

k -Means Clustering

Prüfungsvorbereitung

Aufgabe 1

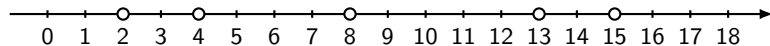
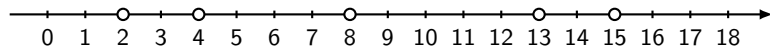
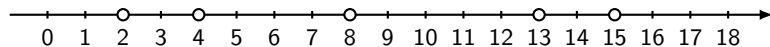
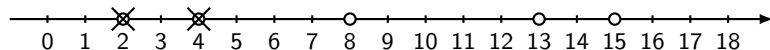
Beschreibe zwei Initialisierungsverfahren für das k -Means Clustering.

Aufgabe 1

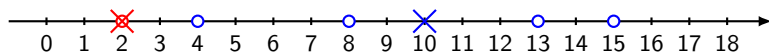
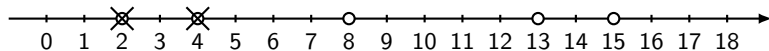
- ▶ Aus den Datenpunkten werden zufällig k Clusterzentren ausgewählt.
- ▶ Jeder Datenpunkt wird zufällig einem der k Cluster zugeordnet.

Aufgabe 2

Führe für die Punkte auf dem Zahlenstrahl ein k -Means Clustering mit $k = 2$ durch. Die initialen Clusterzentren sind durch Kreuze markiert. Gib nach jedem Schritt die Position der Clusterzentren und den Wert des Distortion Measures J an.

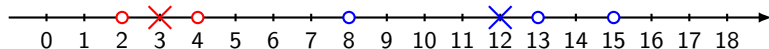


Aufgabe 2



$$C_1 = 2/1 = 2 \quad C_2 = (4 + 8 + 13 + 15)/4 = 10$$

$$J = 0^2 + 6^2 + 2^2 + 3^2 + 5^2 = 74$$



$$C_1 = (2 + 4)/2 = 3 \quad C_2 = (8 + 13 + 15)/3 = 12$$

$$J = 1^2 + 1^2 + 4^2 + 1^2 + 3^2 = 28$$

J : Summe der quadrierten Abstände aller Punkte von ihren Clusterzentren

Aufgabe 3

Erhält man mit dem k -Means-Algorithmus immer ein optimales Clustering? Wenn nein, was kann man tun, um das Resultat zu verbessern?

Aufgabe 3

Nein, der k -Means-Algorithmus findet im Allgemeinen ein *lokales Optimum*. Um das Ergebnis zu verbessern, kann man das Verfahren mehrfach mit jeweils anderen Startwerten durchführen, und dann das Resultat mit dem kleinsten Distortion Measure J wählen.

Aufgabe 4

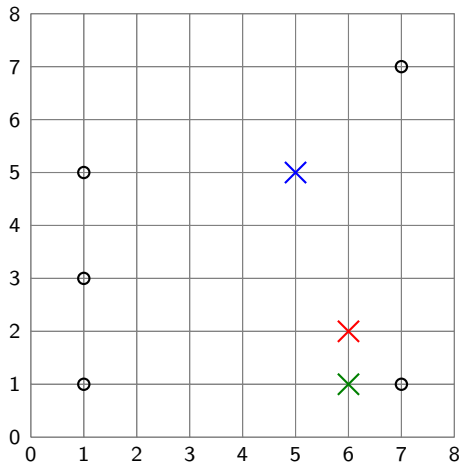
Nenne mindestens zwei Anwendungen des k -Means Clustering in der realen Welt.

Aufgabe 4

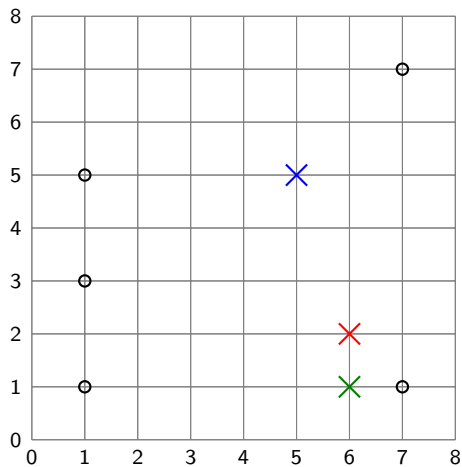
- ▶ Markt- und Kundensegmentierung
- ▶ Bildsegmentierung
- ▶ Dokumente clustern
- ▶ Empfehlungssysteme

Aufgabe 5

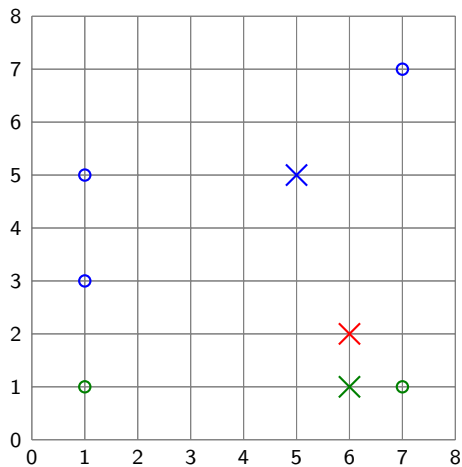
Zeige schrittweise, wie das k -Means-Verfahren die Punkte für die anfänglich gegebenen Startzentren (Kreuze) clustert.



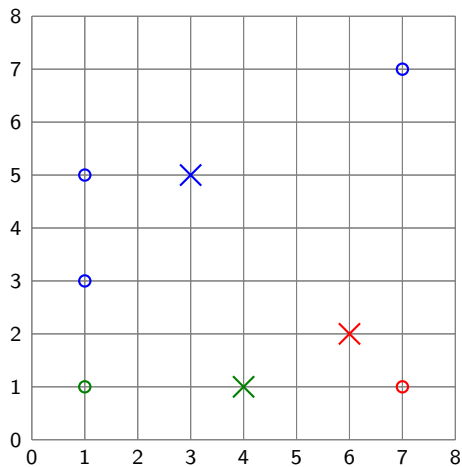
Aufgabe 5



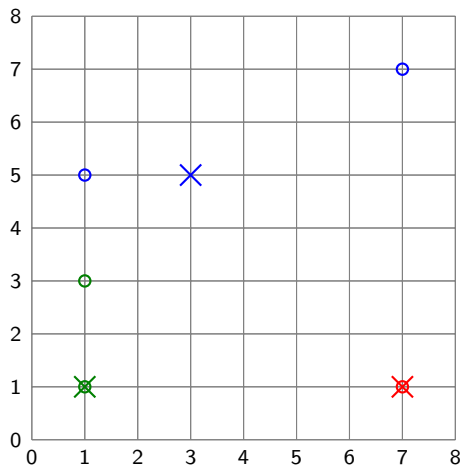
Aufgabe 5 (Schritt 1.1)



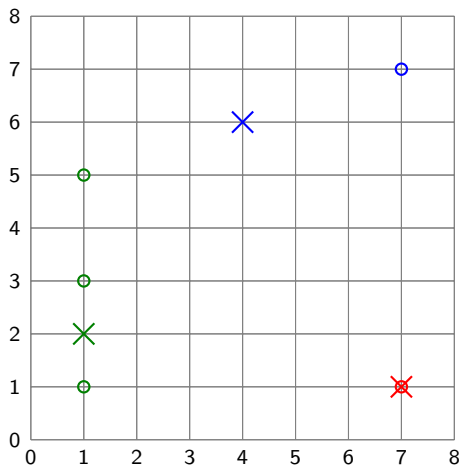
Aufgabe 5 (Schritt 1.2)



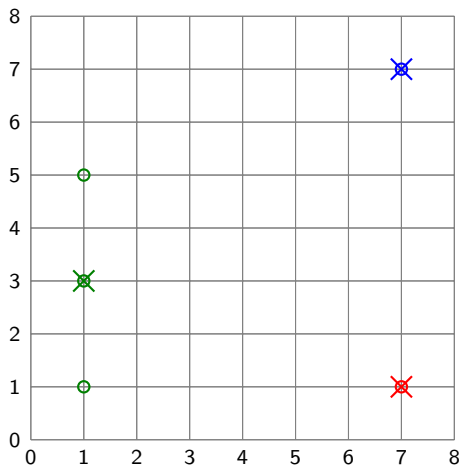
Aufgabe 5 (Schritt 2.1)



Aufgabe 5 (Schritt 2.2)

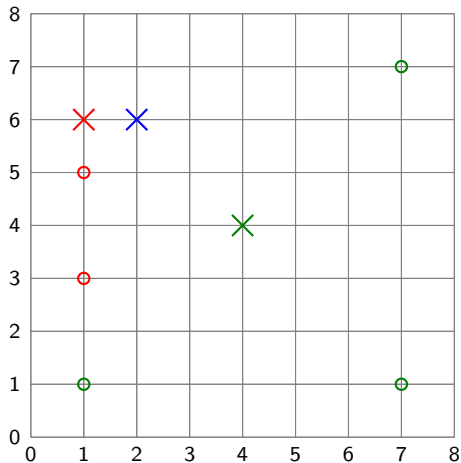


Aufgabe 5 (Schritt 3.1)

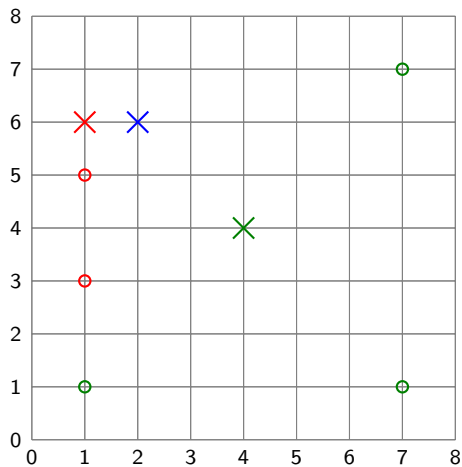


Aufgabe 6

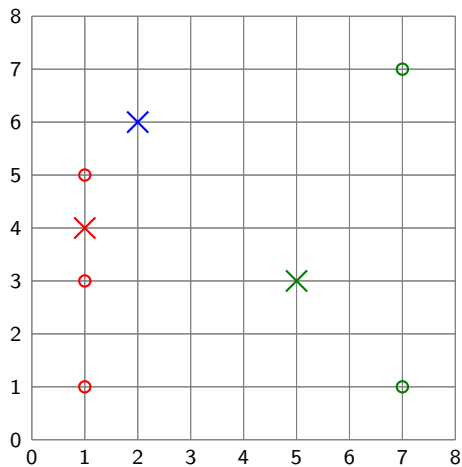
Zeige schrittweise, wie das k -Means-Verfahren die Punkte für die anfänglich gegebenen Startzentren (Kreuze) clustert.



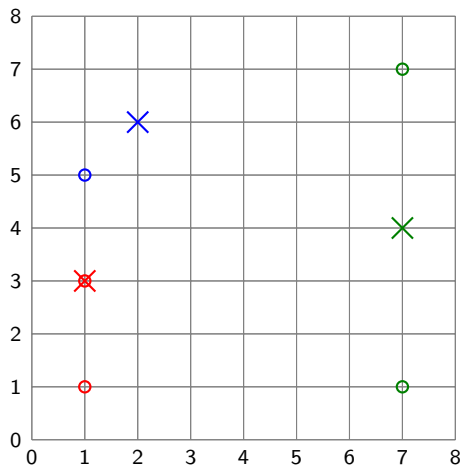
Aufgabe 6



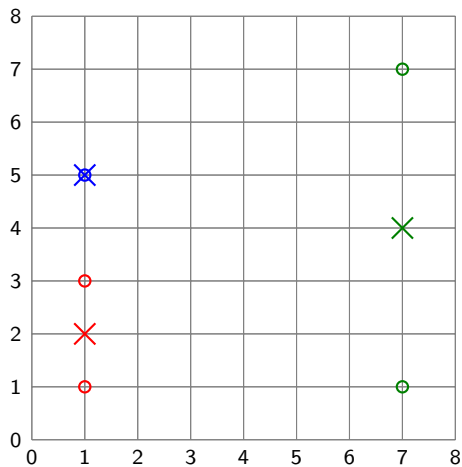
Aufgabe 6 (Schritt 1.1)



Aufgabe 6 (Schritt 1.2)



Aufgabe 6 (Schritt 2.1)



Aufgabe 7

Beachte, dass der k -Means-Algorithmus in den letzten beiden Aufgaben mit den gleichen Daten (Punkten) aber unterschiedlichen Startzentren durchgeführt wurde.

Welche Lösung ist „besser“? Begründe die Antwort quantitativ.

Aufgabe 7

Die Güte des Clusterings kann mit der Summe der Varianzen der einzelnen Cluster (*distortion measure*) gemessen werden.

$$J = \sum_{c_i \in C} \sum_{p_j \in P} |c_i - p_j|^2$$

Aufgabe 5: $J = 0^2 + 0^2 + 2^2 + 2^2 = 8$

Aufgabe 6: $J = 1^2 + 1^2 + 0^2 + 3^2 + 3^2 = 20$

Somit ist das Clustering in Aufgabe 5 besser.