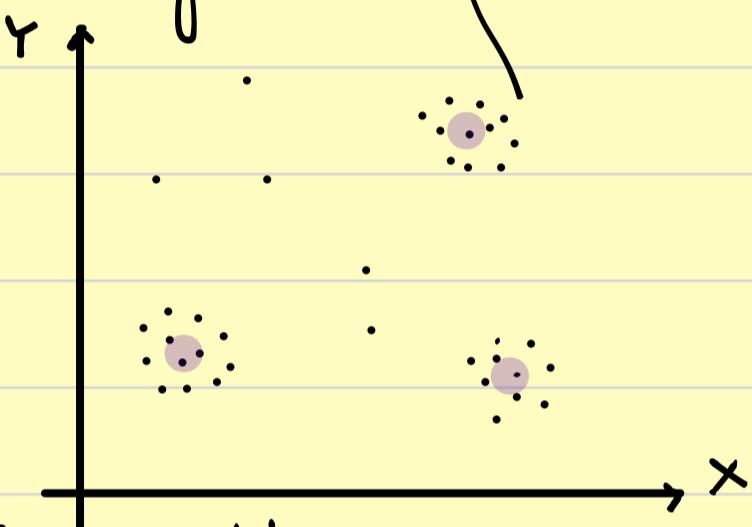


K-Means Clustering

Clustering bedeutet Gruppen bilden: aus Daten sind wir in der Lage ähnliche Subgruppen zu bilden. „Ähnlich“ in diesem Kontext bedeutet „nah“ in dem von den Daten definierten Raum.

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

Beispiel:



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

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

VORTEIL: schnell und effizient.

K-MEANS CLUSTER ALGORITHMUS

SCHRITT 0. ENTSCHEIDUNG ÜBER ANZAHL CLUSTERS: $\dots k \dots$

→ SCHRITT 1. PUNKTE VOM DATASET IN k GRUPPEN TEILEN.

SCHRITT 2. ZENTROID (SCHWERPUNKT) DER GRUPPEN ERMITTLEN.

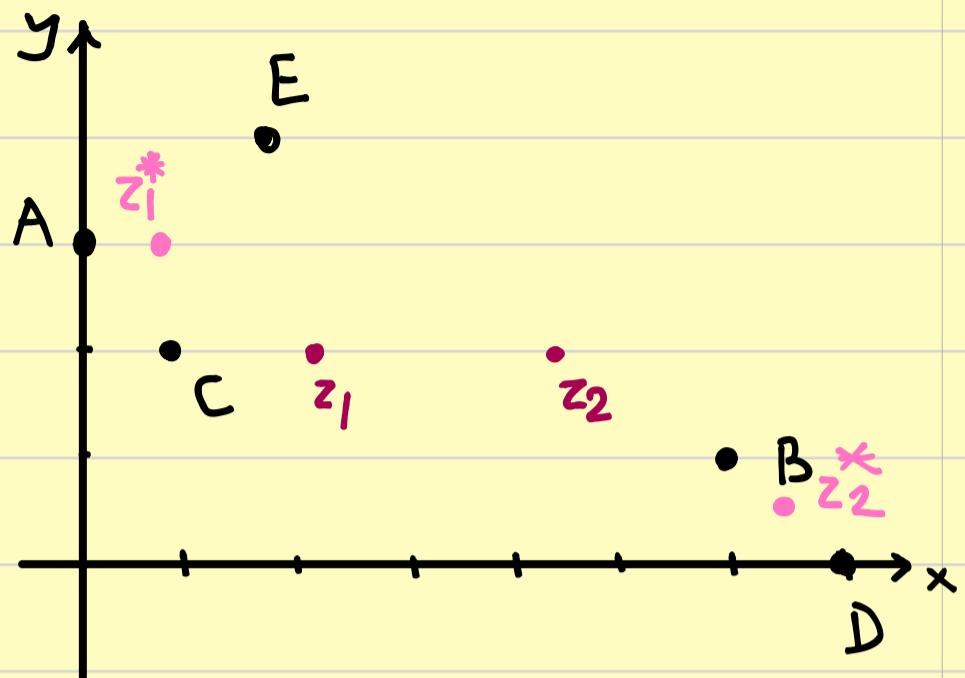
SCHRITT 3. ABSTAND VON DEN PUNKTEN ZUM ZENTROID.

SCHRITT 4. CLUSTERN NACH GERINGSTEN ABSTAND UND NEU

BEI SCHRITT 1 ANFÄNGEN bis ABSTAND zu den ZENTROIDEN KONSTANT ist.

Beispiel.

	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.

zentroide

$$z_1 = \left[\frac{0+6+1}{3}, \frac{3+1+2}{3} \right] = [2.33, 2]$$

$$z_2 = \left[\frac{7+2}{2}, \frac{0+4}{2} \right] = [4.5, 2]$$

SCHRITT 3.

$$d(A, z_1) = \sqrt{(0-2.33)^2 + (3-2)^2} = 2'535 ; d(A, z_2) = \sqrt{(0-4.5)^2 + (3-2)^2} = 4'609$$

$$d(B, z_1) = \sqrt{(6-2.33)^2 + (1-2)^2} = 3'804 ; d(B, z_2) = \sqrt{(6-4.5)^2 + (1-2)^2} = 1'803$$

$$d(C, z_1) = \sqrt{(1-2.33)^2 + (2-2)^2} = 1'33 ; d(C, z_2) = \sqrt{(1-4.5)^2 + (2-2)^2} = 4'5$$

$$d(D, z_1) = \sqrt{(7-2.33)^2 + (0-2)^2} = 5'080 ; d(D, z_2) = \sqrt{(7-4.5)^2 + (0-2)^2} = 3'201$$

$$d(E, z_1) = \sqrt{(2-2.33)^2 + (4-2)^2} = 2'027 ; d(E, z_2) = \sqrt{(2-4.5)^2 + (4-2)^2} = 3'201$$

SCHRITT 4.

Neue Gruppen: $1 \{ A, C, E \} \quad 2 \{ B, D \}$

SCHRITT 2.

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

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

SCHRITT 3.

$$d(A, z_1^*) = \sqrt{(0-1)^2 + (3-3)^2} = 1 ; \quad d(A, z_2^*) = \sqrt{(0-6'5)^2 + (3-0'5)^2} = 6'964$$

$$d(B, z_1^*) = \sqrt{(6-1)^2 + (1-3)^2} = 5'385 ; \quad d(B, z_2^*) = \sqrt{(6-6'5)^2 + (1-0'5)^2} = 0'707$$

$$d(C, z_1^*) = \sqrt{(1-1)^2 + (2-3)^2} = 1 ; \quad d(C, z_2^*) = \sqrt{(1-6'5)^2 + (2-0'5)^2} = 5'701$$

$$d(D, z_1^*) = \sqrt{(7-1)^2 + (0-3)^2} = 6'708 ; \quad d(D, z_2^*) = \sqrt{(7-6'5)^2 + (0-0'5)^2} = 0'707$$

$$d(E, z_1^*) = \sqrt{(2-1)^2 + (4-3)^2} = 0'707 ; \quad d(E, z_2^*) = \sqrt{(2-6'5)^2 + (4-0'5)^2} = 5'701$$

SCHRITT 4.

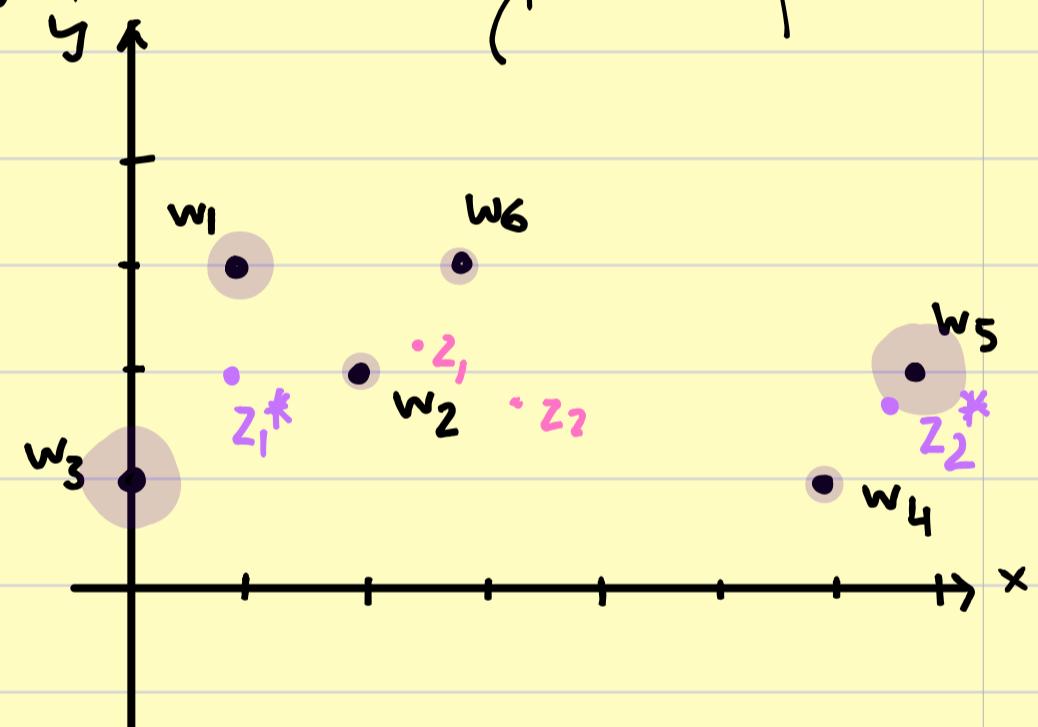
CLUSTERUNG bleibt gleich.

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

ZENTROIDE: $z_1^* = [1, 3] ; \quad z_2^* = [6'5, 0'5]$

Beispiel . Die Positionen von 7 Werkten mit unterschiedlichen Bedarfen an Rohware sind durch ihre Koordinaten auf der Karte bestimmt. Jedes Werk wird von einem der zwei 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	3	3	1



GRUPPEN : 1 { w₁, w₂, w₄ } 2 { w₃, w₅, w₆ }

ZENTROIDE :

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

$$z_2 = \left[\frac{0 \cdot 3 + 7 \cdot 3 + 3 \cdot 1}{3+3+1}, \frac{1 \cdot 3 + 2 \cdot 3 + 3 \cdot 1}{3+3+1} \right] = [3\frac{1}{4}, 1\frac{1}{4}]$$

ABSTÄNDE:

$$d(w_1, z_1) = \sqrt{(1-2'5)^2 + (3-2'25)^2} = 1'677$$

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

$$d(w_2, z_1) = \sqrt{(2-2'5)^2 + (2-2'25)^2} = 0'559$$

$$d(w_2, z_2) = \sqrt{(2-3'43)^2 + (2-1'714)^2} = 1'453$$

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

$$d(w_3, z_2) = \sqrt{(0-3'43)^2 + (1-1'714)^2} = 3'51$$

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

$$d(w_4, z_2) = \sqrt{(6-3'43)^2 + (1-1'714)^2} = 2'674$$

$$d(w_5, z_1) = \sqrt{(7-2'5)^2 + (2-2'25)^2} = 4'51$$

$$d(w_5, z_2) = \sqrt{(7-3'43)^2 + (2-1'714)^2} = 3'58$$

$$d(w_6, z_1) = \sqrt{(3-2'5)^2 + (3-2'25)^2} = 0'901$$

$$d(w_6, z_2) = \sqrt{(3-3'43)^2 + (3-1'714)^2} = 1'33$$

GRUPPEN : 1 { w₁, w₂, w₃, w₆ } 2 { w₄, w₅ }

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:

$$d(w_1, z_1^*) = \sqrt{(1-1)^2 + (3-2)^2} = 1 < d(w_1, z_2^*) = \sqrt{(1-6'75)^2 + (3-1'75)^2}$$

$$d(w_2, z_1^*) = \sqrt{(2-1)^2 + (2-2)^2} = 1 < d(w_2, z_2^*) = \sqrt{(2-6'75)^2 + (2-1'75)^2}$$

$$d(w_3, z_1^*) = \sqrt{(0-1)^2 + (1-2)^2} = 1 < d(w_3, z_2^*) = \sqrt{(0-6'75)^2 + (1-1'75)^2}$$

$$d(w_4, z_1^*) = \sqrt{(6-1)^2 + (1-2)^2} = 5'1 > d(w_4, z_2^*) = \sqrt{(6-6'75)^2 + (1-1'75)^2} = 1'06$$

$$d(w_5, z_1^*) = \sqrt{(7-1)^2 + (2-2)^2} = 6 > d(w_5, z_2^*) = \sqrt{(7-6^{175})^2 + (2-1^{175})^2}$$

$$d(w_6, z_1^*) = \sqrt{(3-1)^2 + (3-2)^2} = 2^{124} < d(w_6, z_2^*) = \sqrt{(3-6^{175})^2 + (3-1^{175})^2} =$$

die Gruppen ändern sich nicht

$$z_1^* = [1, 2] \quad z_2^* = [6^{175}, 1^{175}]$$

Gruppen: 1 {w₁, w₂, w₃, w₆} 2 {w₄, w₅}

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	Häufigkeit	Reklamationen
300	6	10
500	7	20
450	5	11
360	4	22
110	1	7
90	2	13
70	1	2
€	Wie oft im Laden	

Hinweis: Alle Daten zw. $[0,1]$ normieren

Säule für Säule!

$$x^* = \frac{x_i - \bar{x}}{\sigma_x}$$

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