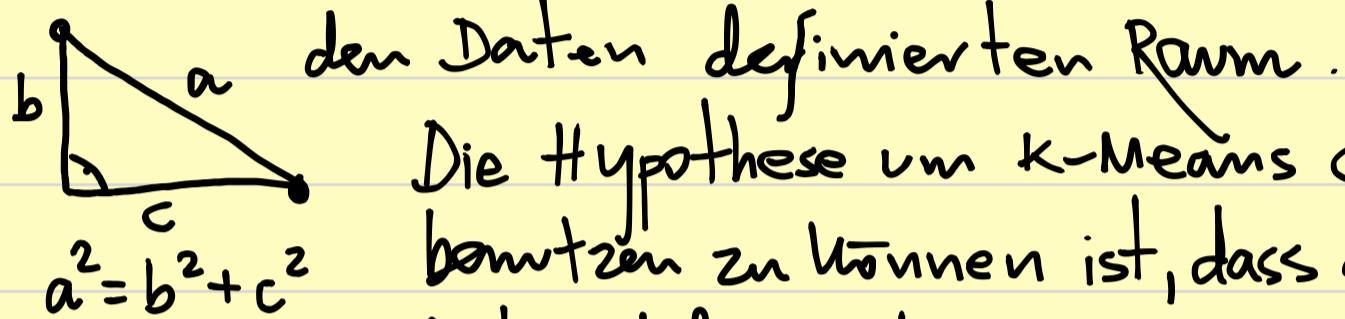


k-Means Clustering

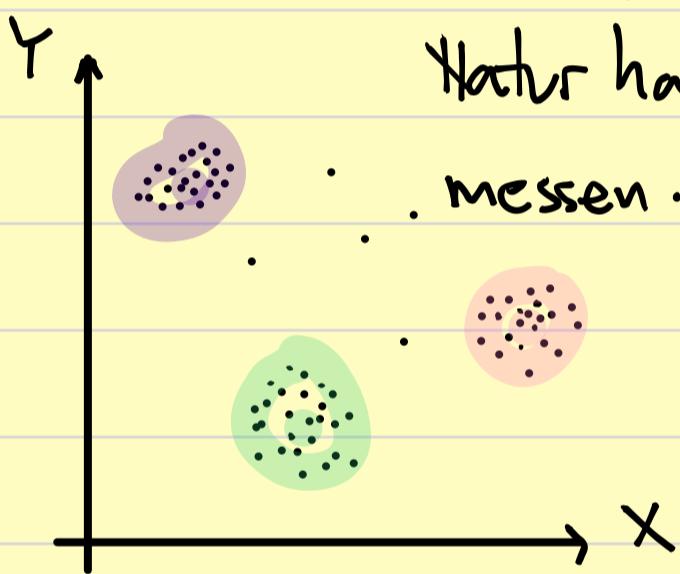
.. CLUSTERING .. bedeutet Gruppenbildung.

Aus Daten sind wir in der Lage **ähnliche** Subgruppen zu bilden. .. Ähnlich bedeutet hier ..nah“ in dem von



den Daten definierten Raum.

Die Hypothese um k-Means Clustering bewezen zu können ist, dass der von den Daten definierten Raum eine euklidische Natur hat. D.h. wir können einen Abstand messen.



Beispiel Amazon Lager positionierung

.. 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.

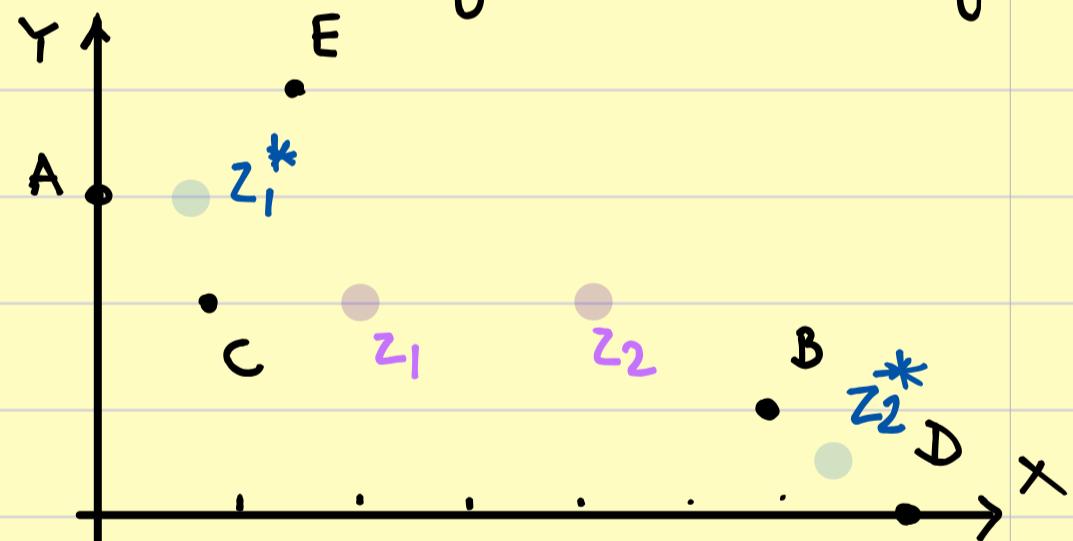
K-MEANS CLUSTERING ALGORITHMUS

SCHRITT 0. Entscheidung über Anzahl cluster: k

- SCHRITT 1. Punkte vom Dataset in k Gruppen teilen
- SCHRITT 2. Zentroid (Schwerpunkt) der Gruppen ermitteln
- SCHRITT 3. Abstand von den Punkten zu den Zentoiden.
- SCHRITT 4. Clustern nach geringsten Abstand und neu bei Schritt 1 anfangen bis Abstand zu den Zentoiden 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.

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



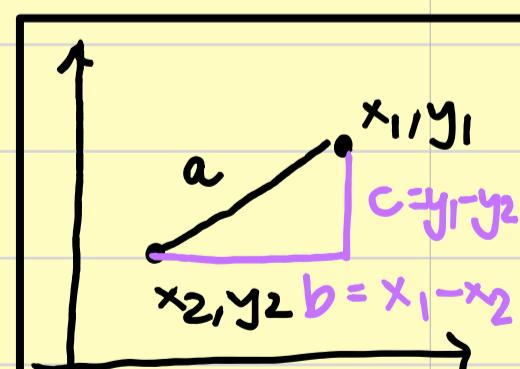
Schritt 0. #clusters . $k = 2$

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

Schritt 2.

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

$$a = \sqrt{b^2 + c^2}$$



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

Schritt 3.

$$d(A, z_1) = \sqrt{(0-2|33)^2 + (3-2)^2} = 2|535; d(A, z_2) = \sqrt{(0-4|5)^2 + (3-2)^2} = 4|609$$

$$d(B, z_1) = \sqrt{(6-2|33)^2 + (1-2)^2} = 3|804; d(B, z_2) = \sqrt{(6-4|5)^2 + (1-2)^2} = 1|803$$

$$d(C, z_1) = \sqrt{(1-2|33)^2 + (2-2)^2} = 1|33; d(C, z_2) = \sqrt{(1-4|5)^2 + (2-2)^2} = 4|5$$

$$d(D, z_1) = \sqrt{(7-2|33)^2 + (0-2)^2} = 5|08; d(D, z_2) = \sqrt{(7-4|5)^2 + (0-2)^2} = 3|2$$

$$d(E, z_1) = \sqrt{(2-2|33)^2 + (4-2)^2} = 2|027; d(E, z_2) = \sqrt{(2-4|5)^2 + (4-2)^2} = 3|2$$

Schritt 4. Neue Clustering:

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

Schritt 2. Neue 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.

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

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

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

$$d(D, z_1^*) = \sqrt{(7-1)^2 + (0-3)^2} = 6|71; d(D, z_2^*) = \sqrt{(7-6|5)^2 + (0-0|5)^2} = 0|7$$

$$d(E_1 z_1^*) = \sqrt{(2-1)^2 + (4-3)^2} = 1'707 ; d(E_1 z_2^*) = \sqrt{(2-6,5)^2 + (4-0,5)^2} = 5'7$$

Schritt 4. clustering bleibt gleich

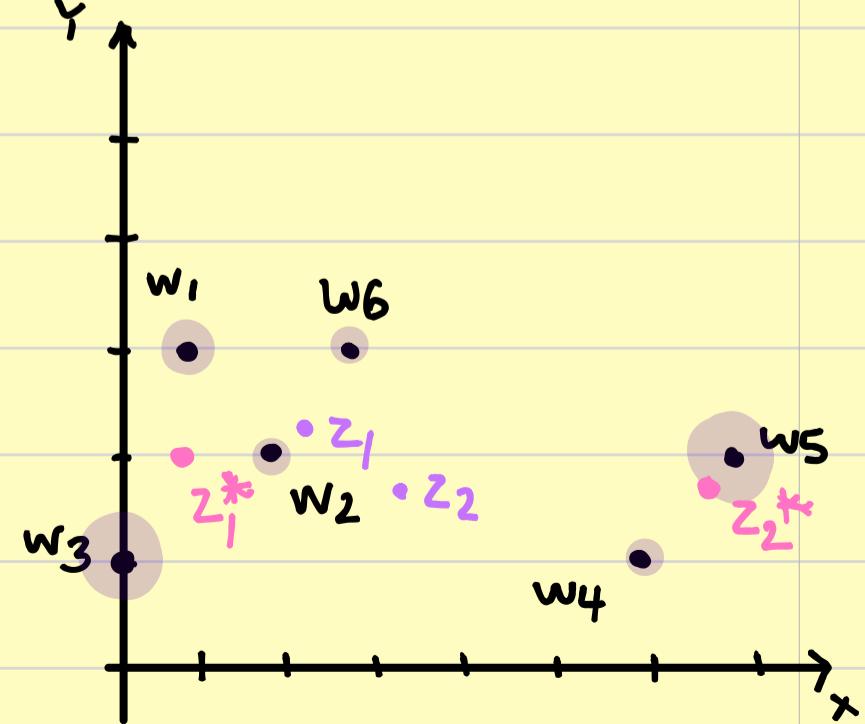
Gruppen: 1. {A,C,E} 2. {B,D}

Zentroide: $z_1^*[1,3]$ $z_2^*[6,5,0,5]$

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 Läger 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 Lagerpositionierung auszusprechen.

DATEN.	X	Y	B
w ₁	1	3	2
w ₂	2	2	1
w ₃	0	1	3
w ₄	6	1	1
w ₅	7	2	3



$w_6 \quad 3 \quad | \quad 3 \quad | \quad 1$

GRUPPEN: 1. $\{w_1, w_2, w_4\}$ 2. $\{w_3, w_5, w_6\}$

ZENTROIDE GEWICHTET:

$$z_1 = \left[\frac{1 \cdot 2 + 2 \cdot 1 + 6 \cdot 1}{2+1+1}, \frac{3 \cdot 2 + 2 \cdot 1 + 1 \cdot 1}{2+1+1} \right] = [2'5, 2'25]$$

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

Abstände:

$$d(w_1, z_1) = \sqrt{(1-2'5)^2 + (3-2'25)^2} = 1'6744; d(w_1, z_2) = \sqrt{(1-3'43)^2 + (3-1'714)^2} = 2'74$$

$$d(w_2, z_1) = \sqrt{(2-2'5)^2 + (2-2'25)^2} = 0'559; d(w_2, z_2) = \sqrt{(2-3'43)^2 + (2-1'714)^2} = 1'45$$

$$d(w_3, z_1) = \sqrt{(0-2'5)^2 + (1-2'25)^2} = 2'795; d(w_3, z_2) = \sqrt{(0-3'43)^2 + (1-1'714)^2} = 3'5$$

$$d(w_4, z_1) = \sqrt{(6-2'5)^2 + (1-2'25)^2} = 3'716; d(w_4, z_2) = \sqrt{(6-3'43)^2 + (1-1'714)^2} = 2'67$$

$$d(w_5, z_1) = \sqrt{(7-2'5)^2 + (2-2'25)^2} = 4'51; d(w_5, z_2) = \sqrt{(7-3'43)^2 + (2-1'714)^2} = 3'58$$

$$d(w_6, z_1) = \sqrt{(3-2'5)^2 + (3-2'25)^2} = 0'901; d(w_6, z_2) = \sqrt{(3-3'43)^2 + (3-1'714)^2} = 1'33$$

Gruppen: 1. $\{w_1, w_2, w_3, w_6\}$ 2. $\{w_4, w_5\}$

Neue Zentroide:

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

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

Abstände:

$$d(w_1, z_1^*) = \sqrt{(1-1)^2 + (3-2)^2} = 1 < d(w_1, z_2^*) = \sqrt{(1-6^{'}75)^2 + (3-1^{'}75)^2}$$
$$d(w_2, z_1^*) = \sqrt{(2-1)^2 + (2-2)^2} = 1 < d(w_2, z_2^*) = \sqrt{(2-6^{'}75)^2 + (2-1^{'}75)^2}$$
$$d(w_3, z_1^*) = \sqrt{(0-1)^2 + (1-2)^2} = 1 < d(w_3, z_2^*) = \sqrt{(0-6^{'}75)^2 + (1-1^{'}75)^2}$$
$$d(w_4, z_1^*) = \sqrt{(6-1)^2 + (1-2)^2} = 5'1 > d(w_4, z_2^*) = \sqrt{(6-6^{'}75)^2 + (1-1^{'}75)^2} = 1'06$$
$$d(w_5, z_1^*) = \sqrt{(7-1)^2 + (2-2)^2} = 6 > d(w_5, z_2^*) = \sqrt{(7-6^{'}75)^2 + (2-1^{'}75)^2}$$
$$d(w_6, z_1^*) = \sqrt{(3-1)^2 + (3-2)^2} = 2'24 < d(w_6, z_2^*) = \sqrt{(3-6^{'}75)^2 + (3-1^{'}75)^2}$$

die Gruppen ändern sich nicht

$$z_1^* = [1, 2] \quad z_2^* = [6^{'}75, 1^{'}75]$$

Gruppen: 1. $\{w_1, w_2, w_3, w_6\}$ 2. $\{w_4, w_5\}$

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

Bitte clustern Sie die Daten in 2 Gruppen und ermitteln Sie die Zentroide der Gruppen.

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 muss normieren sein!

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

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

z.B: Umsatz* = $\begin{cases} \frac{300-70}{500-70}, & \frac{500-70}{500-70}, \\ \frac{450-70}{500-70}, & \frac{360-70}{500-70}, \\ \frac{110-70}{500-70}, & \frac{90-70}{500-70}, \\ \frac{70-70}{500-70} \end{cases}$

zw [0,1]

