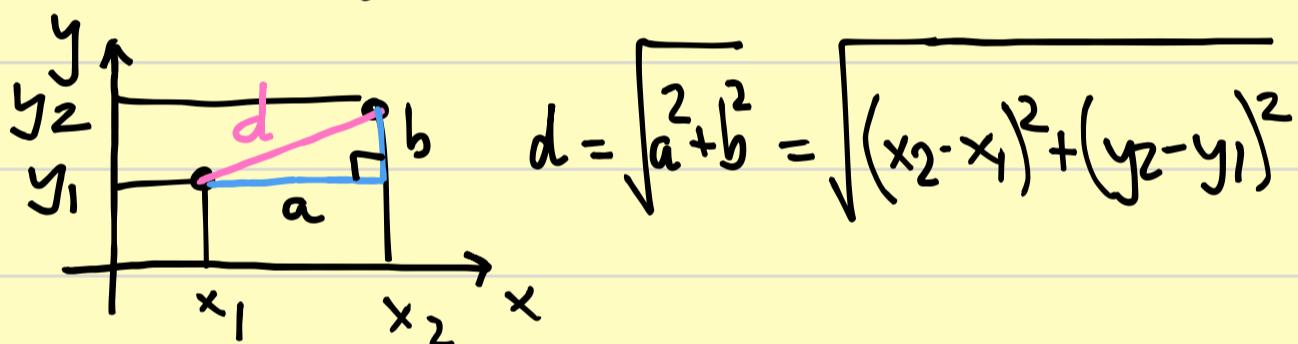


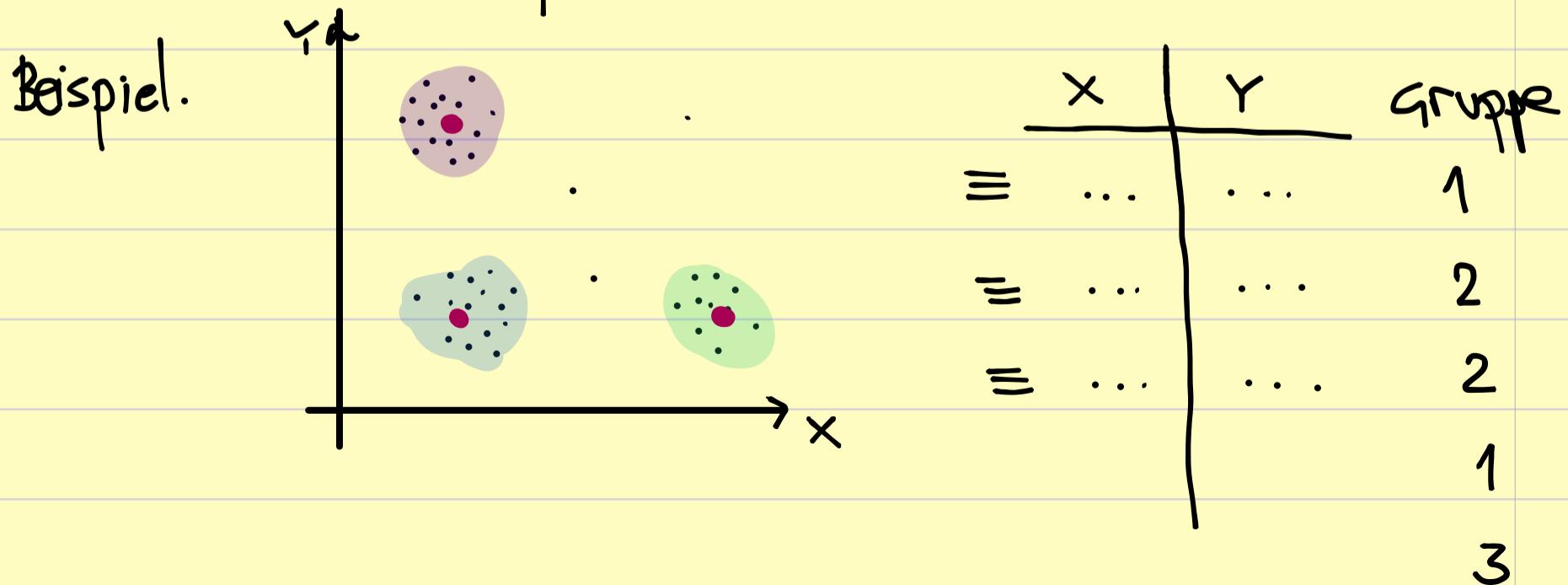
## MASCHINELLES LERNEN

### K-Means Clustering

- Clustering bedeutet „Gruppenbildung“
- Aus euklidischen Daten sind wir in der Lage ähnliche Subgruppen (Cluster) zu bilden.
- „Ähnlich“ bedeutet hier „nah“, in dem von den Daten definierten Raum.



- Die Hypothese um K-Means Clustering anwenden zu können ist, dass der von den Daten definierten Raum eine euklidische Natur hat.



- „K“ sind die Anzahl Gruppen und k-Means Clustering zeigt uns die Position der Punkte mit dem geringsten Abstand zu den Gruppen (Zentroide).

NACHTEIL: wir müssen den Algorithmus sagen, wie viele Gruppen wir haben wollen.

VORTEIL: schnell & effizient.

---

### k-Means Clustering

SCHRITT 0. Entscheidung über Anzahl Clusters ( $k$ )

→ SCHRITT 1. Punkte vom Dataset in  $k$  Gruppen teilen

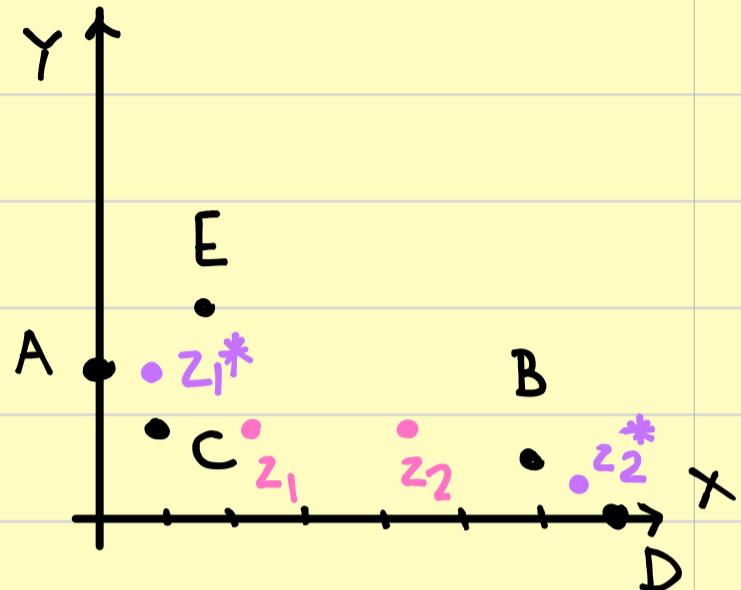
SCHRITT 2. Zentroide (Schwerpunkt) der Gruppen ermitteln.

SCHRITT 3. Abstand von den Punkten zu den Zentroiden.

SCHRITT 4. Clustern nach dem geringsten Abstand und neu bei Schritt 1 anfangen bis Abstand zu den Zentroiden konstant ist.

Beispiel. Gegeben sind die  $(x, y)$  Positionen von 5 Werkten. Bitte ermitteln Sie die optimale Position von 2 Läger, angenommen Alle Werke haben den gleichen Bedarf.  $y \uparrow$

	x	y
A	0	3
B	6	1
C	1	2
D	7	0
E	2	4



☒ Sind die Daten euklidisch? Ja, Abstand kann gemessen werden.

Schritt 0. # Glusters  $\equiv K = 2$  : wir suchen 2 Läger.

Schritt 1. 1.  $\{A, B, C\}$  2.  $\{D, E\}$

Schritt 2.

$$z_1 = \left[ \frac{x_A + x_B + x_C}{3}, \frac{y_A + y_B + y_C}{3} \right] = \left[ \frac{0+6+1}{3}, \frac{3+1+2}{3} \right] =$$

$$z_1 = [2^1 33, 2]$$

$$z_2 = \left[ \frac{x_D + x_E}{2}, \frac{y_D + y_E}{2} \right] = \left[ \frac{7+2}{2}, \frac{0+4}{2} \right] = [4^1 5, 2]$$

Schritt 3.

$$\begin{aligned} d_{A,z_1} &= \sqrt{(0-2^1 33)^2 + (3-2)^2} = 2^1 535 ; \quad d_{A,z_2} = \sqrt{(0-4^1 5)^2 + (3-2)^2} = 4^1 609 \\ d_{B,z_1} &= \sqrt{(6-2^1 33)^2 + (1-2)^2} = 3^1 804 ; \quad d_{B,z_2} = \sqrt{(6-4^1 5)^2 + (1-2)^2} = 1^1 803 \end{aligned}$$

$$\cdot d_{C,z_1} = \sqrt{(1-2)^2 + (3-3)^2} = 1'33; d_{C,z_2} = \sqrt{(-4-5)^2 + (2-2)^2} = 4'5$$

$$\cdot d_{D,z_1} = \sqrt{(7-2)^2 + (3-3)^2} = 5'08; d_{D,z_2} = \sqrt{(7-4)^2 + (5-2)^2} = 3'2$$

$$\cdot d_{E,z_1} = \sqrt{(2-2)^2 + (4-2)^2} = 2'027; d_{E,z_2} = \sqrt{(2-4)^2 + (4-2)^2} = 3'2$$

## Schritt 4. Neue Clusters

$$1^*. \{ A, C, E \} \quad 2^*. \{ B, D \}$$

## Schritt 2. Zentroide

$$z_1^* = \left[ \frac{0+1+2}{3}, \frac{3+2+4}{3} \right] = [1, 3]$$

$$z_2^* = \left[ \frac{6+7}{2}, \frac{1+0}{2} \right] = [6'5, 0'5]$$

## Schritt 3.

$$d_{A,z_1^*} = \sqrt{(0-1)^2 + (3-3)^2} = 1; d_{A,z_2^*} = \sqrt{(0-6'5)^2 + (3-0'5)^2} = 6'9$$

$$d_{B,z_1^*} = \sqrt{(6-1)^2 + (1-3)^2} = 5'38; d_{B,z_2^*} = \sqrt{(6-6'5)^2 + (1-0'5)^2} = 0'707$$

$$d_{C,z_1^*} = \sqrt{(1-1)^2 + (2-3)^2} = 1; d_{C,z_2^*} = \sqrt{(1-6'5)^2 + (2-0'5)^2} = 5'7$$

$$d_{D,z_1^*} = \sqrt{(7-1)^2 + (0-3)^2} = 6'7; d_{D,z_2^*} = \sqrt{(7-6'5)^2 + (0-0'5)^2} = 0'707$$

$$d_{E,z_1^*} = \sqrt{(2-1)^2 + (4-3)^2} = 0'707; d_{E,z_2^*} = \sqrt{(2-6'5)^2 + (4-0'5)^2} = 5'7$$

## Schritt 4. Cluster-Bildung

$$1^*. \{ A, C, E \} \quad 2^*. \{ B, D \} \quad \checkmark$$

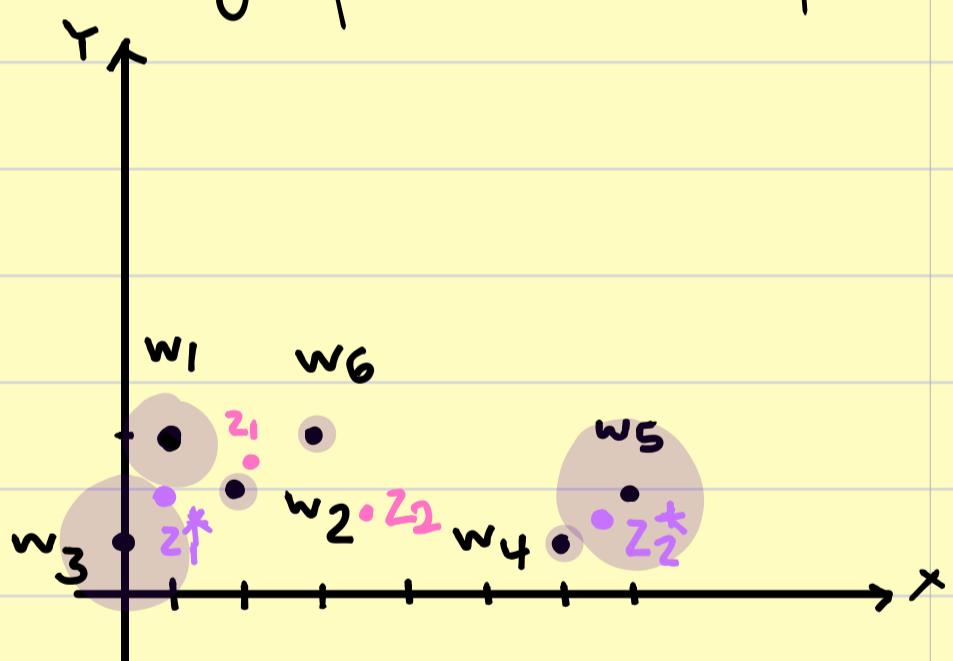
Zentroide: Position der Läger:

$$z_1^* = [1, 3] \quad z_2^* = [6^{\circ}5, 0^{\circ}5] \quad \checkmark$$

**Beispiel.** Die Positionen von 6 Werkstätten mit unterschiedlichen Bedarfen an Rohwaren sind durch ihre Koordinaten auf der Karte bestimmt. Jedes Werk wird von einem der 2 geplanten Lägen beliefert. Um die Fahrtkosten zu minimieren sollten die Läger so positioniert werden, dass sowohl die Werke möglichst nah sind, als auch die Bedarfe berücksichtigt werden. Bitte nutzen Sie einen geeigneten Algorithmus um der Geschäftsführung eine Empfehlung für die Lagerposition auszusprechen.

Daten:

	x	y	z
w <sub>1</sub>	1	3	2
w <sub>2</sub>	2	2	1
w <sub>3</sub>	0	1	3
w <sub>4</sub>	6	1	1
w <sub>5</sub>	7	2	3
w <sub>6</sub>	3	3	1



GRUPPEN: 1. {w<sub>1</sub>, w<sub>2</sub>, w<sub>4</sub>}    2. {w<sub>3</sub>, w<sub>5</sub>, w<sub>6</sub>}

Zentroide:

$$z_1 = \left[ \frac{1 \cdot 2 + 2 \cdot 1 + 6 \cdot 1}{2+1+1}, \frac{3 \cdot 2 + 2 \cdot 1 + 1 \cdot 1}{2+1+1} \right] = [2^{\circ}5, 2^{\circ}5]$$

$$z_2 = \left[ \frac{0 \cdot 3 + 7 \cdot 3 + 3 \cdot 1}{3+3+1}, \frac{1 \cdot 3 + 2 \cdot 3 + 3 \cdot 1}{3+3+1} \right] = [3^{\circ}43, 1^{\circ}714]$$

Abstände:

$$d_{w_1, z_1} = \sqrt{(1-2^15)^2 + (3-2^125)^2} = 1'674; d_{w_1, z_2} = \sqrt{(1-3^143)^2 + (3-1^1714)^2} = 2'74$$

$$d_{w_2, z_1} = \sqrt{(2-2^15)^2 + (2-2^125)^2} = 0'559; d_{w_2, z_2} = \sqrt{(2-3^143)^2 + (2-1^1714)^2} = 1'45$$

$$d_{w_3, z_1} = \sqrt{(0-2^15)^2 + (1-2^125)^2} = 2'79; d_{w_3, z_2} = \sqrt{(0-3^143)^2 + (1-1^1714)^2} = 3'5$$

$$d_{w_4, z_1} = \sqrt{(6-2^15)^2 + (1-2^125)^2} = 3'716; d_{w_4, z_2} = \sqrt{(7-3^143)^2 + (2-1^1714)^2} = 2'67$$

$$d_{w_5, z_1} = \sqrt{(7-2^15)^2 + (2-2^125)^2} = 4'51; d_{w_5, z_2} = \sqrt{(7-3^143)^2 + (2-1^1714)^2} = 3'58$$

$$d_{w_6, z_1} = \sqrt{(3-2^15)^2 + (3-2^125)^2} = 0'1; d_{w_6, z_2} = \sqrt{\frac{(3-3^143)^2 + (3-1^1714)^2}{2+1+3+1}} = 1'33$$

Clusters: 1\*  $\{w_1, w_2, w_3, w_6\}$   
2\*  $\{w_4, w_5\}$

Neuer Zentroide:

$$z_1^* = \left[ \frac{1 \cdot 2 + 2 \cdot 1 + 0 \cdot 3 + 3 \cdot 1}{2+1+3+1}, \frac{3 \cdot 2 + 2 \cdot 1 + 1 \cdot 3 + 3 \cdot 1}{2+1+3+1} \right] =$$

$$z_1^* = [1, 2]$$

$$z_2^* = \left[ \frac{6 \cdot 1 + 7 \cdot 3}{1+3}, \frac{1 \cdot 1 + 2 \cdot 3}{1+3} \right] = [6'75, 1'75]$$

Abstände:

$$d_{w_1, z_1^*} = 1 < d_{w_1, z_2^*} \quad d_{w_4, z_1^*} = 5'1 > d_{w_4, z_2^*} = 1'06$$

$$d_{w_2, z_1^*} = 1 < d_{w_2, z_2^*} \quad d_{w_5, z_1^*} = 6 > d_{w_5, z_2^*} = 0'3$$

$$d_{w_3, z_1^*} = 1 < d_{w_3, z_2^*} \quad d_{w_6, z_1^*} = 2'24 < d_{w_6, z_2^*}$$

Die Gruppen ändern sich also nicht:

$$z_1^* = [1, 2] ; z_2^* = [6'75, 1'75]$$

$$\text{Gruppen: } \begin{array}{l} 1^* \left\{ w_1, w_2, w_3, w_6 \right\} \\ 2^* \left\{ w_4, w_5 \right\} \end{array}$$

Übung. Gegeben werden 3 Kennzahlen zur Beschreibung von 2 Kundengruppen: Umsatz, Häufigkeit, # Reklamationen.

- Bitte clustern Sie die Kundendaten und ermitteln Sie die Zentroide.
- Welche Interpretation haben in dem Fall die Zentroide?

	K1	K2	K3	K4	K5	K6	K7
Umsatz	300	500	450	360	110	90	70
Häufigkeit	6	7	5	4	1	2	1
# Rekla	10	20	11	22	7	13	2

# 1. SCHRITT NORMIEREN!

$$x_i^* = \frac{x_i - \mu_x}{\sigma_x}$$

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

Umsatz\*  $\frac{300-70}{500-70}$   $\frac{500-70}{500-70}$   $\frac{450-70}{500-70}$  ...  $[0,1]$

Häufigkeit\*  $\frac{6-1}{7-1}$   $\frac{7-1}{7-1}$   $\frac{5-1}{7-1}$  ...  $[0,1]$

#Rekl.  $\frac{10-2}{22-2}$   $\frac{20-2}{22-2}$   $\frac{11-2}{22-2}$  ...  $[0,1]$

