Friedrich-Alexander-Universität
Technische Fakultät

Department of Computer Science
Computer Science 6
(Data Management)

Knowledge Discovery in Databases with Exercises
Summer Semester 2024

Exercise Sheet 5: Clustering

About this Exercise Sheet

This exercise sheet focuses on the content of lecture *8. Clustering*.

It includes both theoretical exercises on K-means (Exercise 1) and DBSCAN (Exercise 2) and a practical data science exercise (Exercise 3).

The exercise sheet is designed for a two-week period, during which the tasks can be completed flexibly.

The sample solution will be published after the two weeks have elapsed.

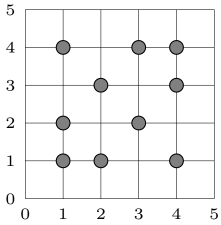
Preparation

Before participating in the exercise, you must prepare the following:

1. **Install Python and pip on your computer**
   - Detailed instructions can be found in `1-Introduction-Python-Pandas.pdf`.
2. **Download provided additional files**
   - Download `Additional-Files-Student.zip` from StudOn
   - Extract it to a folder of your choice.
3. **Install required Python packages**
   - Open a terminal and navigate to the folder where you extracted the files.
   - Run the command `pip install -r requirements.txt` within the extracted additional files folder to install the required Python packages.

Exercise 1: K-means

Given is a set of points in a two-dimensional space:



Points:

● (1,1), (1,2), (1,4), (2,1), (2,3), (3,2), (3,4), (4,1), (4,3), (4,4)

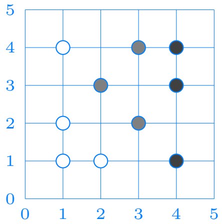
Use **K-means** to cluster the given points into three clusters. Use the **Euclidean distance** as the metric defining the similarity between points.

Write down **all** intermediate steps.

1. **Create an initial partitioning of the points**

   There are many ways to create an initial partitioning. All of them are valid as long as there are three non-empty partitions.

   In this sample solution, we assign the points to the partitions from left to right with roughly equal partition sizes:



Points:

○ (1,1), (1,2), (1,4), (2,1)

● (2,3), (3,2), (3,4)

● (4,1), (4,3), (4,4)

2. **Calculate the centroids of the partitions**

   - **White:**

$$\left.\begin{aligned} x &= \tfrac{1+1+1+2}{4} = 1.25 \\ y &= \tfrac{1+2+4+1}{4} = 2 \end{aligned}\right\} \Rightarrow (1.25, 2)$$

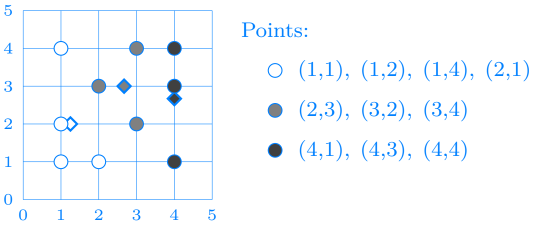
   - **Gray:**

$$\left.\begin{aligned} x &= \tfrac{2+3+3}{3} \approx 2.67 \\ y &= \tfrac{3+2+4}{3} = 3 \end{aligned}\right\} \Rightarrow (2.67, 3)$$

   - **Darkgray:**

$$\left.\begin{aligned} x &= \tfrac{4+4+4}{3} = 4 \\ y &= \tfrac{1+3+4}{3} \approx 2.67 \end{aligned}\right\} \Rightarrow (4, 2.67)$$

The centroids in the coordinate system:



3. **Reassign the points to the partitions based on the new centroids**

For each point, calculate the distance to each centroid and assign the point to the partition with the closest centroid:

- **Point** $(1, 1)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(1,1)\leftrightarrow(1.25,2)} &= \sqrt{(1-1.25)^2 + (1-2.0)^2} \approx 1.03 \\
Distance_{(1,1)\leftrightarrow(2.67,3)} &= \sqrt{(1-2.67)^2 + (1-3.0)^2} \approx 2.6 \\
Distance_{(1,1)\leftrightarrow(4,2.67)} &= \sqrt{(1-4.0)^2 + (1-2.67)^2} \approx 3.43
\end{aligned}$$

  The point $(1, 1)$ is closest to the centroid $(1.25, 2)$.

- **Point** $(1, 2)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(1,2)\leftrightarrow(1.25,2)} &= \sqrt{(1-1.25)^2 + (2-2.0)^2} \approx 0.25 \\
Distance_{(1,2)\leftrightarrow(2.67,3)} &= \sqrt{(1-2.67)^2 + (2-3.0)^2} \approx 1.94 \\
Distance_{(1,2)\leftrightarrow(4,2.67)} &= \sqrt{(1-4.0)^2 + (2-2.67)^2} \approx 3.07
\end{aligned}$$

  The point $(1, 2)$ is closest to the centroid $(1.25, 2)$.

- **Point** $(1, 4)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(1,4)\leftrightarrow(1.25,2)} &= \sqrt{(1-1.25)^2 + (4-2.0)^2} \approx 2.02 \\
Distance_{(1,4)\leftrightarrow(2.67,3)} &= \sqrt{(1-2.67)^2 + (4-3.0)^2} \approx 1.94 \\
Distance_{(1,4)\leftrightarrow(4,2.67)} &= \sqrt{(1-4.0)^2 + (4-2.67)^2} \approx 3.28
\end{aligned}$$

  The point $(1, 4)$ is closest to the centroid $(2.67, 3)$.

- **Point** $(2,1)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(2,1)\leftrightarrow(1.25,2)} &= \sqrt{(2-1.25)^2+(1-2.0)^2} \approx 1.25\\
Distance_{(2,1)\leftrightarrow(2.67,3)} &= \sqrt{(2-2.67)^2+(1-3.0)^2} \approx 2.11\\
Distance_{(2,1)\leftrightarrow(4,2.67)} &= \sqrt{(2-4.0)^2+(1-2.67)^2} \approx 2.6
\end{aligned}
$$

  The point $(2,1)$ is closest to the centroid $(1.25,2)$.

- **Point** $(2,3)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(2,3)\leftrightarrow(1.25,2)} &= \sqrt{(2-1.25)^2+(3-2.0)^2} \approx 1.25\\
Distance_{(2,3)\leftrightarrow(2.67,3)} &= \sqrt{(2-2.67)^2+(3-3.0)^2} \approx 0.67\\
Distance_{(2,3)\leftrightarrow(4,2.67)} &= \sqrt{(2-4.0)^2+(3-2.67)^2} \approx 2.03
\end{aligned}
$$

  The point $(2,3)$ is closest to the centroid $(2.67,3)$.

- **Point** $(3,2)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(3,2)\leftrightarrow(1.25,2)} &= \sqrt{(3-1.25)^2+(2-2.0)^2} \approx 1.75\\
Distance_{(3,2)\leftrightarrow(2.67,3)} &= \sqrt{(3-2.67)^2+(2-3.0)^2} \approx 1.05\\
Distance_{(3,2)\leftrightarrow(4,2.67)} &= \sqrt{(3-4.0)^2+(2-2.67)^2} \approx 1.2
\end{aligned}
$$

  The point $(3,2)$ is closest to the centroid $(2.67,3)$.

- **Point** $(3,4)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(3,4)\leftrightarrow(1.25,2)} &= \sqrt{(3-1.25)^2+(4-2.0)^2} \approx 2.66\\
Distance_{(3,4)\leftrightarrow(2.67,3)} &= \sqrt{(3-2.67)^2+(4-3.0)^2} \approx 1.05\\
Distance_{(3,4)\leftrightarrow(4,2.67)} &= \sqrt{(3-4.0)^2+(4-2.67)^2} \approx 1.67
\end{aligned}
$$

  The point $(3,4)$ is closest to the centroid $(2.67,3)$.

- **Point** $(4, 1)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(4,1)\leftrightarrow(1.25,2)} &= \sqrt{(4-1.25)^2+(1-2.0)^2} \approx 2.93 \\
Distance_{(4,1)\leftrightarrow(2.67,3)} &= \sqrt{(4-2.67)^2+(1-3.0)^2} \approx 2.4 \\
Distance_{(4,1)\leftrightarrow(4,2.67)} &= \sqrt{(4-4.0)^2+(1-2.67)^2} \approx 1.67
\end{aligned}$$

  The point $(4, 1)$ is closest to the centroid $(4, 2.67)$.

- **Point** $(4, 3)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(4,3)\leftrightarrow(1.25,2)} &= \sqrt{(4-1.25)^2+(3-2.0)^2} \approx 2.93 \\
Distance_{(4,3)\leftrightarrow(2.67,3)} &= \sqrt{(4-2.67)^2+(3-3.0)^2} \approx 1.33 \\
Distance_{(4,3)\leftrightarrow(4,2.67)} &= \sqrt{(4-4.0)^2+(3-2.67)^2} \approx 0.33
\end{aligned}$$

  The point $(4, 3)$ is closest to the centroid $(4, 2.67)$.

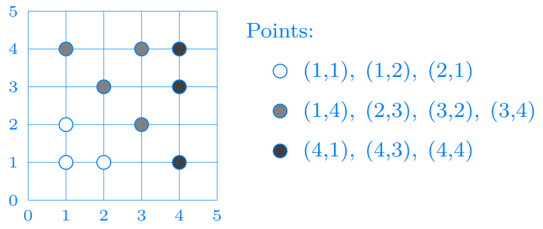
- **Point** $(4, 4)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(4,4)\leftrightarrow(1.25,2)} &= \sqrt{(4-1.25)^2+(4-2.0)^2} \approx 3.4 \\
Distance_{(4,4)\leftrightarrow(2.67,3)} &= \sqrt{(4-2.67)^2+(4-3.0)^2} \approx 1.67 \\
Distance_{(4,4)\leftrightarrow(4,2.67)} &= \sqrt{(4-4.0)^2+(4-2.67)^2} \approx 1.33
\end{aligned}$$

  The point $(4, 4)$ is closest to the centroid $(4, 2.67)$.

Overall, the new partitioning is:



Since the partitioning did change, the algorithm has not converged.

4. **Compute the new centroids**

   - **White:**

$$\left.\begin{aligned} x &= \tfrac{1+1+2}{3} \approx 1.33 \\ y &= \tfrac{1+2+1}{3} \approx 1.33 \end{aligned}\right\} \Rightarrow (1.33, 1.33)$$

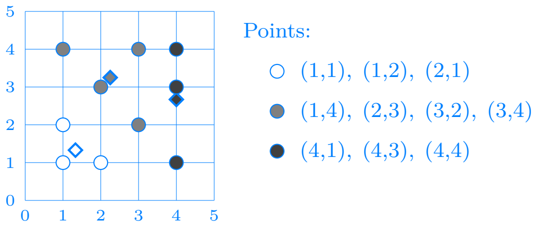
   - **Gray:**

$$\left.\begin{aligned} x &= \tfrac{1+2+3+3}{4} = 2.25 \\ y &= \tfrac{4+3+2+4}{4} = 3.25 \end{aligned}\right\} \Rightarrow (2.25, 3.25)$$

   - **Darkgray:**

$$\left.\begin{aligned} x &= \tfrac{4+4+4}{3} = 4 \\ y &= \tfrac{1+3+4}{3} \approx 2.67 \end{aligned}\right\} \Rightarrow (4, 2.67)$$

The centroids in the coordinate system:



5. **Reassign the points to the partitions based on the new centroids**

For each point, calculate the distance to each centroid and assign the point to the partition with the closest centroid:

   - **Point** $(1, 1)$**:**

     Calculate the distances:

$$\begin{aligned} Distance_{(1,1)\leftrightarrow(1.33,1.33)} &= \sqrt{(1-1.33)^2 + (1-1.33)^2} \approx 0.47 \\ Distance_{(1,1)\leftrightarrow(2.25,3.25)} &= \sqrt{(1-2.25)^2 + (1-3.25)^2} \approx 2.6 \\ Distance_{(1,1)\leftrightarrow(4,2.67)} &= \sqrt{(1-4.0)^2 + (1-2.67)^2} \approx 3.43 \end{aligned}$$

     The point $(1, 1)$ is closest to the centroid $(1.33, 1.33)$.

- **Point** $(1, 2)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(1,2)\leftrightarrow(1.33,1.33)} &= \sqrt{(1-1.33)^2+(2-1.33)^2} \approx 0.75\\
Distance_{(1,2)\leftrightarrow(2.25,3.25)} &= \sqrt{(1-2.25)^2+(2-3.25)^2} \approx 1.77\\
Distance_{(1,2)\leftrightarrow(4,2.67)} &= \sqrt{(1-4.0)^2+(2-2.67)^2} \approx 3.07
\end{aligned}
$$

  The point $(1, 2)$ is closest to the centroid $(1.33, 1.33)$.

- **Point** $(1, 4)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(1,4)\leftrightarrow(1.33,1.33)} &= \sqrt{(1-1.33)^2+(4-1.33)^2} \approx 2.69\\
Distance_{(1,4)\leftrightarrow(2.25,3.25)} &= \sqrt{(1-2.25)^2+(4-3.25)^2} \approx 1.46\\
Distance_{(1,4)\leftrightarrow(4,2.67)} &= \sqrt{(1-4.0)^2+(4-2.67)^2} \approx 3.28
\end{aligned}
$$

  The point $(1, 4)$ is closest to the centroid $(2.25, 3.25)$.

- **Point** $(2, 1)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(2,1)\leftrightarrow(1.33,1.33)} &= \sqrt{(2-1.33)^2+(1-1.33)^2} \approx 0.75\\
Distance_{(2,1)\leftrightarrow(2.25,3.25)} &= \sqrt{(2-2.25)^2+(1-3.25)^2} \approx 2.26\\
Distance_{(2,1)\leftrightarrow(4,2.67)} &= \sqrt{(2-4.0)^2+(1-2.67)^2} \approx 2.6
\end{aligned}
$$

  The point $(2, 1)$ is closest to the centroid $(1.33, 1.33)$.

- **Point** $(2, 3)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(2,3)\leftrightarrow(1.33,1.33)} &= \sqrt{(2-1.33)^2+(3-1.33)^2} \approx 1.8\\
Distance_{(2,3)\leftrightarrow(2.25,3.25)} &= \sqrt{(2-2.25)^2+(3-3.25)^2} \approx 0.35\\
Distance_{(2,3)\leftrightarrow(4,2.67)} &= \sqrt{(2-4.0)^2+(3-2.67)^2} \approx 2.03
\end{aligned}
$$

  The point $(2, 3)$ is closest to the centroid $(2.25, 3.25)$.

- **Point** $(3, 2)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(3,2)\leftrightarrow(1.33,1.33)} &= \sqrt{(3-1.33)^2+(2-1.33)^2} \approx 1.8 \\
Distance_{(3,2)\leftrightarrow(2.25,3.25)} &= \sqrt{(3-2.25)^2+(2-3.25)^2} \approx 1.46 \\
Distance_{(3,2)\leftrightarrow(4,2.67)} &= \sqrt{(3-4.0)^2+(2-2.67)^2} \approx 1.2
\end{aligned}
$$

  The point $(3, 2)$ is closest to the centroid $(4.0, 2.67)$.

- **Point** $(3, 4)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(3,4)\leftrightarrow(1.33,1.33)} &= \sqrt{(3-1.33)^2+(4-1.33)^2} \approx 3.14 \\
Distance_{(3,4)\leftrightarrow(2.25,3.25)} &= \sqrt{(3-2.25)^2+(4-3.25)^2} \approx 1.06 \\
Distance_{(3,4)\leftrightarrow(4,2.67)} &= \sqrt{(3-4.0)^2+(4-2.67)^2} \approx 1.67
\end{aligned}
$$

  The point $(3, 4)$ is closest to the centroid $(2.25, 3.25)$.

- **Point** $(4, 1)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(4,1)\leftrightarrow(1.33,1.33)} &= \sqrt{(4-1.33)^2+(1-1.33)^2} \approx 2.69 \\
Distance_{(4,1)\leftrightarrow(2.25,3.25)} &= \sqrt{(4-2.25)^2+(1-3.25)^2} \approx 2.85 \\
Distance_{(4,1)\leftrightarrow(4,2.67)} &= \sqrt{(4-4.0)^2+(1-2.67)^2} \approx 1.67
\end{aligned}
$$

  The point $(4, 1)$ is closest to the centroid $(4.0, 2.67)$.

- **Point** $(4, 3)$**:**

  Calculate the distances:

$$
\begin{aligned}
Distance_{(4,3)\leftrightarrow(1.33,1.33)} &= \sqrt{(4-1.33)^2+(3-1.33)^2} \approx 3.14 \\
Distance_{(4,3)\leftrightarrow(2.25,3.25)} &= \sqrt{(4-2.25)^2+(3-3.25)^2} \approx 1.77 \\
Distance_{(4,3)\leftrightarrow(4,2.67)} &= \sqrt{(4-4.0)^2+(3-2.67)^2} \approx 0.33
\end{aligned}
$$

  The point $(4, 3)$ is closest to the centroid $(4.0, 2.67)$.

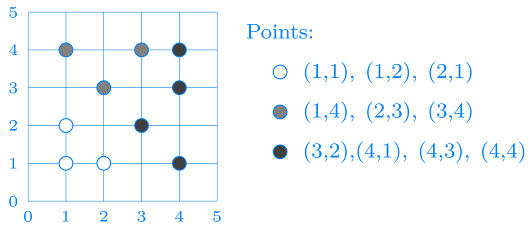
- **Point** $(4,4)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(4,4)\leftrightarrow(1.33,1.33)} &= \sqrt{(4-1.33)^2+(4-1.33)^2} \approx 3.77\\
Distance_{(4,4)\leftrightarrow(2.25,3.25)} &= \sqrt{(4-2.25)^2+(4-3.25)^2} \approx 1.9\\
Distance_{(4,4)\leftrightarrow(4,2.67)} &= \sqrt{(4-4.0)^2+(4-2.67)^2} \approx 1.33
\end{aligned}$$

  The point $(4,4)$ is closest to the centroid $(4.0, 2.67)$.

Overall, the new partitioning is:



Since the partitioning did change, the algorithm has not converged.

6. **Compute the new centroids**

- **White:**

$$\left.\begin{aligned} x &= \tfrac{1+1+2}{3} \approx 1.33\\ y &= \tfrac{1+2+1}{3} \approx 1.33\end{aligned}\right\} \Rightarrow (1.33, 1.33)$$

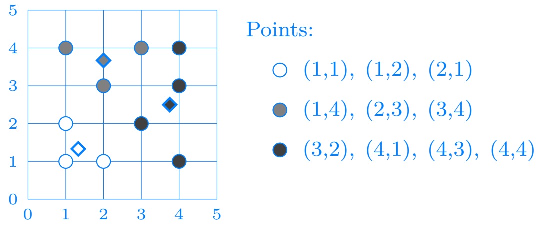
- **Gray:**

$$\left.\begin{aligned} x &= \tfrac{1+2+3}{3} = 2\\ y &= \tfrac{4+3+4}{3} \approx 3.67\end{aligned}\right\} \Rightarrow (2, 3.67)$$

- **Darkgray:**

$$\left.\begin{aligned} x &= \tfrac{3+4+4+4}{4} = 3.75\\ y &= \tfrac{2+1+3+4}{4} = 2.5\end{aligned}\right\} \Rightarrow (3.75, 2.5)$$

The centroids in the coordinate system:



. . . .

7. **Reassign the points to the partitions based on the new centroids**

   For each point, calculate the distance to each centroid and assign the point to the partition with the closest centroid:

   - **Point** $(1,1)$**:**

     Calculate the distances:

     $$Distance_{(1,1)\leftrightarrow(1.33,1.33)} = \sqrt{(1-1.33)^2 + (1-1.33)^2} \approx 0.47$$
     $$Distance_{(1,1)\leftrightarrow(2.0,3.67)} = \sqrt{(1-2.0)^2 + (1-3.67)^2} \approx 2.85$$
     $$Distance_{(1,1)\leftrightarrow(3.75,2.5)} = \sqrt{(1-3.75)^2 + (1-2.5)^2} \approx 3.13$$

     The point $(1,1)$ is closest to the centroid $(1.33,1.33)$.

   - **Point** $(1,2)$**:**

     Calculate the distances:

     $$Distance_{(1,2)\leftrightarrow(1.33,1.33)} = \sqrt{(1-1.33)^2 + (2-1.33)^2} \approx 0.75$$
     $$Distance_{(1,2)\leftrightarrow(2.0,3.67)} = \sqrt{(1-2.0)^2 + (2-3.67)^2} \approx 1.94$$
     $$Distance_{(1,2)\leftrightarrow(3.75,2.5)} = \sqrt{(1-3.75)^2 + (2-2.5)^2} \approx 2.8$$

     The point $(1,2)$ is closest to the centroid $(1.33,1.33)$.

   - **Point** $(1,4)$**:**

     Calculate the distances:

     $$Distance_{(1,4)\leftrightarrow(1.33,1.33)} = \sqrt{(1-1.33)^2 + (4-1.33)^2} \approx 2.69$$
     $$Distance_{(1,4)\leftrightarrow(2.0,3.67)} = \sqrt{(1-2.0)^2 + (4-3.67)^2} \approx 1.05$$
     $$Distance_{(1,4)\leftrightarrow(3.75,2.5)} = \sqrt{(1-3.75)^2 + (4-2.5)^2} \approx 3.13$$

     The point $(1,4)$ is closest to the centroid $(2.0,3.67)$.

   - **Point** $(2,1)$**:**

     Calculate the distances:

$$Distance_{(2,1)\leftrightarrow(1.33,1.33)} = \sqrt{(2-1.33)^2+(1-1.33)^2} \approx 0.75$$
$$Distance_{(2,1)\leftrightarrow(2.0,3.67)} = \sqrt{(2-2.0)^2+(1-3.67)^2} \approx 2.67$$
$$Distance_{(2,1)\leftrightarrow(3.75,2.5)} = \sqrt{(2-3.75)^2+(1-2.5)^2} \approx 2.3$$

The point $(2, 1)$ is closest to the centroid $(1.33, 1.33)$.

- **Point** $(2, 3)$**:**

  Calculate the distances:

$$Distance_{(2,3)\leftrightarrow(1.33,1.33)} = \sqrt{(2-1.33)^2+(3-1.33)^2} \approx 1.8$$
$$Distance_{(2,3)\leftrightarrow(2.0,3.67)} = \sqrt{(2-2.0)^2+(3-3.67)^2} \approx 0.67$$
$$Distance_{(2,3)\leftrightarrow(3.75,2.5)} = \sqrt{(2-3.75)^2+(3-2.5)^2} \approx 1.82$$

  The point $(2, 3)$ is closest to the centroid $(2.0, 3.67)$.

- **Point** $(3, 2)$**:**

  Calculate the distances:

$$Distance_{(3,2)\leftrightarrow(1.33,1.33)} = \sqrt{(3-1.33)^2+(2-1.33)^2} \approx 1.8$$
$$Distance_{(3,2)\leftrightarrow(2.0,3.67)} = \sqrt{(3-2.0)^2+(2-3.67)^2} \approx 1.94$$
$$Distance_{(3,2)\leftrightarrow(3.75,2.5)} = \sqrt{(3-3.75)^2+(2-2.5)^2} \approx 0.9$$

  The point $(3, 2)$ is closest to the centroid $(3.75, 2.5)$.

- **Point** $(3, 4)$**:**

  Calculate the distances:

$$Distance_{(3,4)\leftrightarrow(1.33,1.33)} = \sqrt{(3-1.33)^2+(4-1.33)^2} \approx 3.14$$
$$Distance_{(3,4)\leftrightarrow(2.0,3.67)} = \sqrt{(3-2.0)^2+(4-3.67)^2} \approx 1.05$$
$$Distance_{(3,4)\leftrightarrow(3.75,2.5)} = \sqrt{(3-3.75)^2+(4-2.5)^2} \approx 1.68$$

  The point $(3, 4)$ is closest to the centroid $(2.0, 3.67)$.

- **Point** $(4, 1)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(4,1)\leftrightarrow(1.33,1.33)} &= \sqrt{(4-1.33)^2+(1-1.33)^2} \approx 2.69 \\
Distance_{(4,1)\leftrightarrow(2.0,3.67)} &= \sqrt{(4-2.0)^2+(1-3.67)^2} \approx 3.33 \\
Distance_{(4,1)\leftrightarrow(3.75,2.5)} &= \sqrt{(4-3.75)^2+(1-2.5)^2} \approx 1.52
\end{aligned}$$

  The point $(4, 1)$ is closest to the centroid $(3.75, 2.5)$.

- **Point** $(4, 3)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(4,3)\leftrightarrow(1.33,1.33)} &= \sqrt{(4-1.33)^2+(3-1.33)^2} \approx 3.14 \\
Distance_{(4,3)\leftrightarrow(2.0,3.67)} &= \sqrt{(4-2.0)^2+(3-3.67)^2} \approx 2.11 \\
Distance_{(4,3)\leftrightarrow(3.75,2.5)} &= \sqrt{(4-3.75)^2+(3-2.5)^2} \approx 0.56
\end{aligned}$$

  The point $(4, 3)$ is closest to the centroid $(3.75, 2.5)$.

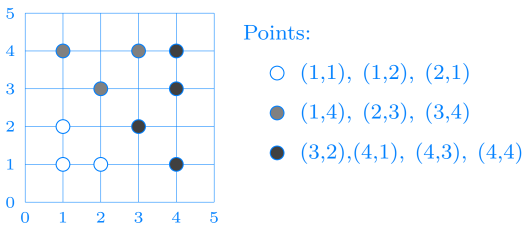
- **Point** $(4, 4)$**:**

  Calculate the distances:

$$\begin{aligned}
Distance_{(4,4)\leftrightarrow(1.33,1.33)} &= \sqrt{(4-1.33)^2+(4-1.33)^2} \approx 3.77 \\
Distance_{(4,4)\leftrightarrow(2.0,3.67)} &= \sqrt{(4-2.0)^2+(4-3.67)^2} \approx 2.03 \\
Distance_{(4,4)\leftrightarrow(3.75,2.5)} &= \sqrt{(4-3.75)^2+(4-2.5)^2} \approx 1.52
\end{aligned}$$

  The point $(4, 4)$ is closest to the centroid $(3.75, 2.5)$.

Overall, the partitioning still is:

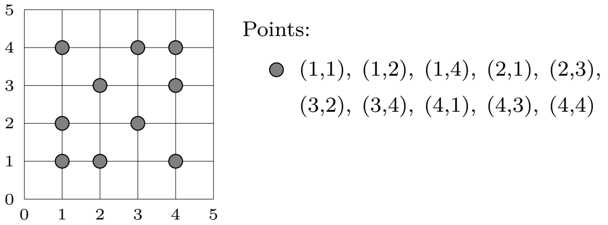


Since the partitioning did not change, the algorithm has converged.

Exercise 2: DBSCAN

Task 1: Basic Terms

Given is a set of points in a two-dimensional space:



Task 1.1: Core Points

Determine whether $(1,1)$, $(2,1)$, $(2,3)$, and $(1,4)$ are **core points** if a density based clustering algorithm like **DBSCAN** is initialized with $\varepsilon = 1$ and $MinPts = 2$ and applied on the given point set. The distance is calculated using the Euclidean distance.

To determine the core points, we need to calculate the distance of each point to all other points. If the distance is less than or equal to $\varepsilon = 1$, the point is considered a core point. The number of points within the ε-neighborhood of a point must be greater than or equal to $MinPts = 2$.

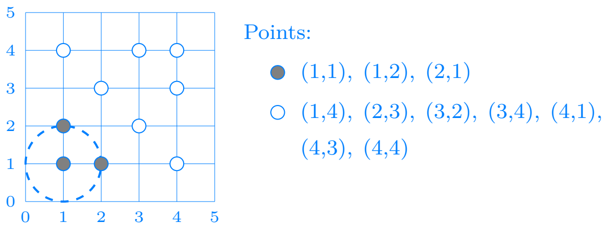
- **Point $(1,1)$:**

$$Distance_{(1,1)\leftrightarrow(1,1)} = \sqrt{(1-1)^2 + (1-1)^2} = 0$$
$$Distance_{(1,1)\leftrightarrow(1,2)} = \sqrt{(1-1)^2 + (1-2)^2} = 1$$
$$Distance_{(1,1)\leftrightarrow(1,4)} = \sqrt{(1-1)^2 + (1-4)^2} = 3$$
$$Distance_{(1,1)\leftrightarrow(2,1)} = \sqrt{(1-2)^2 + (1-1)^2} = 1$$
$$Distance_{(1,1)\leftrightarrow(2,3)} = \sqrt{(1-2)^2 + (1-3)^2} \approx 2.24$$
$$Distance_{(1,1)\leftrightarrow(3,2)} = \sqrt{(1-3)^2 + (1-2)^2} \approx 2.24$$
$$Distance_{(1,1)\leftrightarrow(3,4)} = \sqrt{(1-3)^2 + (1-4)^2} \approx 3.61$$
$$Distance_{(1,1)\leftrightarrow(4,1)} = \sqrt{(1-4)^2 + (1-1)^2} = 3$$
$$Distance_{(1,1)\leftrightarrow(4,3)} = \sqrt{(1-4)^2 + (1-3)^2} \approx 3.61$$
$$Distance_{(1,1)\leftrightarrow(4,4)} = \sqrt{(1-4)^2 + (1-4)^2} \approx 4.24$$

There are 3 points within the ε-neighborhood of point $(1,1)$, thus it is a core point.

. . . .

This can also be visualized in the coordinate system:



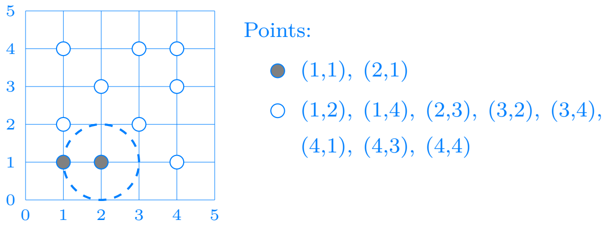
Please note, that the point itself is also part of the ε-neighborhood and thus is included in the count.

- **Point $(2,1)$:**

$$
\begin{aligned}
Distance_{(2,1)\leftrightarrow(1,1)} &= \sqrt{(2-1)^2+(1-1)^2} = 1 \\
Distance_{(2,1)\leftrightarrow(1,2)} &= \sqrt{(2-1)^2+(1-2)^2} \approx 1.41 \\
Distance_{(2,1)\leftrightarrow(1,4)} &= \sqrt{(2-1)^2+(1-4)^2} \approx 3.16 \\
Distance_{(2,1)\leftrightarrow(2,1)} &= \sqrt{(2-2)^2+(1-1)^2} = 0 \\
Distance_{(2,1)\leftrightarrow(2,3)} &= \sqrt{(2-2)^2+(1-3)^2} = 2 \\
Distance_{(2,1)\leftrightarrow(3,2)} &= \sqrt{(2-3)^2+(1-2)^2} \approx 1.41 \\
Distance_{(2,1)\leftrightarrow(3,4)} &= \sqrt{(2-3)^2+(1-4)^2} \approx 3.16 \\
Distance_{(2,1)\leftrightarrow(4,1)} &= \sqrt{(2-4)^2+(1-1)^2} = 2 \\
Distance_{(2,1)\leftrightarrow(4,3)} &= \sqrt{(2-4)^2+(1-3)^2} \approx 2.83 \\
Distance_{(2,1)\leftrightarrow(4,4)} &= \sqrt{(2-4)^2+(1-4)^2} \approx 3.61
\end{aligned}
$$

There are 2 points within the ε-neighborhood of point $(2,1)$, thus it is a core point.

This can also be visualized in the coordinate system:

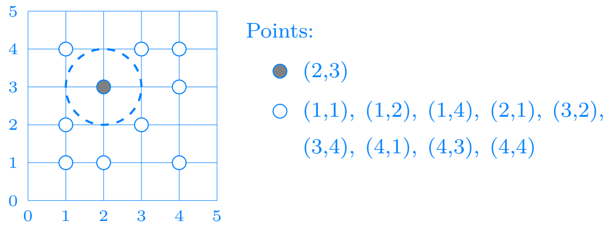


- **Point** $(2,3)$**:**

$$
\begin{aligned}
Distance_{(2,3)\leftrightarrow(1,1)} &= \sqrt{(2-1)^2+(3-1)^2} \approx 2.24\\
Distance_{(2,3)\leftrightarrow(1,2)} &= \sqrt{(2-1)^2+(3-2)^2} \approx 1.41\\
Distance_{(2,3)\leftrightarrow(1,4)} &= \sqrt{(2-1)^2+(3-4)^2} \approx 1.41\\
Distance_{(2,3)\leftrightarrow(2,1)} &= \sqrt{(2-2)^2+(3-1)^2} = 2\\
Distance_{(2,3)\leftrightarrow(2,3)} &= \sqrt{(2-2)^2+(3-3)^2} = 0\\
Distance_{(2,3)\leftrightarrow(3,2)} &= \sqrt{(2-3)^2+(3-2)^2} \approx 1.41\\
Distance_{(2,3)\leftrightarrow(3,4)} &= \sqrt{(2-3)^2+(3-4)^2} \approx 1.41\\
Distance_{(2,3)\leftrightarrow(4,1)} &= \sqrt{(2-4)^2+(3-1)^2} \approx 2.83\\
Distance_{(2,3)\leftrightarrow(4,3)} &= \sqrt{(2-4)^2+(3-3)^2} = 2\\
Distance_{(2,3)\leftrightarrow(4,4)} &= \sqrt{(2-4)^2+(3-4)^2} \approx 2.24
\end{aligned}
$$

There is only 1 point within the ε-neighborhood of point $(2,3)$, thus it is not a core point.

This can also be visualized in the coordinate system:

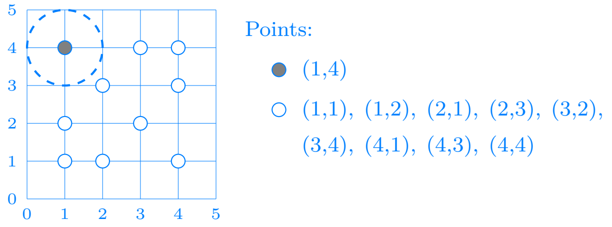


- **Point** $(1,4)$**:**

$$
\begin{aligned}
Distance_{(1,4)\leftrightarrow(1,1)} &= \sqrt{(1-1)^2+(4-1)^2} = 3\\
Distance_{(1,4)\leftrightarrow(1,2)} &= \sqrt{(1-1)^2+(4-2)^2} = 2\\
Distance_{(1,4)\leftrightarrow(1,4)} &= \sqrt{(1-1)^2+(4-4)^2} = 0\\
Distance_{(1,4)\leftrightarrow(2,1)} &= \sqrt{(1-2)^2+(4-1)^2} \approx 3.16\\
Distance_{(1,4)\leftrightarrow(2,3)} &= \sqrt{(1-2)^2+(4-3)^2} \approx 1.41\\
Distance_{(1,4)\leftrightarrow(3,2)} &= \sqrt{(1-3)^2+(4-2)^2} \approx 2.83\\
Distance_{(1,4)\leftrightarrow(3,4)} &= \sqrt{(1-3)^2+(4-4)^2} = 2\\
Distance_{(1,4)\leftrightarrow(4,1)} &= \sqrt{(1-4)^2+(4-1)^2} \approx 4.24\\
Distance_{(1,4)\leftrightarrow(4,3)} &= \sqrt{(1-4)^2+(4-3)^2} \approx 3.16\\
Distance_{(1,4)\leftrightarrow(4,4)} &= \sqrt{(1-4)^2+(4-4)^2} = 3
\end{aligned}
$$

There is only 1 point within the ε-neighborhood of point $(1,4)$, thus it is not a core point.

This can also be visualized in the coordinate system:



Task 1.2: Direct Density Reachability

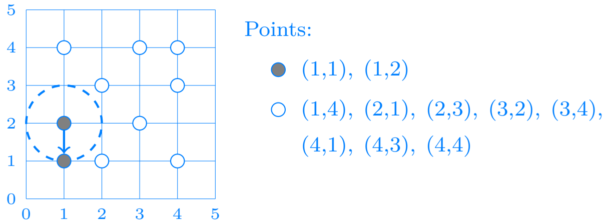
Determine which of the points in the point set are **directly density reachable** from the core point $(1,2)$ if a density based clustering algorithm like **DBSCAN** is initialized with $\varepsilon = 1$ and $MinPts = 2$. The distance is calculated using the Euclidean distance.

To determine which points are directly density reachable from a core point, we need to calculate the distance of each point to the core point. If the distance is less than or equal to $\varepsilon = 1$, the point is directly density reachable.

$$\begin{aligned}
Distance_{(1,2)\leftrightarrow(1,1)} &= \sqrt{(1-1)^2+(2-1)^2} = 1\\
Distance_{(1,2)\leftrightarrow(1,4)} &= \sqrt{(1-1)^2+(2-4)^2} = 2\\
Distance_{(1,2)\leftrightarrow(2,1)} &= \sqrt{(1-2)^2+(2-1)^2} \approx 1.41\\
Distance_{(1,2)\leftrightarrow(2,3)} &= \sqrt{(1-2)^2+(2-3)^2} \approx 1.41\\
Distance_{(1,2)\leftrightarrow(3,2)} &= \sqrt{(1-3)^2+(2-2)^2} = 2\\
Distance_{(1,2)\leftrightarrow(3,4)} &= \sqrt{(1-3)^2+(2-4)^2} \approx 2.83\\
Distance_{(1,2)\leftrightarrow(4,1)} &= \sqrt{(1-4)^2+(2-1)^2} \approx 3.16\\
Distance_{(1,2)\leftrightarrow(4,3)} &= \sqrt{(1-4)^2+(2-3)^2} \approx 3.16\\
Distance_{(1,2)\leftrightarrow(4,4)} &= \sqrt{(1-4)^2+(2-4)^2} \approx 3.60
\end{aligned}$$

Only the point $(1,1)$ is directly density reachable from the core point $(1,2)$.

This can also be visualized in the coordinate system:



Task 1.3: Density Reachability

Task 1.3.1: Basic Density Reachability

Determine whether $(1,1)$, $(2,1)$, $(2,3)$, and $(4,4)$ are **density reachable** from the core point $(1,2)$ if a density based clustering algorithm like **DBSCAN** is initialized with $\varepsilon = 1$ and $MinPts = 2$. The distance is calculated using the Euclidean distance.

A point is density reachable from a core point if there is a chain of direct density reachable points leading from the core point to the point. The distance between each pair of points in the chain must be less than or equal to $\varepsilon = 1$ and each point in the chain (with the final point being an exception) must be a core point.

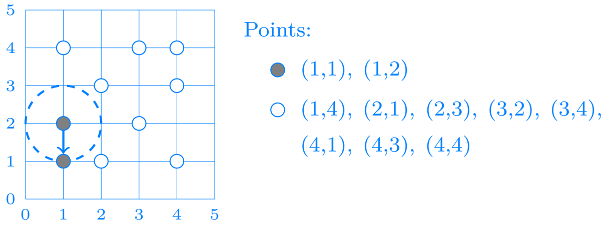
In theory we would have to check all possible chains of direct density reachable points. However, as a human it is possible to estimate which point chains might be possible just by looking at the points. In this sample solution, we will only check, whether a chain we think is possible is actually possible.

- **Point** $(1, 1)$**:**

  The point $(1, 1)$ is directly density reachable from the core point $(1, 2)$ (as shown in Task 1.2) Thus, it is also density reachable.

  ....

  This can also be visualized in the coordinate system:

  ....

- **Point** $(2, 1)$**:**

  The point $(2, 1)$ is not directly density reachable from the core point $(1, 2)$. However, it seems like there is a chain of points leading from the core point $(1, 2)$ to the point $(1, 1)$ and then to the point $(2, 1)$.

  To prove that this chain is possible, we have to prove the following:

  1. The point $(1, 2)$ is a core point:

  $$Distance_{(1,2)\leftrightarrow(1,1)} = \sqrt{(1-1)^2 + (2-1)^2} = 1$$
  $$Distance_{(1,2)\leftrightarrow(1,2)} = \sqrt{(1-1)^2 + (2-2)^2} = 0$$

  2. The point $(1, 1)$ is directly density reachable from the point $(1, 2)$:

  $$Distance_{(1,2)\leftrightarrow(1,1)} = \sqrt{(1-1)^2 + (2-1)^2} = 1$$

  3. The point $(1, 1)$ is a core point:

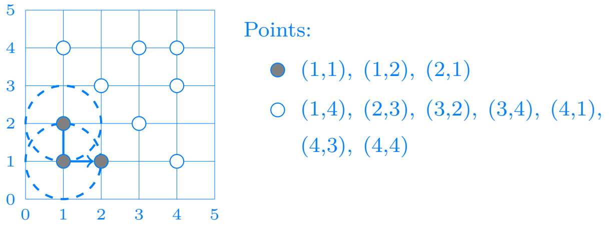
  $$Distance_{(1,1)\leftrightarrow(1,1)} = \sqrt{(1-1)^2 + (1-1)^2} = 0$$
  $$Distance_{(1,1)\leftrightarrow(1,2)} = \sqrt{(1-1)^2 + (1-2)^2} = 1$$
  $$Distance_{(1,1)\leftrightarrow(2,1)} = \sqrt{(1-2)^2 + (1-1)^2} = 1$$

  4. The point $(2, 1)$ is directly density reachable from the point $(1, 1)$:

  $$Distance_{(1,1)\leftrightarrow(2,1)} = \sqrt{(1-2)^2 + (1-1)^2} = 1$$

Thus, the point $(2,1)$ is density reachable from $(1,2)$.

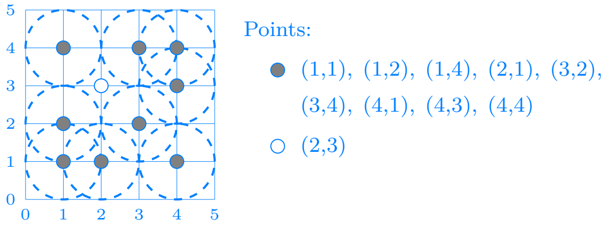
This can also be visualized in the coordinate system:



- **Point** $(2,3)$**:**

  The point $(2,3)$ is not within the ε-neighborhood of any other point. Thus, it is not density reachable from the core point $(1,2)$.

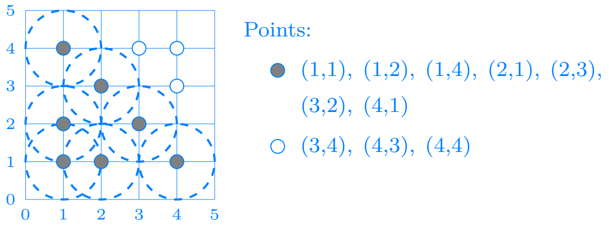
  This can also be visualized in the coordinate system:



- **Point** $(4,4)$**:**

  While the point $(4,4)$ seems to be density reachable from the points $(3,4)$ and $(4,3)$, neither of these points are in the ε-neighborhood of any other point. Thus, the point $(4,4)$ is not density reachable from the core point $(1,2)$.

  This can also be visualized in the coordinate system:



Task 1.3.2: Reversal of Density Reachability

Determine whether $(3,4)$ is density reachable from $(4,4)$ and whether $(4,4)$ is density reachable from $(3,4)$ if a density based clustering algorithm like **DBSCAN** is initialized with $\varepsilon = 1$ and $MinPts = 3$. The distance is calculated using the Euclidean distance.

Be careful: $MinPts$ was increased in this task. Thus you have to reevaluate whether points are core points or not.

Density reachability is directional, since all points in the chain **excluding** the final point must be core points.

- **$(3,4)$ from $(4,4)$:**

  To determine whether $(3,4)$ is density reachable from $(4,4)$, we have to check the following:

  1. The point $(4,4)$ is a core point:

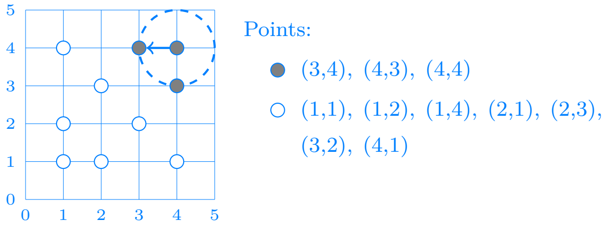
  $$\begin{aligned} Distance_{(4,4)\leftrightarrow(4,4)} &= \sqrt{(4-4)^2 + (4-4)^2} = 0 \\ Distance_{(4,4)\leftrightarrow(3,4)} &= \sqrt{(4-3)^2 + (4-4)^2} = 1 \\ Distance_{(4,4)\leftrightarrow(4,3)} &= \sqrt{(4-4)^2 + (4-3)^2} = 1 \end{aligned}$$

  2. The point $(3,4)$ is directly density reachable from the point $(4,4)$:

  $$Distance_{(4,4)\leftrightarrow(3,4)} = \sqrt{(4-3)^2 + (4-4)^2} = 1$$

  Thus, $(3,4)$ is density reachable from $(4,4)$.

  This can also be visualized in the coordinate system:



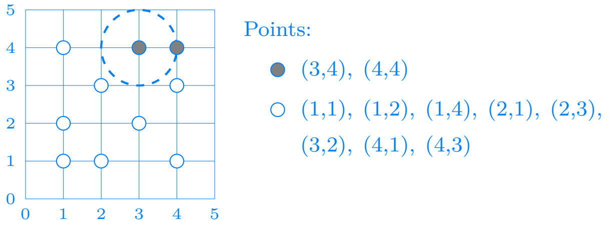
- **$(4,4)$ from $(3,4)$:**

  While checking the density reachability from $(3,4)$ to $(4,4)$, we have to first check whether $(3,4)$ is a core point:

$$Distance_{(3,4)\leftrightarrow(3,4)} = \sqrt{(3-3)^2 + (4-4)^2} = 0$$
$$Distance_{(3,4)\leftrightarrow(4,4)} = \sqrt{(3-4)^2 + (4-4)^2} = 1$$

As $(3,4)$ is not a core point (only 2 points in the ε-neighborhood), $(4,4)$ is not density reachable from $(3,4)$.

This can also be visualized in the coordinate system:



Points:

● (3,4), (4,4)

○ (1,1), (1,2), (1,4), (2,1), (2,3), (3,2), (4,1), (4,3)

Task 1.4: Density Connectivity

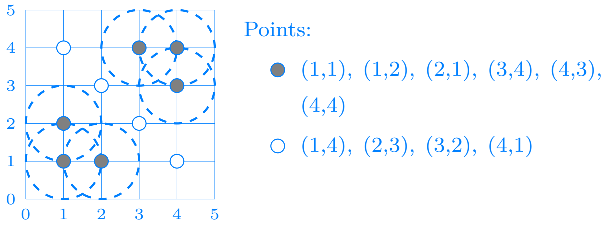
Determine whether $(1,1)$, $(3,2)$, $(4,3)$, and $(4,4)$ are **density connected** to the point $(3,4)$ if a density based clustering algorithm like **DBSCAN** is initialized with $\varepsilon = 1$ and $MinPts = 3$. The distance is calculated using the Euclidean distance.

Points are density connected if there is a core point from which both points are density reachable. Thus, we have to check whether there is a core point from which both points are density reachable.

- **Point $(1,1)$:**

  Since the group of the points $(1,1)$, $(1,2)$, and $(2,1)$ and the group of the points $(3,4)$, $(4,3)$, and $(4,4)$ don't have any points outside of their group in their ε-neighborhoods (and are therefore also not in the ε-neighborhoods of points outside of their group), there is no core point from which both $(1,1)$ and $(3,4)$ are both density reachable. Thus, $(1,1)$ is not density connected to $(3,4)$.

  This can also be visualized in the coordinate system:



Points:

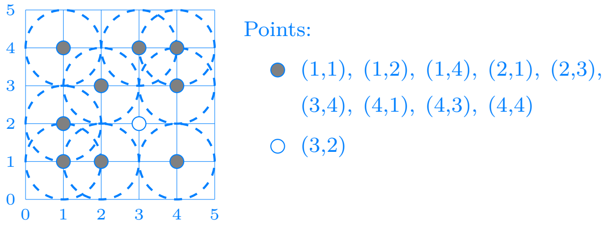
● (1,1), (1,2), (2,1), (3,4), (4,3), (4,4)

○ (1,4), (2,3), (3,2), (4,1)

- **Point** $(3,2)$**:**

  As $(3,2)$ is not in the ε-neighborhood of any other point, it can not be reached from any core point. Thus, $(3,2)$ is not density connected to $(3,4)$.

  This can also be visualized in the coordinate system:

- **Point** $(4,3)$**:**

  While neither $(4,3)$ nor $(3,4)$ seem to be core points, they seem to be density reachable from the core point $(4,4)$. To prove that $(4,3)$ is density connected to $(3,4)$, we have to show the following:

  1. The point $(4,4)$ is a core point:

  $$\begin{aligned} Distance_{(4,4)\leftrightarrow(4,4)} &= \sqrt{(4-4)^2+(4-4)^2} = 0 \\ Distance_{(4,4)\leftrightarrow(4,3)} &= \sqrt{(4-4)^2+(4-3)^2} = 1 \\ Distance_{(4,4)\leftrightarrow(3,4)} &= \sqrt{(4-3)^2+(4-4)^2} = 1 \end{aligned}$$

  2. The point $(4,3)$ is directly density reachable from the point $(4,4)$:

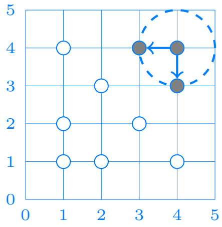
  $$Distance_{(4,4)\leftrightarrow(4,3)} = \sqrt{(4-4)^2+(4-3)^2} = 1$$

  3. The point $(3,4)$ is directly density reachable from the point $(4,4)$:

  $$Distance_{(4,4)\leftrightarrow(3,4)} = \sqrt{(4-3)^2+(4-4)^2} = 1$$

  Thus, $(4,3)$ is density connected to $(3,4)$.

  This can also be visualized in the coordinate system:



Points:

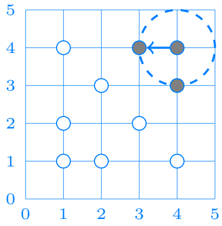
● (3,4), (4,3), (4,4)

○ (1,1), (1,2), (1,4), (2,1), (2,3), (3,2), (4,1)

- **Point $(4, 4)$:**

  As we already have shown that $(4, 4)$ is a core point and that $(3, 4)$ is directly density reachable from $(4, 4)$, we can conclude that $(4, 4)$ is density connected to $(3, 4)$.

  This can also be visualized in the coordinate system:



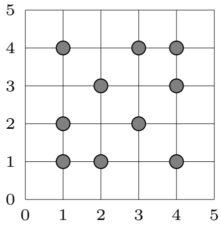
Points:

● (3,4), (4,4), (4,3)

○ (1,1), (1,2), (1,4), (2,1), (2,3), (3,2), (4,1)

Task 2: Application of DBSCAN

Given is a set of points in a two-dimensional space:



Points:

● (1,1), (1,2), (1,4), (2,1), (2,3), (3,2), (3,4), (4,1), (4,3), (4,4)

Apply the **DBSCAN** algorithm known from the lecture on the given point set while using $\varepsilon = 1$ and $MinPts = 2$.

Write down **all** intermediate steps.

The DBSCAN algorithm can either be structured recursively or iteratively.

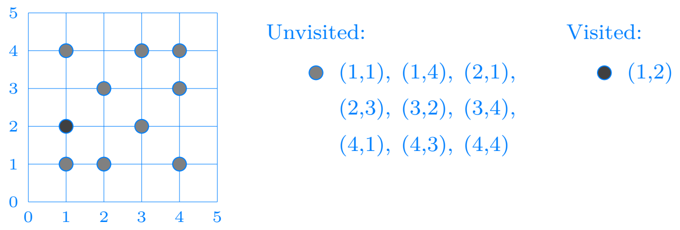
In this sample solution, we will structure it iteratively, as it is less nested and therefore easier to write down.

1. **Select $(1, 2)$ as Random Point:**

   Every point can be selected as the starting point. In this sample solution, we randomly decided to use $(1, 2)$ as the starting point:

a) **Mark $(1, 2)$ as Visited:**

Points should only be visited once. This is important to avoid infinite loops:



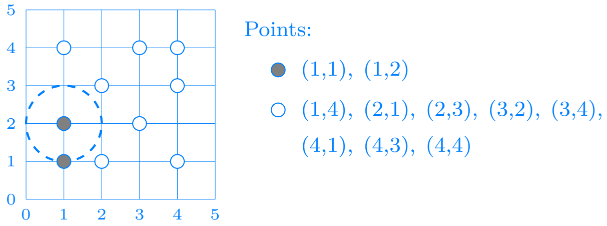
b) **Check if $(1, 2)$ is a Core Point:**

If there are at least 2 points in the ε-neighborhood of $(1, 2)$, $(1, 2)$ is a core point:

$$Distance_{(1,2)\leftrightarrow(1,2)} = \sqrt{(1-1)^2 + (2-2)^2} = 0$$
$$Distance_{(1,2)\leftrightarrow(1,1)} = \sqrt{(1-1)^2 + (2-1)^2} = 1$$

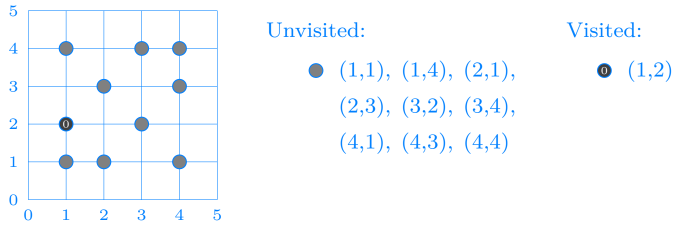
Therefore, $(1, 2)$ is a core point.

Shown in the coordinate system:



c) **Create a New Cluster and Add $(1, 2)$:**

In this sample solution, we simply name this cluster "0":



d) **Add Points in the ε-Neighborhood of $(1, 2)$ to the Candidate Set N:**

Only $(1, 1)$ is within the does ε-Neighborhood of $(1, 2)$ and does not belong to a cluster yet:

$$Distance_{(1,2)\leftrightarrow(1,1)} = \sqrt{(1-1)^2 + (2-1)^2} = 1$$

So only $(1,1)$ is added to the candidate set N:

$$N = \{(1,1)\}$$

e) **Iterate through the Candidate Set N (for Cluster 0):**

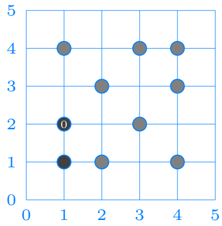
The candidate set N is iterated through until it is empty. All points visited in this iteration will be added to cluster 0, since this iteration was started with point $(1,2)$:

i. **For $(1,1)$:**

A. **Remove $(1,1)$ from the Candidate Set N:**

$$N = \{\}$$

B. **Mark $(1,1)$ as Visited:**



Unvisited:

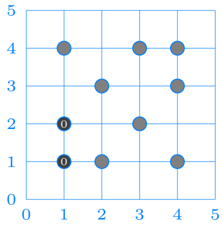
● (1,4), (2,1), (2,3), (3,2), (3,4), (4,1), (4,3), (4,4)

Visited:

● (1,1)

⓪ (1,2)

C. **Add $(1,1)$ to Cluster 0:**



Unvisited:

● (1,4), (2,1), (2,3), (3,2), (3,4), (4,1), (4,3), (4,4)

Visited:

⓪ (1,2), (1,1)

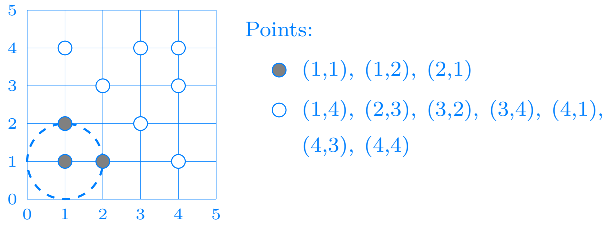
D. **Check if $(1,1)$ is a Core Point:**

If there are at least 2 points in the ε-neighborhood of $(1,1)$, $(1,1)$ is a core point:

$$Distance_{(1,1)\leftrightarrow(1,1)} = \sqrt{(1-1)^2 + (1-1)^2} = 0$$
$$Distance_{(1,1)\leftrightarrow(1,2)} = \sqrt{(1-1)^2 + (1-2)^2} = 1$$
$$Distance_{(1,1)\leftrightarrow(2,1)} = \sqrt{(1-2)^2 + (1-1)^2} = 1$$

Therefore, $(1,1)$ is a core point.

Shown in the coordinate system:



E. **Add Points in the ε-Neighborhood of $(1,1)$ to the Candidate Set N:**

Only $(2,1)$ is within the ε-neighborhood of $(1,1)$ and does not belong to a cluster yet:

$$Distance_{(1,1)\leftrightarrow(2,1)} = \sqrt{(1-2)^2 + (1-1)^2} = 1$$

So only $(2,1)$ is added to the candidate set N:

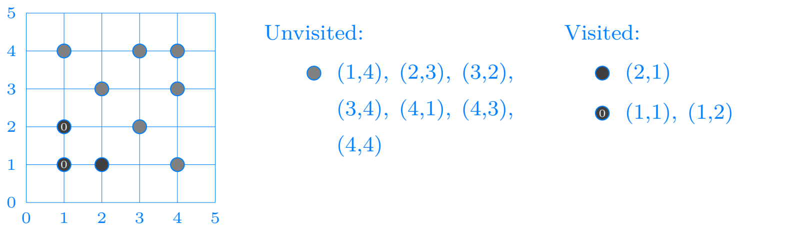
$$N = \{(2,1)\}$$

ii. **For $(2,1)$:**

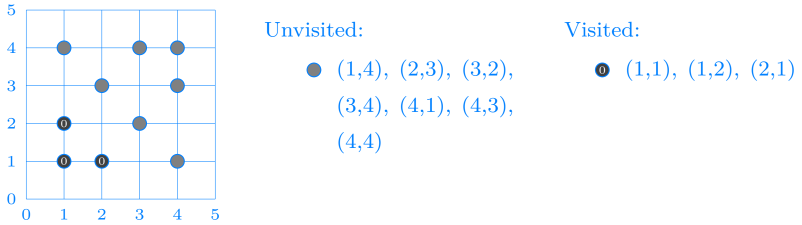
A. **Remove $(2,1)$ from the Candidate Set N:**

$$N = \{\}$$

B. **Mark $(2,1)$ as Visited:**



C. **Add** $(2,1)$ **to Cluster** 0**:**



D. **Check if** $(2,1)$ **is a Core Point:**

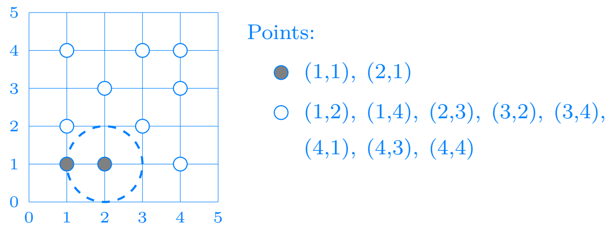
If there are at least 2 points in the ε-neighborhood of $(2,1)$, $(2,1)$ is a core point:

$$Distance_{(2,1)\leftrightarrow(1,1)} = \sqrt{(2-1)^2+(1-1)^2} = 1$$
$$Distance_{(2,1)\leftrightarrow(2,1)} = \sqrt{(2-2)^2+(1-1)^2} = 0$$

Therefore, $(2,1)$ is a core point.

. . . .

Shown in the coordinate system:



. . . .

E. **Add Points in the** ε**-Neighborhood of** $(2,1)$ **to the Candidate Set** N**:**

There are no unvisited points in the ε-neighborhood of $(2,1)$.

Thus, the candidate set N remains empty.

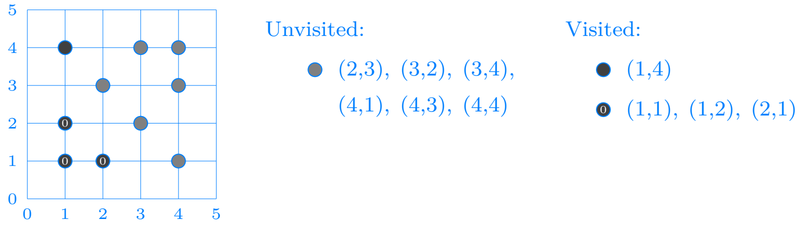
iii. **Stop the Iteration:**

The candidate set N is empty, thus the iteration is stopped.

2. **Select** $(1, 4)$ **as Random Point:**

   Every unvisited point can be selected as the next point to visit. In this sample solution, we decided to use $(1, 4)$ as the next point:

   a) **Mark** $(1, 4)$ **as Visited:**



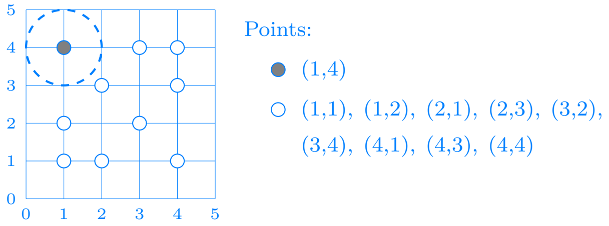
   b) **Check if** $(1, 4)$ **is a Core Point:**

   If there are at least 2 points in the ε-neighborhood of $(1, 4)$, $(1, 4)$ is a core point:

$$Distance_{(1,4)\leftrightarrow(1,4)} = \sqrt{(1-1)^2 + (4-4)^2} = 0$$

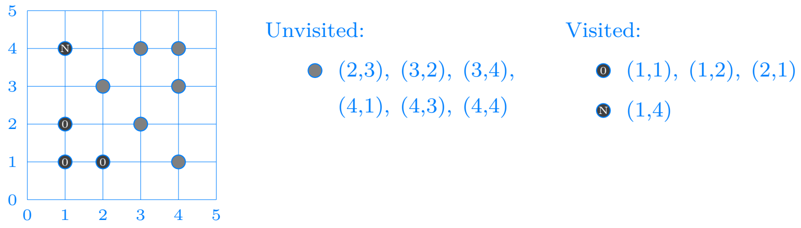
   Only $(1, 4)$ is in the ε-neighborhood of $(1, 4)$, thus $(1, 4)$ is not a core point.

   Shown in the coordinate system:



   c) **Mark** $(1, 4)$ **as Noise:**

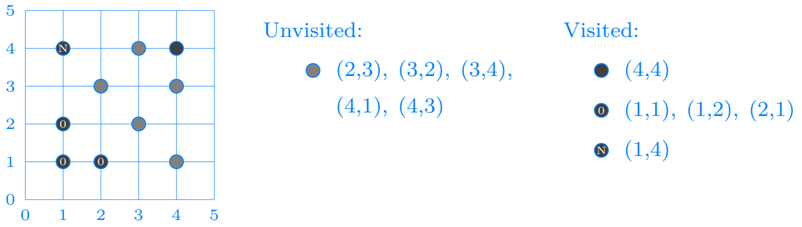
   Since $(1, 4)$ is not a core point, it is marked as noise:



3. **Select** $(4, 4)$ **as Random Point:**

   Every unvisited point can be selected as the next point to visit. In this sample solution, we decided to use $(4, 4)$ as the next point:

   a) **Mark** $(4, 4)$ **as Visited:**



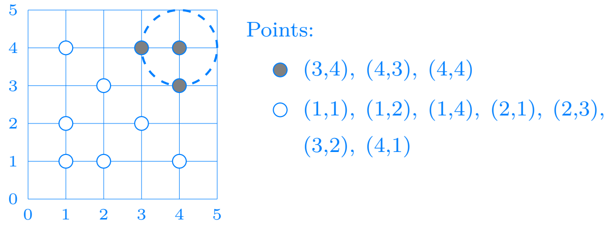
b) **Check if $(4, 4)$ is a Core Point:**

If there are at least 2 points in the ε-neighborhood of $(4, 4)$, $(4, 4)$ is a core point:

$$\begin{aligned}
Distance_{(4,4)\leftrightarrow(4,4)} &= \sqrt{(4-4)^2 + (4-4)^2} = 0\\
Distance_{(4,4)\leftrightarrow(3,4)} &= \sqrt{(4-3)^2 + (4-4)^2} = 1\\
Distance_{(4,4)\leftrightarrow(4,3)} &= \sqrt{(4-4)^2 + (4-3)^2} = 1
\end{aligned}$$

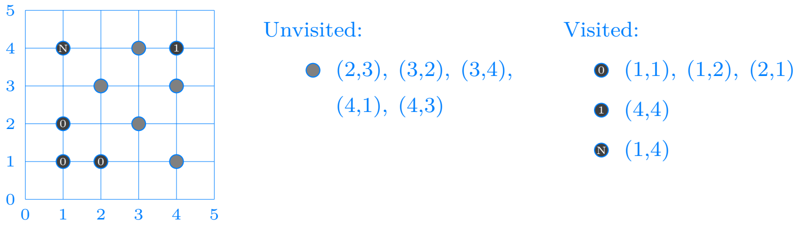
Therefore, $(4, 4)$ is a core point.

Shown in the coordinate system:



c) **Create a New Cluster and Add $(4, 4)$:**

In this sample solution, we simply name this cluster "1":



d) **Add Points in the ε-Neighborhood of $(4, 4)$ to the Candidate Set N:**

Both $(3, 4)$ and $(4, 3)$ are within the ε-neighborhood of $(4, 4)$ and do not belong to a cluster yet:

$$Distance_{(4,4)\leftrightarrow(3,4)} = \sqrt{(4-3)^2 + (4-4)^2} = 1$$
$$Distance_{(4,4)\leftrightarrow(4,3)} = \sqrt{(4-4)^2 + (4-3)^2} = 1$$

Thus, both $(3,4)$ and $(4,3)$ are added to the candidate set N:

$$N = \{(3,4),(4,3)\}$$

e) **Iterate through the Candidate Set N (for Cluster 1):**

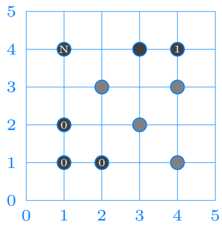
The candidate set N is iterated through until it is empty. All points visited in this iteration will be added to cluster 1, since this iteration was started with point $(4,4)$:

i. **For $(3,4)$:**

A. **Remove $(3,4)$ from the Candidate Set N:**

$$N = \{(4,3)\}$$

B. **Mark $(3,4)$ as Visited:**



Unvisited:

● (2,3), (3,2), (4,1), (4,3)

Visited:

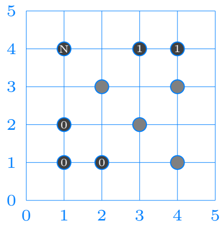
● (3,4)

0 (1,1), (1,2), (2,1)

1 (4,4)

N (1,4)

C. **Add $(3,4)$ to Cluster 1:**



Unvisited:

● (2,3), (3,2), (4,1), (4,3)

Visited:

0 (1,1), (1,2), (2,1)

1 (3,4), (4,4)

N (1,4)

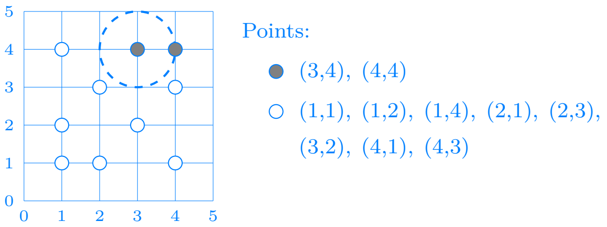
D. **Check if $(3,4)$ is a Core Point:**

If there are at least 2 points in the ε-neighborhood of $(3,4)$, $(3,4)$ is a core point:

$$Distance_{(3,4)\leftrightarrow(3,4)} = \sqrt{(3-3)^2+(4-4)^2} = 0$$
$$Distance_{(3,4)\leftrightarrow(4,4)} = \sqrt{(3-4)^2+(4-4)^2} = 1$$

Therefore, $(3,4)$ is a core point.

Shown in the coordinate system:



E. **Add Points in the ε-Neighborhood of $(3,4)$ to the Candidate Set N:**

There are no unvisited points in the ε-neighborhood of $(3,4)$.

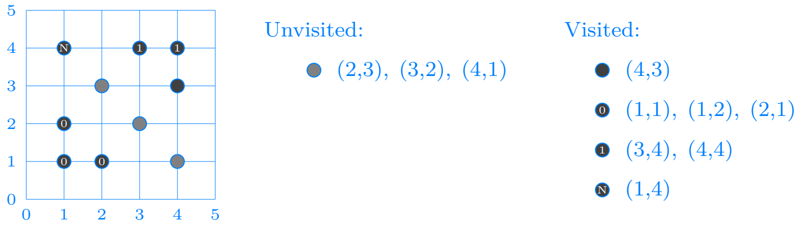
Thus, the candidate set N remains the same.

ii. **For $(4,3)$:**

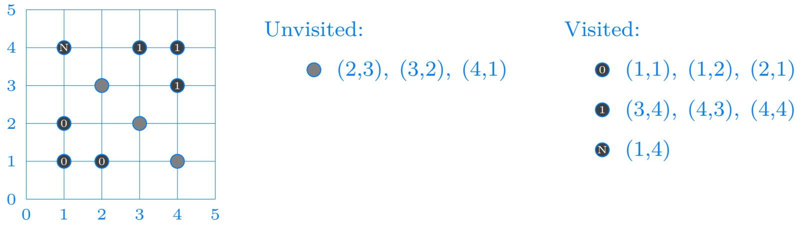
A. **Remove $(4,3)$ from the Candidate Set N:**

$$N = \{\}$$

B. **Mark $(4,3)$ as Visited:**



C. **Add $(4,3)$ to Cluster 1:**



D. **Check if $(4,3)$ is a Core Point:**

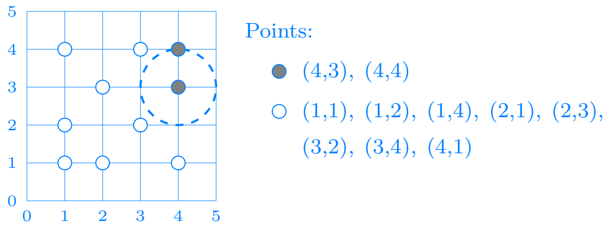
If there are at least 2 points in the ε-neighborhood of $(4,3)$, $(4,3)$ is a core point:

$$Distance_{(4,3)\leftrightarrow(4,3)} = \sqrt{(4-4)^2 + (3-3)^2} = 0$$
$$Distance_{(4,3)\leftrightarrow(4,4)} = \sqrt{(4-4)^2 + (3-4)^2} = 1$$

Therefore, $(4,3)$ is a core point.

. . . .

Shown in the coordinate system:



. . . .

E. **Add Points in the ε-Neighborhood of $(4,3)$ to the Candidate Set N:**

There are no unvisited points in the ε-neighborhood of $(4,3)$.

Thus, the candidate set N remains empty.

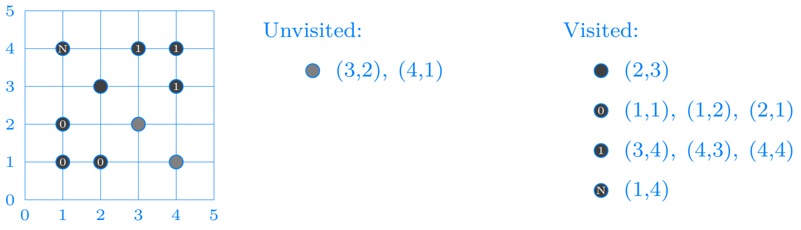
iii. **Stop the Iteration:**

The candidate set N is empty, thus the iteration is stopped.

f) **Select $(2,3)$ as Random Point:**

Every unvisited point can be selected as the next point to visit. In this sample solution, we decided to use $(2,3)$ as the next point:

i. **Mark $(2,3)$ as Visited:**



ii. **Check if $(2,3)$ is a Core Point:**

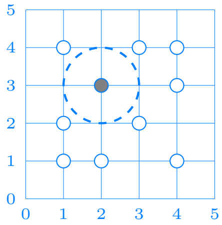
If there are at least 2 points in the ε-neighborhood of $(2,3)$, $(2,3)$ is a core point:

$$Distance_{(2,3)\leftrightarrow(2,3)} = \sqrt{(2-2)^2 + (3-3)^2} = 0$$

Only $(2,3)$ is in the ε-neighborhood of $(2,3)$, thus $(2,3)$ is not a core point.

. . . .

Shown in the coordinate system:



Points:

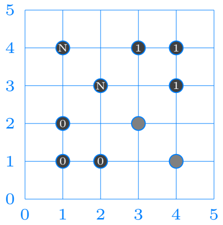
● (2,3)

○ (1,1), (1,2), (1,4), (2,1), (3,2), (3,4), (4,1), (4,3), (4,4)

. . . .

iii. **Mark** $(2,3)$ **as Noise:**

Since $(2,3)$ is not a core point, it is marked as noise:



Unvisited:

● (3,2), (4,1)

Visited:

0 (1,1), (1,2), (2,1)

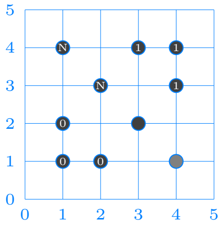
1 (3,4), (4,3), (4,4)

N (1,4), (2,3)

g) **Select** $(3,2)$ **as Random Point:**

Every unvisited point can be selected as the next point to visit. In this sample solution, we decided to use $(3,2)$ as the next point:

i. **Mark** $(3,2)$ **as Visited:**



Unvisited:

● (4,1)

Visited:

● (3,2)

0 (1,1), (1,2), (2,1)

1 (3,4), (4,3), (4,4)

N (1,4), (2,3)

ii. **Check if** $(3,2)$ **is a Core Point:**

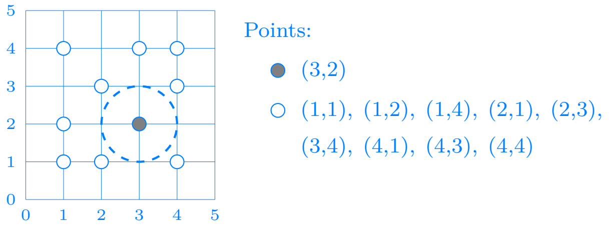
If there are at least 2 points in the ε-neighborhood of $(3,2)$, $(3,2)$ is a core point:

$$Distance_{(3,2)\leftrightarrow(3,2)} = \sqrt{(3-3)^2 + (2-2)^2} = 0$$

Only $(3,2)$ is in the ε-neighborhood of $(3,2)$, thus $(3,2)$ is not a core point.

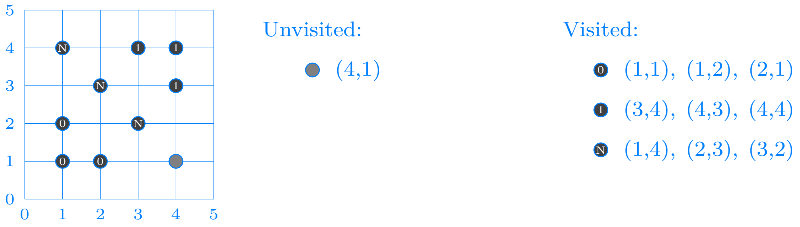
. . . .

Shown in the coordinate system:



iii. **Mark** $(3,2)$ **as Noise:**

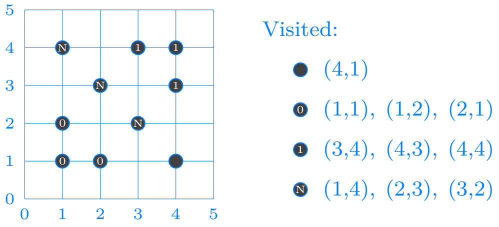
Since $(3,2)$ is not a core point, it is marked as noise:



h) **Select** $(4,1)$ **as Random Point:**

Every unvisited point can be selected as the next point to visit. In this sample solution, we decided to use $(4,1)$ as the next point:

i. **Mark** $(4,1)$ **as Visited:**



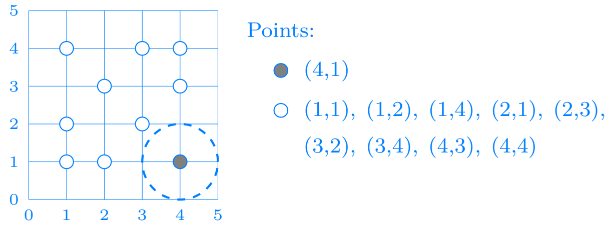
ii. **Check if** $(4,1)$ **is a Core Point:**

If there are at least 2 points in the ε-neighborhood of $(4,1)$, $(4,1)$ is a core point:

$$Distance_{(4,1)\leftrightarrow(4,1)} = \sqrt{(4-4)^2 + (1-1)^2} = 0$$

Only $(4,1)$ is in the ε-neighborhood of $(4,1)$, thus $(4,1)$ is not a core point.

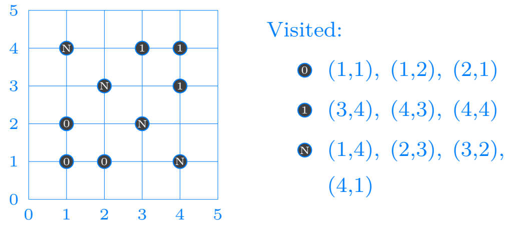
Shown in the coordinate system:



. . . .

iii. **Mark** $(4,1)$ **as Noise:**

Since $(4,1)$ is not a core point, it is marked as noise:



i) **Stop the Algorithm:**

All points have been visited, thus the algorithm is stopped.

The final results are:

$$\begin{aligned} \text{Cluster 0} &: \{(1,1),(1,2),(2,1)\} \\ \text{Cluster 1} &: \{(3,4),(4,3),(4,4)\} \\ \text{Noise} &: \{(1,4),(2,3),(3,2),(4,1)\} \end{aligned}$$

Exercise 3: Clustering in Python

This exercise comprises practical data science tasks and thus utilizes a Jupyter Notebook:

1. Open `Clustering-in-Python.ipynb`.
2. Take a look at the tasks (blue boxes) in the notebook and try to solve them.

If you are unfamiliar with how to open a Jupyter Notebook, please refer to Exercise 1 of `1-Introduction-Python-Pandas.pdf`.

The solution to the exercise can be found in `Additional-Files-Solution.zip`.