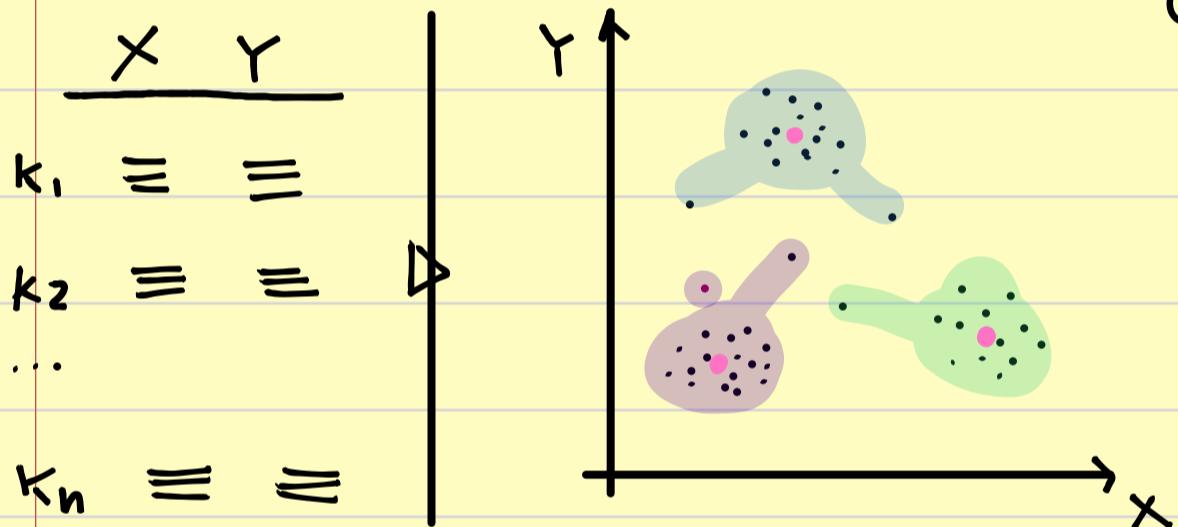


## MASCHINELLES LERNEN

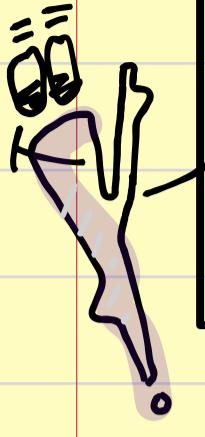
### k-Means Clustering

**Beispiel.** Ein Dienstleister der Paketzustellung möchte, anhand der Kundenadressen, die Kunden in Gebieten für Ihre Mitarbeiter (3) teilen. Dafür würden die  $x, y$  Koordinaten der Kunden zusammengetragen.



- Clustering bedeutet „Gruppenbildung“
- Aus euklidischen Daten (wir können einen Abstand messen) sind wir in der Lage ähnliche Subgruppen (Clusters) zu bilden.
- „Ähnlich“ bedeutet in dem Kontext „nah“, in dem von den Daten definierten Raum.

$$c = \sqrt{a^2 + b^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



• Die Hypothese um K-Means Clustering anzuwenden ist, dass der von den Daten definierten Raum eine euklidische Natur hat (wir können einen Abstand messen).

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

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

VORTEIL : schnell & effizient.

## K-Means Clustering Algorithmus.

Schritt 0 . Entscheidung über Anzahl Clusters (k)

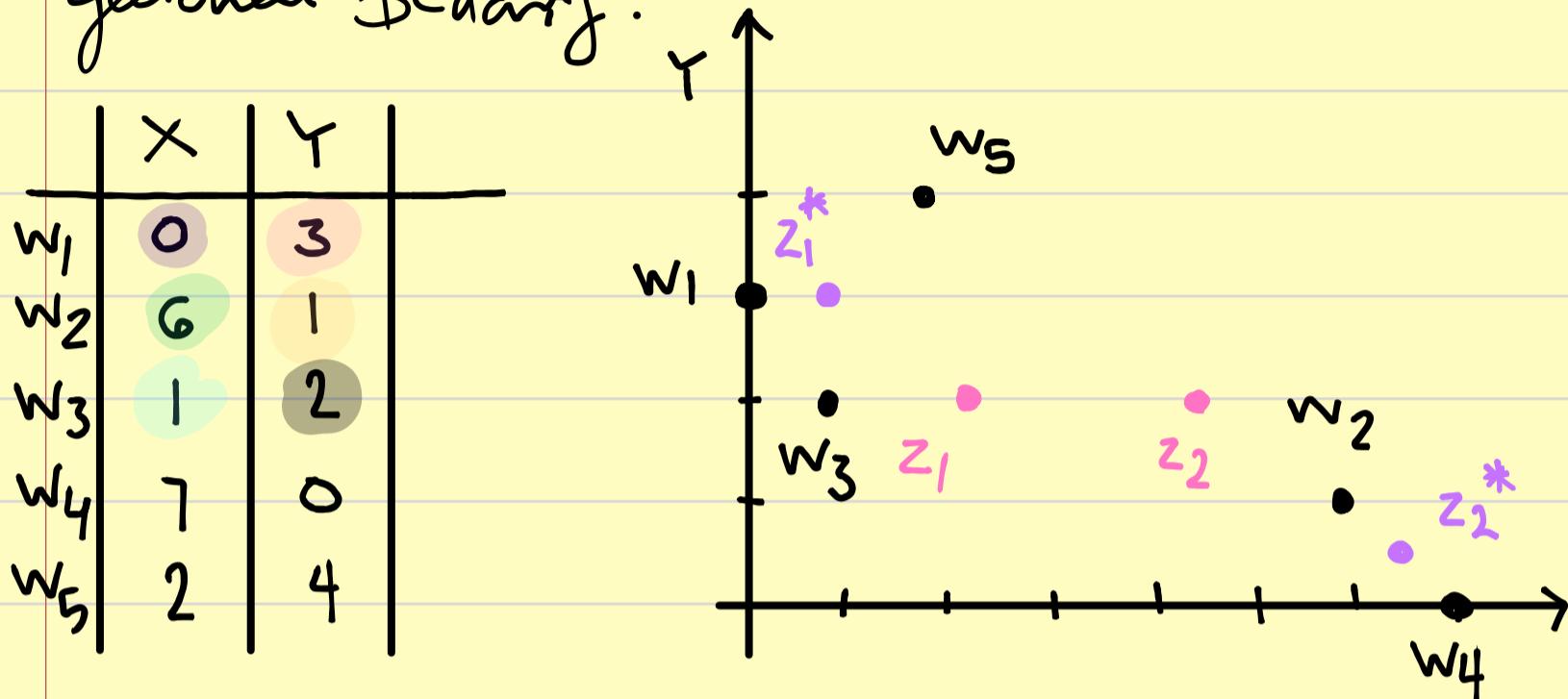
→ Schritt 1 . Punkte vom Dataset in k Gruppen teilen.

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 der Abstand zu den Zentroiden konstant ist.

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



Sind die Daten Euklidisch?

Ja, Abstand kann gemessen werden.

Schritt 0. #Clusters.  $k=2 \equiv$  wir suchen 2 Cluster!

Schritt 1. 1.  $\{w_1, w_2, w_3\}$  2.  $\{w_4, w_5\}$

Schritt 2.

$$z_1 = \left[ \frac{0+6+1}{3}, \frac{3+1+2}{3} \right] = [2^1 \dot{3}3, 2]$$

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

Schritt 3.

$$d_{w_1, z_1} = \sqrt{(0-2^1 \dot{3}3)^2 + (3-2)^2} = 2^1 \dot{5}35; d_{w_1, z_2} = \sqrt{(0-4^1 \dot{5})^2 + (3-2)^2} = 4^1 \dot{6}1$$

$$d_{w_2, z_1} = \sqrt{(6-2^1 \dot{3}3)^2 + (1-2)^2} = 3^1 \dot{8}04; d_{w_2, z_2} = \sqrt{(6-4^1 \dot{5})^2 + (1-2)^2} = 1^1 \dot{8}03$$

$$d_{w_3, z_1} = \sqrt{(1-2^1 \dot{3}3)^2 + (2-2)^2} = 1^1 \dot{3}3; d_{w_3, z_2} = \sqrt{(1-4^1 \dot{5})^2 + (2-2)^2} = 4^1 \dot{5}$$

$$d_{w4,z_1} = \sqrt{(7-2^133)^2 + (0-2)^2} = 5^108; d_{w4,z_2} = \sqrt{(7-4^15)^2 + (0-2)^2} = 3^12$$

$$d_{w5,z_1} = \sqrt{(2-2^133)^2 + (4-2)^2} = 2^1027; d_{w5,z_2} = \sqrt{(2-4^15)^2 + (4-2)^2} = 3^12$$

#### Schritt 4. Neue Clusters

$$1^*. \{w_1, w_3, w_5\} \quad 2^*. \{w_2, w_4\}$$

#### Schritt 2. Zentroide

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

$$z_2^* = \left[ \frac{6+7}{2}, \frac{1+0}{2} \right] = [6^15, 0^15]$$

#### Schritt 3. Abstände.

$$d_{w1,z_1^*} = \sqrt{(0-1)^2 + (3-3)^2} = 1; d_{w1,z_2^*} = \sqrt{(0-6^15)^2 + (3-0^15)^2} = 6^19$$

$$d_{w2,z_1^*} = \sqrt{(6-1)^2 + (1-3)^2} = 5^138; d_{w2,z_2^*} = \sqrt{(6-6^15)^2 + (1-0^15)^2} = 0^1707$$

$$d_{w3,z_1^*} = \sqrt{(1-1)^2 + (2-3)^2} = 1; d_{w3,z_2^*} = \sqrt{(1-6^15)^2 + (2-0^15)^2} = 5^17$$

$$d_{w4,z_1^*} = \sqrt{(7-1)^2 + (0-3)^2} = 6^17; d_{w4,z_2^*} = \sqrt{(7-6^15)^2 + (0-0^15)^2} = 0^1707$$

$$d_{w5,z_1^*} = \sqrt{(2-1)^2 + (4-3)^2} = 0^1707; d_{w5,z_2^*} = \sqrt{(2-6^15)^2 + (4-0^15)^2} = 5^17$$

$$\text{Schritt 4. } 1^{**} = 1^* = \{w_1, w_3, w_5\}$$

$$2^{**} = z^* = \{w_2, w_4\}$$



$$\text{Lagerposition: } z_1^* = [1, 3] \quad z_2^* = [6^15, 0^15] \quad \checkmark$$

Beispiel. Die Positionen von 6 Werkten 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 Lägen so positioniert werden, dass sowohl die Werke möglichst nah sind, als auch die Bedarfe berücksichtigt werden. Bitte nutzen Sie einen geeigneten Algorithmus und der GF eine Empfehlung für die Lagerposition auszusprechen.

Daten:

	X	Y	B
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

$$k = 2$$

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:

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

$$z_1 = [2'5, 2'25]$$

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

Abstände:

$$d_{w_1, z_1} = \sqrt{(1-2'5)^2 + (3-2'25)^2} = 1'674; d_{w_1, z_2} = \dots = 2'74$$

$$d_{w_2, z_1} = \dots = 0'559; d_{w_2, z_2} = \dots = 1'45$$

$$d_{w_3, z_1} = \dots = 2'79; d_{w_3, z_2} = \dots = 3'5$$

$$d_{w_4, z_1} = \dots = 3'716; d_{w_4, z_2} = \dots = 2'67$$

$$d_{w_5, z_1} = \dots = 4'51 ; d_{w_5, z_2} = \dots = 3'58$$

$$d_{w_6, z_1} = \dots = 0'9 ; d_{w_6, z_2} = \dots = 1'33$$

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

Neue 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:

$$d_{w_1, z_1^*} = \dots = 1 < d_{w_1, z_2^*}$$

$$d_{w_2, z_1^*} = \dots = 1 < d_{w_2, z_2^*}$$

$$d_{w_3, z_1^*} = \dots = 1 < d_{w_3, z_2^*}$$

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

$$d_{w_5, z_1^*} = \dots = 6 > d_{w_5, z_2^*} = 0'3$$

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

Die Gruppen ändern sich nicht:

$$1^{**} = 1^* \rightarrow z_1^* = [1, 2]$$

$$2^{**} = 2^* \rightarrow z_2^* = [6'75, 1'75]$$

$$1^* = \{ w_1, w_2, w_3, w_6 \}$$

$$2^* = \{ w_4, w_5 \}$$

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

- Bitte Clustern Sie die Kunden in 2 Gruppen und ermitteln Sie die Zentroide.
- Welche Management Interpretation haben die Zentroide in dem Fall?

Daten:

	K1	K2	K3	K4	K5	K6	K7
Umsatz	300	500	450	360	110	90	70
U-Häufigkeit	60	70	50	40	18	20	10
# Rekla	10	6	11	8	2	7	3

Daten von Kennzahlensystemen müssen zunächst normiert werden !!

$$1. \ x_i^* = \frac{x_i - \mu_x}{\sigma_x}$$

$$2. \ x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

Beispiel Umsatz:

$$1. \ \mu_x = \frac{\sum x_i}{n} = \frac{300+500+450+360+110+90+70}{7} = 268'57$$

$$\sigma_x = \sqrt{\frac{\sum (x_i - \mu_x)^2}{n-1}} = \sqrt{\frac{(300-268'57)^2 + (500-268'57)^2 + \dots + (70-268'57)^2}{7-1}}$$

$$\text{Umsatz}^* \quad \frac{300-268'57}{\sigma_x} \quad \frac{500-268'57}{\sigma_x} \quad \dots \quad \frac{70-268'57}{\sigma_x}$$

	u1	u2	u3	u4	u5	u6	u7
2. Umsatz	300	500	450	360	110	90	70

Umsatz\*\*  $\frac{300-70}{500-70}$   $\frac{500-70}{500-70}$   $\frac{450-70}{500-70}$   $\frac{360-70}{500-70}$   $\frac{110-70}{500-70}$   $\frac{90-70}{500-70}$   $\frac{70-70}{500-70}$

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

2. k-Means-Cluster mit den normierten Zahlen!

