Maschinelles Lernen II

Bilal Al Homsi

ML II

- Unüberwachtes Lernen:
  - Clusteranalyse:
    - k-means
    - k nearest neighbor
    - hierarchisches Clustering
- Überwachtes Lernen:
- Support Vector Machines
- Deep Learning:
  - Künstliche Neuronale Netze

- Der Unterschied dazwischen ist, dass es beim Überwachten Lernen Daten mit Labels genutzt werden, beim Unüberwachten Lernen nicht.

Unüberwachtes Lernen

- Clusteranalyse:
  - K-Means

- Es wird bei diesem Ansatz versucht, Strukturen und Muster in den unbekannten Daten zu finden.
- Der Algorithmus benötigt für das Clustering Eingabedaten und eine geeignete Anzahl an Clustern, denen die Daten zugeordnet werden können. Das Verfahren terminiert nach so vielen Schritten dann, wenn es in der aktuellen Iteration keine Änderungen an den Daten in den Clustern festgestellt werden.

K-Means Algorithmus (Pseudocode)

- Input:
  - **X**: Datenpunkte, **k**: Anzahl der Cluster, **d(x, y)**: Distanzfunktion
- Output:
  - **C**: Clusterzentren
- Pseudocode:
  1. Initialisiere Clusterzentren **C** zufällig aus **X**.
  2. Wiederhole, bis Konvergenz erreicht ist:
     a. Weise jedem Datenpunkt **x** dem nächsten Clusterzentrum **c** zu.
     b. Aktualisiere die Clusterzentren **C** basierend auf den zugewiesenen Datenpunkten.
  3. Die Menge der Clusterzentren **C** sind die endgültigen Cluster.
4. Gib **C** als Ergebnis zurück.

Funktionen

Distanzfunktion

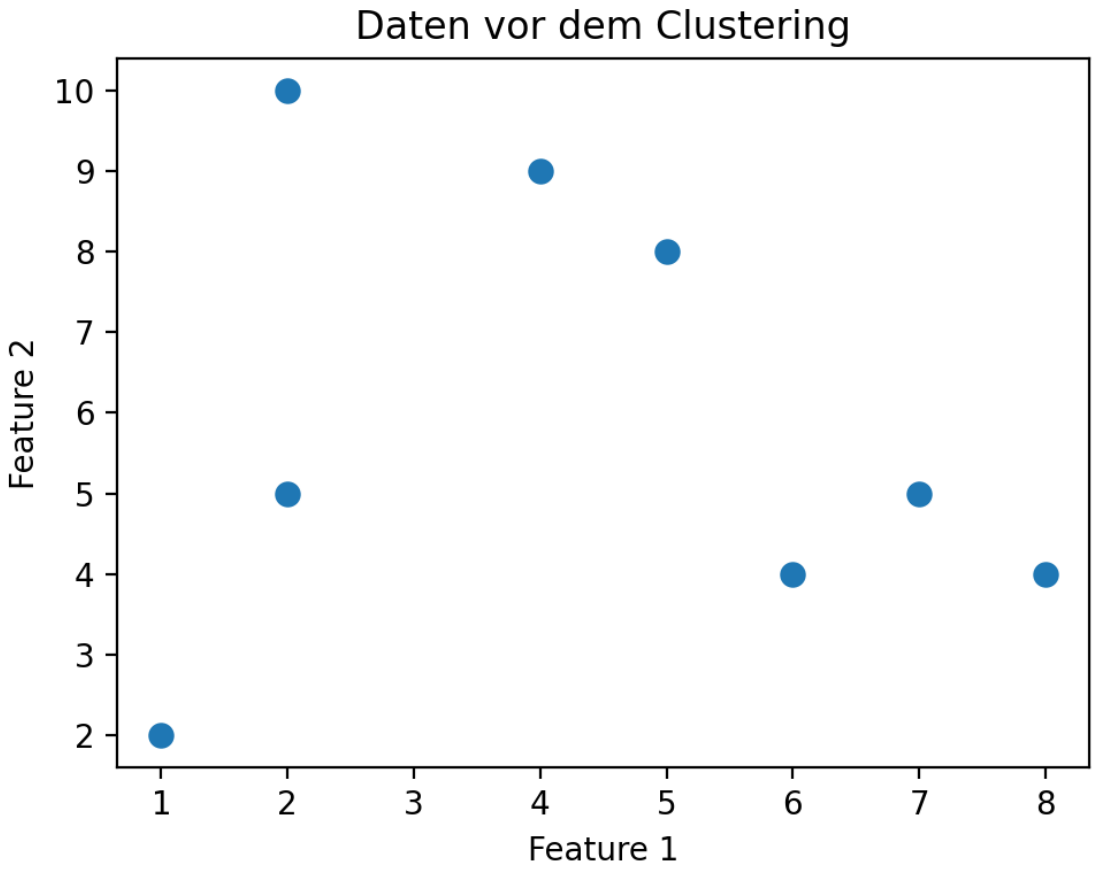
$$d(x_i, x_j) = |x_i - x_j|$$

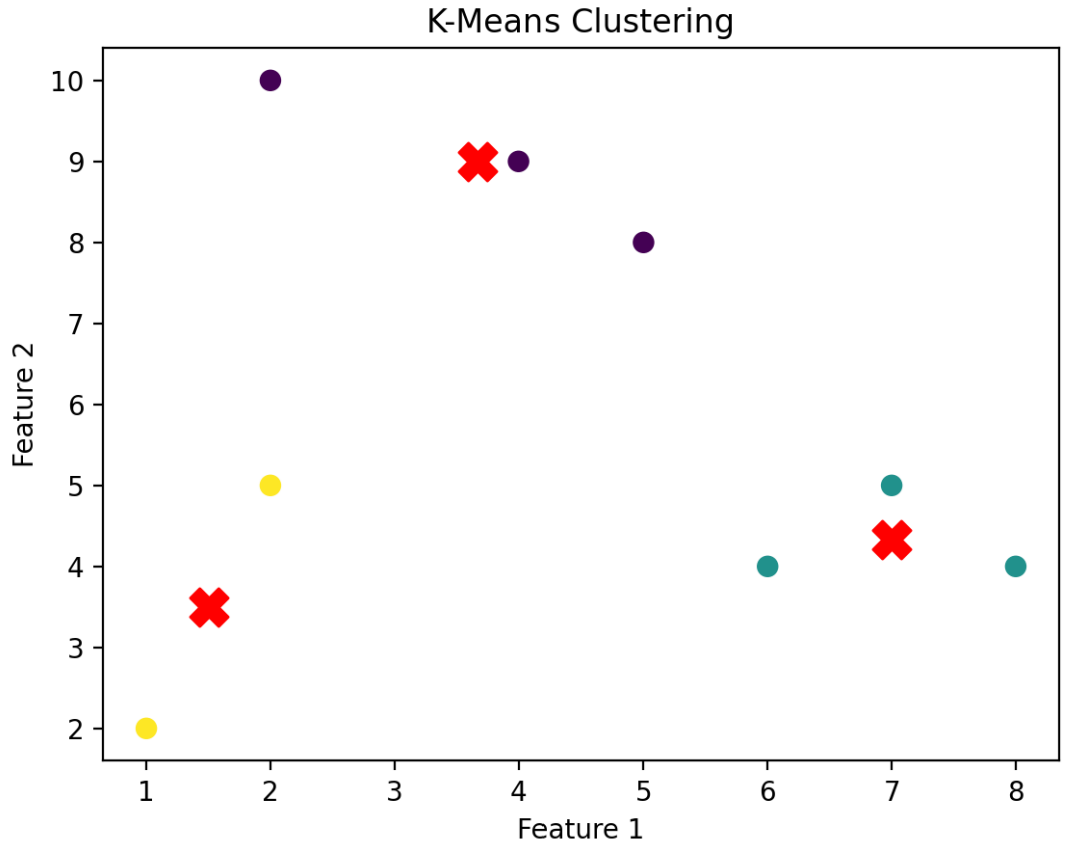
Berechnungsfunktion der Clusterzentren

$$C_i = \frac{1}{|X_i|} \sum_{x \in X_i} x$$

Daten vor und nach dem Clustering

X = [(2, 10), (2, 5), (8, 4), (5, 8), (7, 5), (6, 4), (1, 2), (4, 9)], C = 3

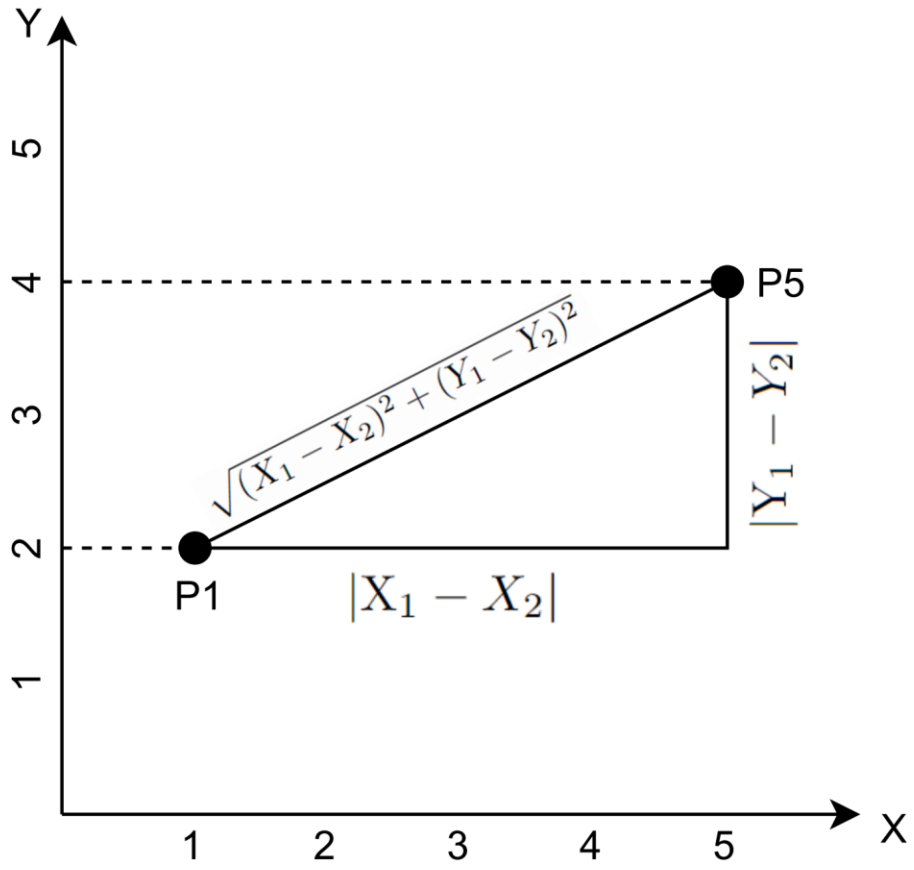




Hierarchisches Clustering

- Ist ein weiteres Verfahren zum Clustering. Dazu werden sowohl eine Abstandstabelle, die im Laufe des Verfahrens aktualisiert wird, als auch ein sogenannten Dendrogram erstellt, welches die Beziehung zwischen den Datenpunkten bzw. Clustern darstellt.
- Für dieses Verfahten benötigt man:
  - Datenpunkte
  - Eine Methode zur Ermittlung der Abstände
  - Eine Methode zur Festlegung, von welchem Datenpunkt man den Abstand zwischen zwei Clustern.

Methoden zur Ermittlung der Abstände zwischen zwei Punkten

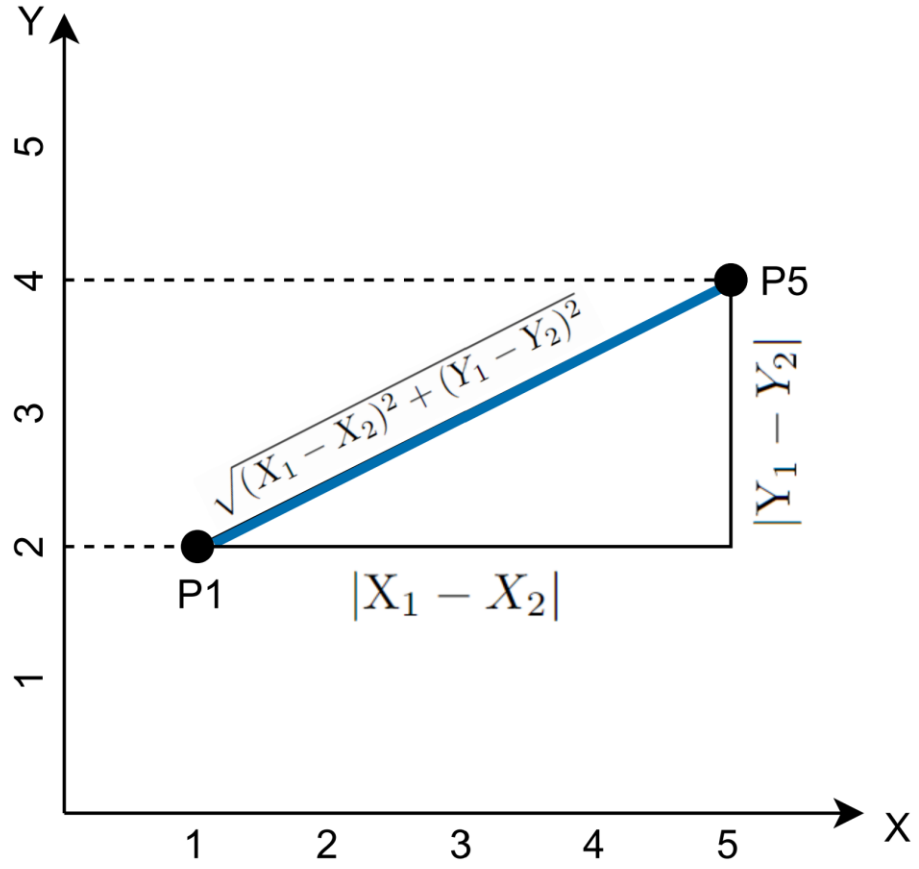


a) **Euklidischer Abstand:** $\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$

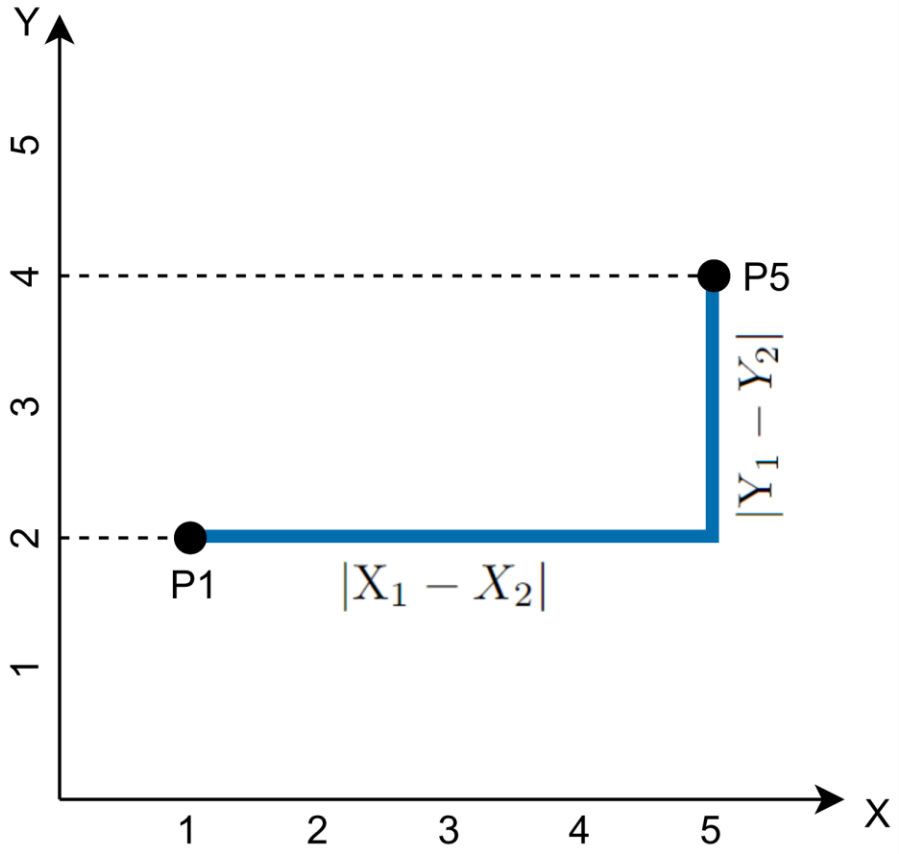
b) **Manhattan-Distanz:** $|X_1 - X_2| + |Y_1 - Y_2|$

c) **Maximale Distanz:** $\mathbf{Max}((|X_1 - X_2|), (|Y_1 - Y_2|))$

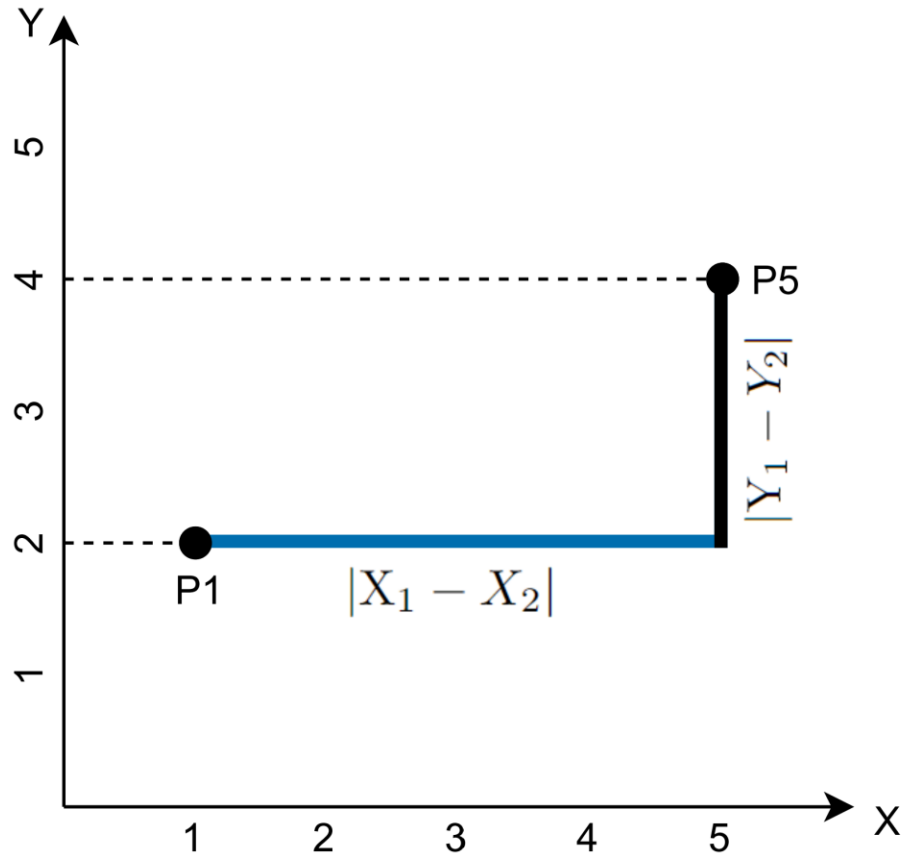
Euklidischer Abstand



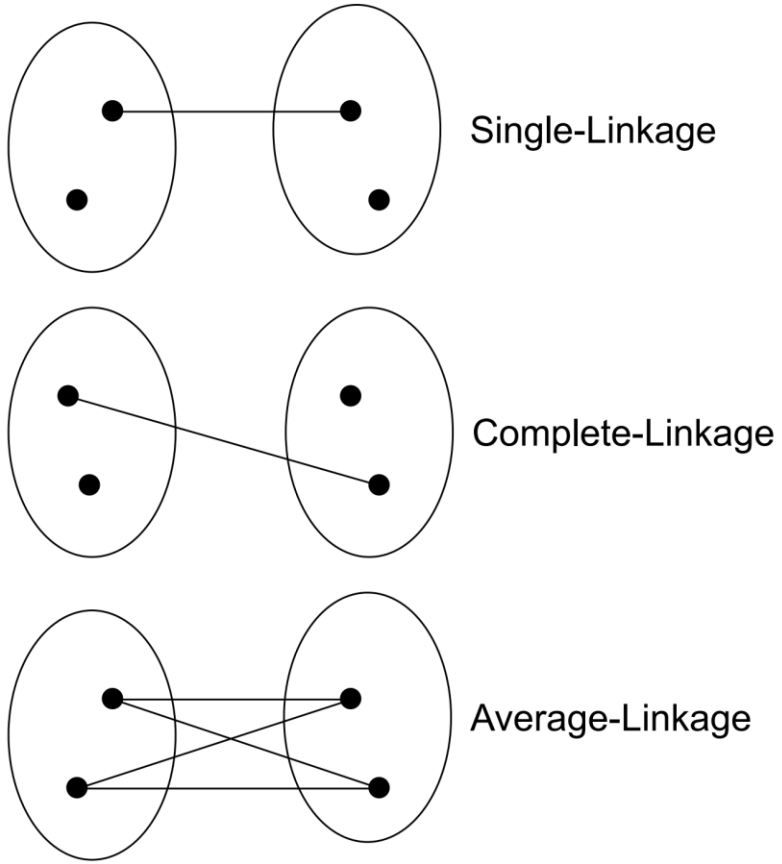
Manhatten Distanz



Maximale Distanz



Methoden zur Bestimmung der Abstände zwischen zwei Punkten



Algorithmus

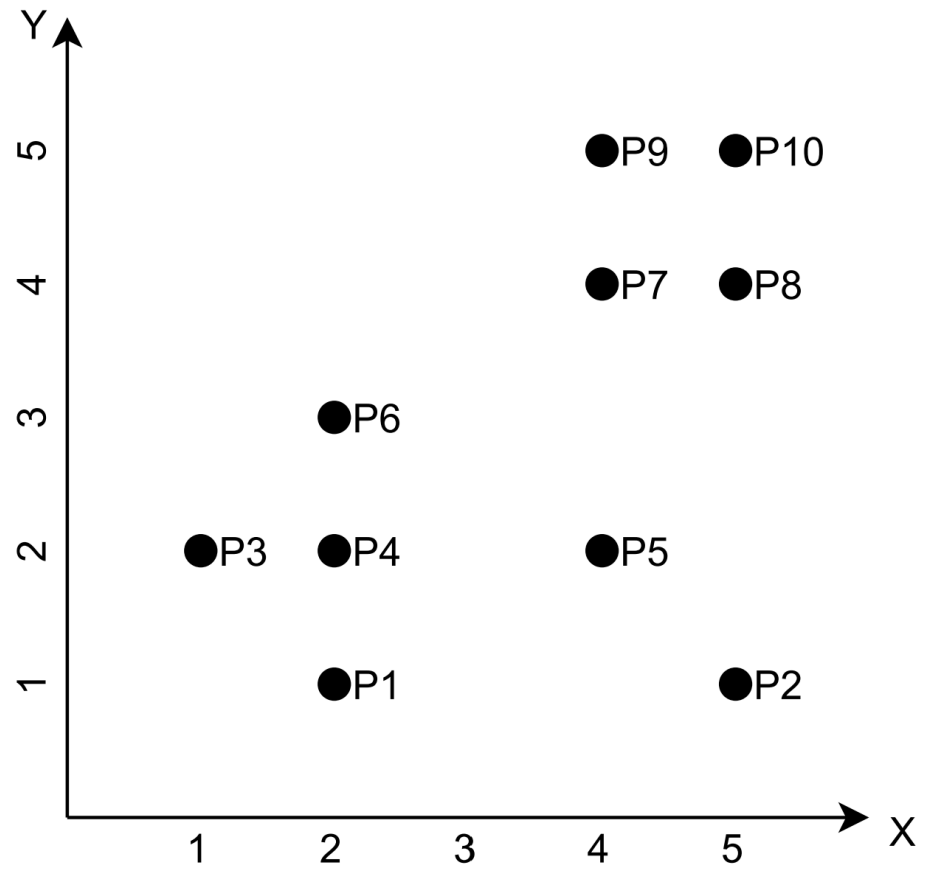
- Jeder Datenpunkt wird als ein Cluster betrachtet.
- Es wird der Abstand zwischen den einzelnen Datenpunkten berechnet und in eine Tabelle eingetragen.
  - **Euklidischer Abstand, Manhatten-Distanz** oder **maximale Distanz**
- Es werden die zwei Punkte mit dem minimalen Abstand zu einem Cluster fusioniert.
  - Wenn mehrere Punkte den gleichen Abstand haben, dann können zwei Punkte wahllos fusioniert werden.
- Um den Abstand zwischen zwei Clustern (Cluster hat mind. zwei Datenpunkte) zu bestimmen, wird eine der folgenden Methoden genutzt:
  - **Single-Linkage, Complete-Linkage** oder **Average-Linkage**
- Die Abstandstabelle wird nach jeder Iteration aktualisiert.
- Nach jeder Iteration wird ein neues Cluster in dem **Dendrogram** dargestellt.
- Das Verfahren terminiert, wenn alles Datenpunkte in einem Cluster enthalten sind.

Beispiel

Gegeben ist folgender Datensatz:

- X = [(2, 1) (5, 1), (1, 2), (2, 2), (4, 2), (2, 3), (4, 4), (5, 4), (4, 5), (5, 5)]

- Zunächst werden Datenpunkte in ein Koordinatensystem eingetragen.
- Es wird die Methode **maximale Distanz** verwendet.
- Es wird die Methode **Single-Lingkage** verwendet.



Beispiel

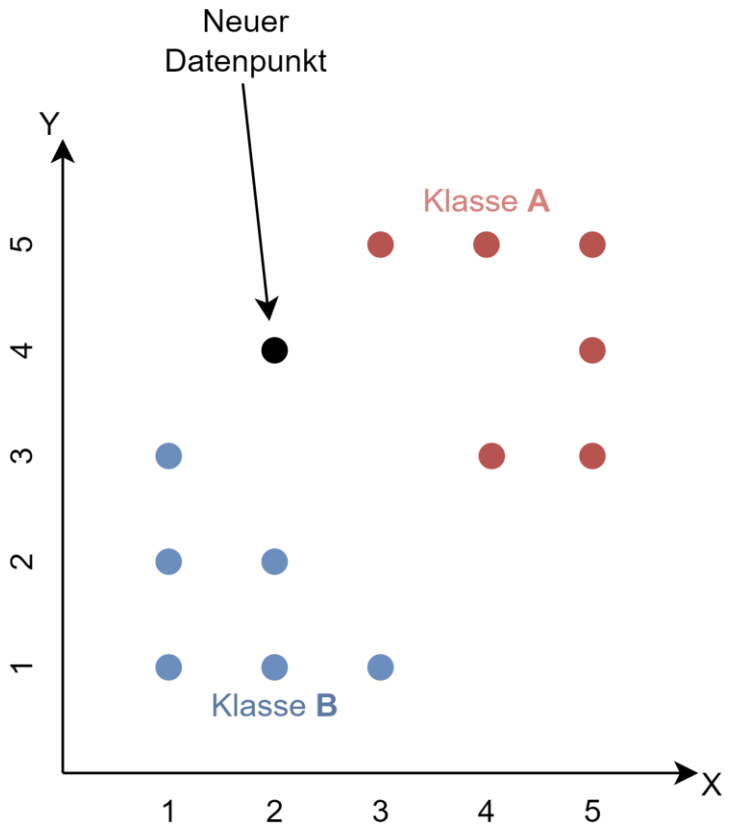
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	0									
P2	3	0								
P3	1	4	0							
P4	1	3	1	0						
P5	2	1	3	2	0					
P6	2	3	1	1	2	0				
P7	3	3	3	2	2	2	0			
P8	3	3	4	3	2	3	1	0		
P9	4	4	3	3	3	2	1	1	0	
P10	4	4	4	3	3	3	1	1	1	0

K-nearest neighbor KNN (K-Nächste-Nachbarn-Klassifikation)

- Das ist ein überwachtes Klassifikationsverfahren, welches für die Klassenzuordnung unbekannter Datenpunkte unter Berücksichtigung der nächsten ***K*** Nachbarn.
- Für die Ermittlung der Abstände zwischen dem neuen unbekannten Datenpunkt und allen übrigen Datenpunkten wird der **euklidische Abstand** verwendet.
- Es werden ***K*** Datenpunkte betrachtet, deren Abstände zum neuen Datenpunkt am kleinsten sind.
- Zur Klassenzuordnung wird Klassen mit der maximalen Anzahl an Datenpunkten.

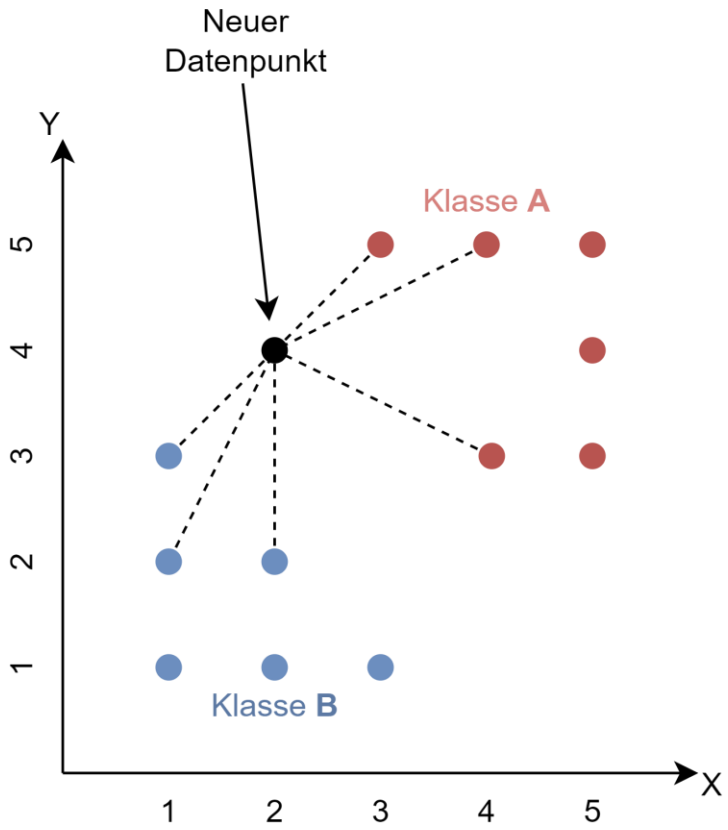
Beispiel

- Gegeben ist eine Klassenverteilung von geclusterten Datenpunkten
- Es gibt einen neuen Datenpunkt, der einer der gegebenen Klassen zuzuordnen wird.



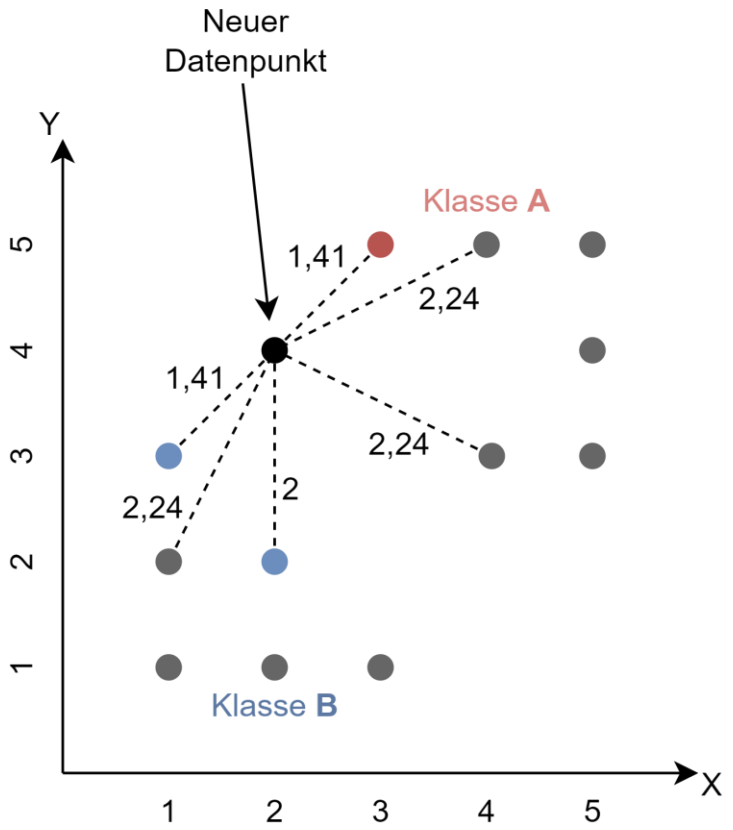
Beispiel

- Dafür werden die Abstände zwischen dem neuen Datenpunkt und allen übrigen Datenpunkten. Diese werden mit der Formel der **Euklidischen Distanz** bestimmt.



Beispiel

- Für dieses Beispiel sei ***K=3***
- Es werden die 3 nächsten Nachbarn mit dem minimalen Abstand gesucht.
- Von der Klasse **B** wurden zwei Datenpunkte und von der Klasse **A** einen Datenpunkt ermittelt.
- Max(*A*=**1**, *B*=**2**) = **B**



Beispiel

- Der neue Datenpunkt wird somit der Klasse **B** zugeordnet.