow E$ is not part of this cluster.
  - E in cluster $\{JKLMN\}$?
    $d(E, centroid_{\{JKLMN\}}) = \sqrt{(\frac{3-10.5}{0.45})^2 + (\frac{2-2.3}{0.6})^2} = \sqrt{(-16.67)^2 + (-0.5)^2} = \sqrt{277.89 + 0.25} =$
    $\sqrt{278.14} \approx 16.68 > 2.5 \Rightarrow E$ is not part of this cluster.

  $E \notin$ any DS (cluster).

We have read all points from the second chunk: C is a part of the $\{ABCD\}$ cluster, J is in the $\{JKLMN\}$ cluster.

The points $G(4, 5)$, $H(5, 5)$, $E(3, 2)$ as well as the points from retained sets (RS) - $F(3, 3)$ and $I(8, 9)$ can be clustered with **hierarchical clustering** approach (complete linkage), if they are close enough:

Note, that diameter of the DS clusters cannot be calculated without reading from the disc (what we want to avoid). We do not preserve points in the main memory, but the summary: $N, SUM, SUMSQ$. $centroid$, σ^2 and σ can be derived from the summary. Possible solution would be to use standard deviation for diameter calculation.

- $diameter_{\{ABCD\}} = d(C, B) = \sqrt{2.5}$
- $diameter_{\{JKLMN\}} = d(K, M) = \sqrt{2.5}$

$\Rightarrow average\ diameter \approx 1.58$

$min\ D = D(\{E\}, \{F\}) = d(E, F) = D(\{G\}, \{H\}) = d(G, H) = 1.$
$1 < 2 * 1.58 \Rightarrow$ these pairs of points are close enough and we can combine $E(3, 2)$ and $F(3, 3)$ to the clusters.

$min\ D = D(\{EF\}, \{GH\}) = d(E, H) = \sqrt{13} \approx 3.61.$
$3.61 > 2 * 1.58 \Rightarrow$ these two clusters can not be combined to the one bigger cluster and we have as result of hierarchical clustering two compressed sets (CS) $\{EF\}$, $\{GH\}$ and one retained set (RS) $I(8, 9)$.

Finally, if this is the last chunk of input data, we need to do something with the compressed and retained sets. We can treat them as outliers, and never cluster them at all. Or, we can assign each point in the retained set to the cluster of the nearest centroid. We can combine each minicluster with the cluster whose centroid is closest to the centroid of the minicluster.

We decide to treat compressed and retained sets as outliers. The result of clustering can be found in the Figure 4

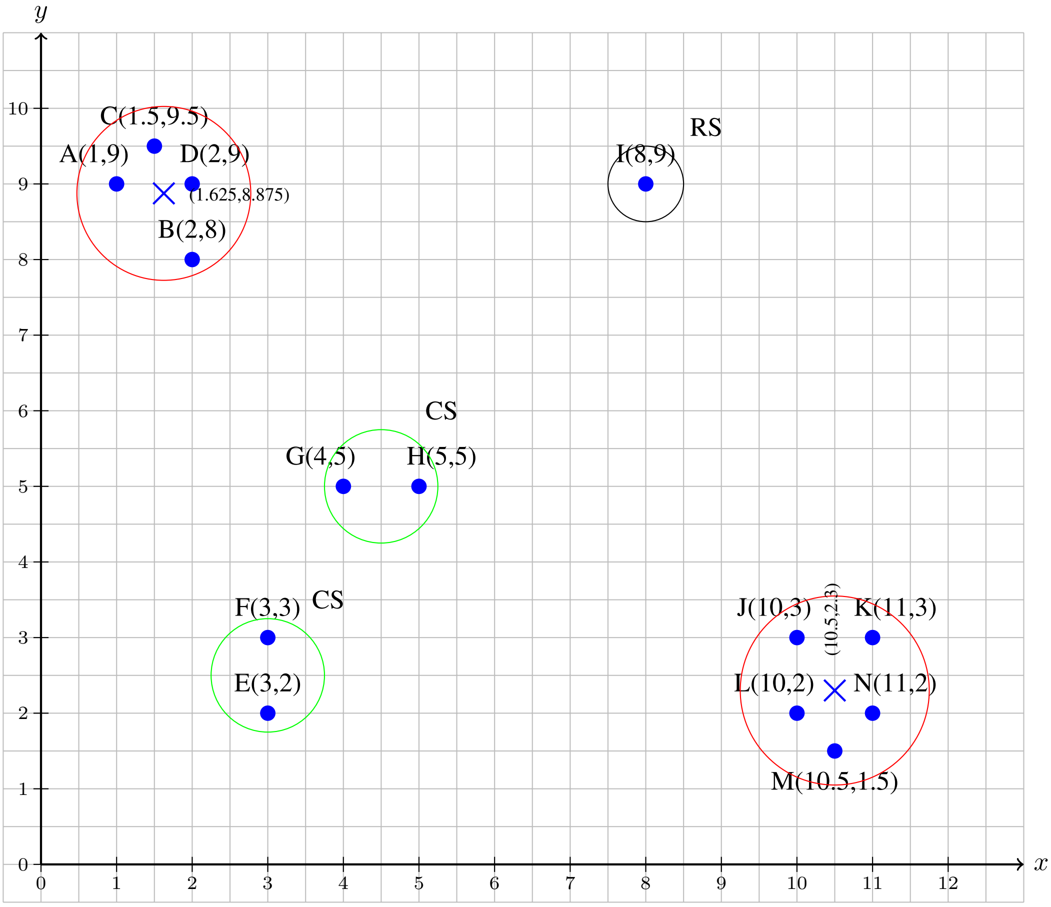

Abbildung 4: Result after chunk 2.