Data Mining Tutorial

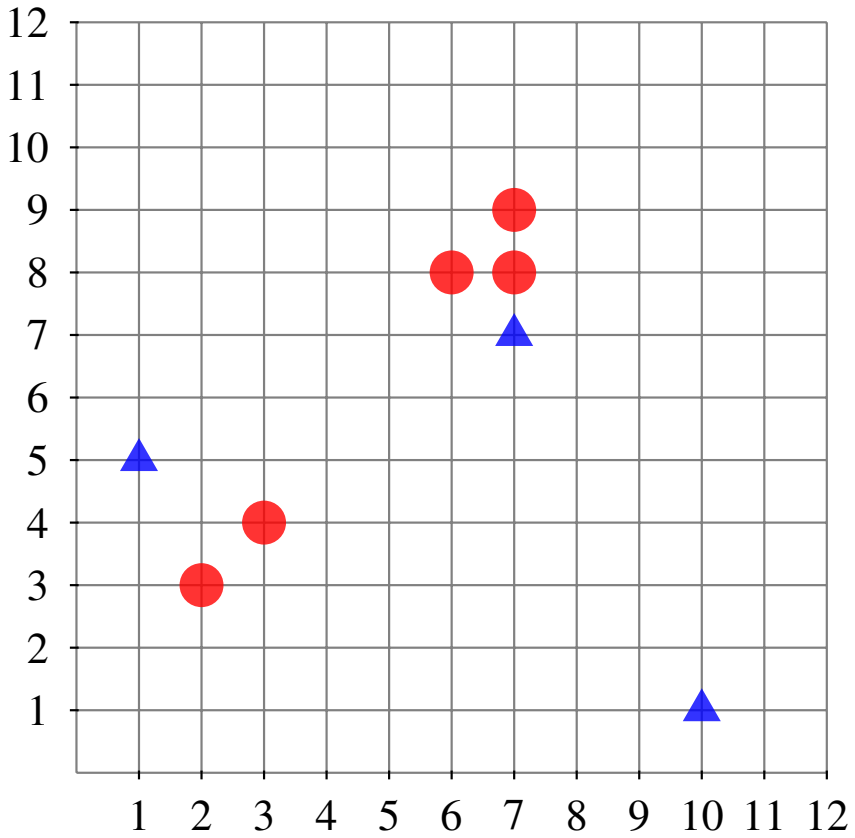
Clusteranalyse – Teil I

Erich Schubert, Arthur Zimek

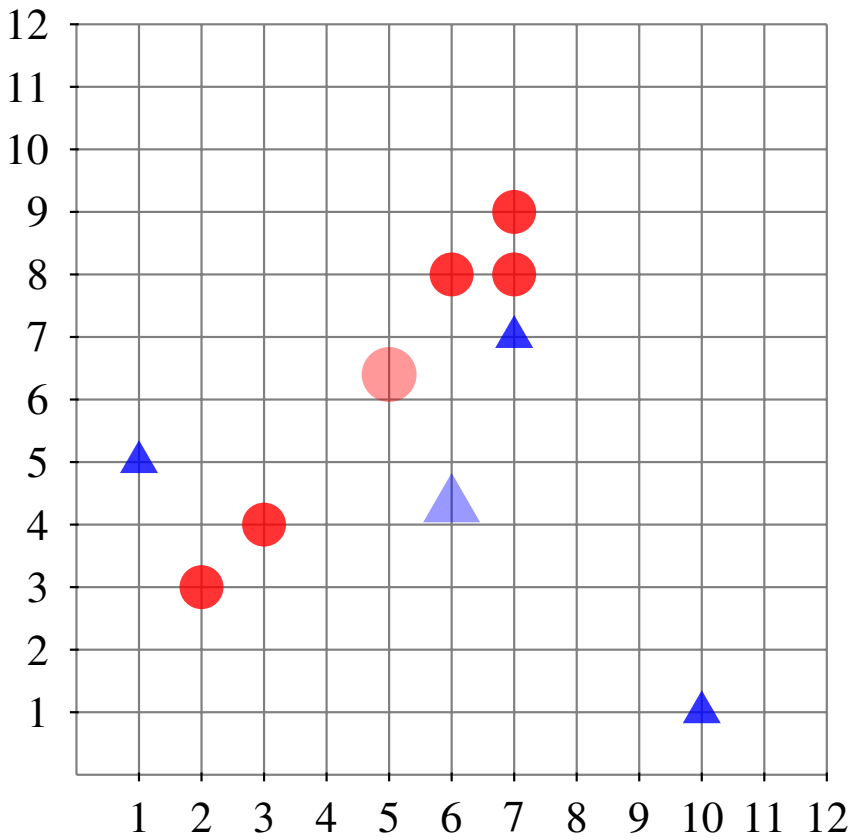
Ludwig-Maximilians-Universität München

14. und 15.12.2017 — KDD Übung

k-Means Clustering – Lloyd/Forgy Algorithmus



k-Means Clustering – Lloyd/Forgy Algorithmus

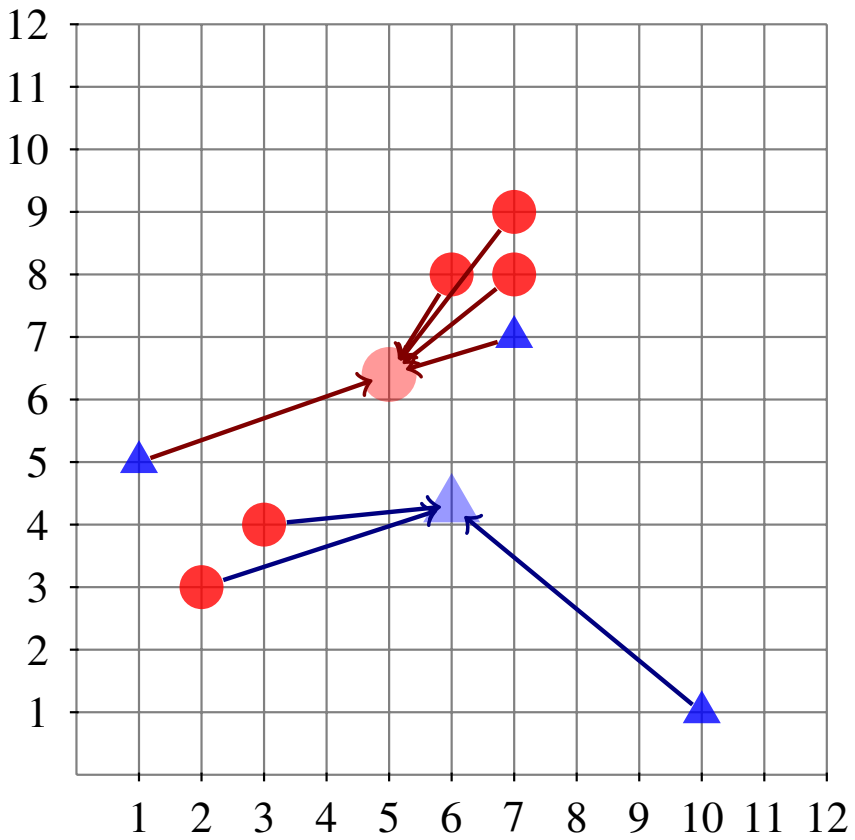


Zentroide neu berechnen:

$$\mu \approx (6.0, 4.3)$$

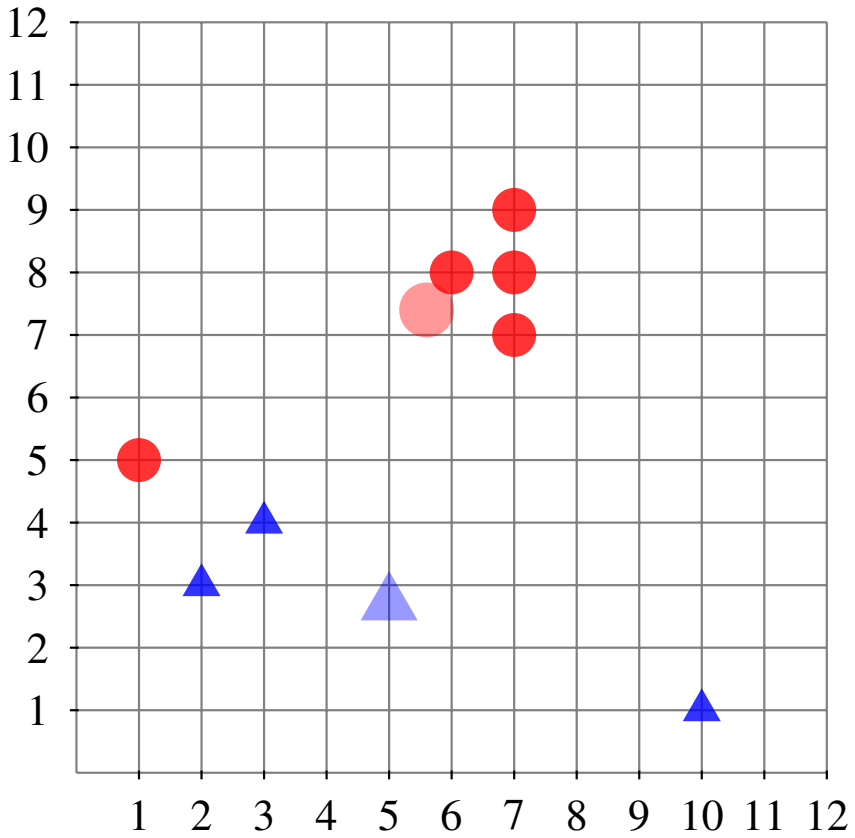
$$\mu \approx (5.0, 6.4)$$

k-Means Clustering – Lloyd/Forgy Algorithmus



Punkte neu zuordnen

k-Means Clustering – Lloyd/Forgy Algorithmus

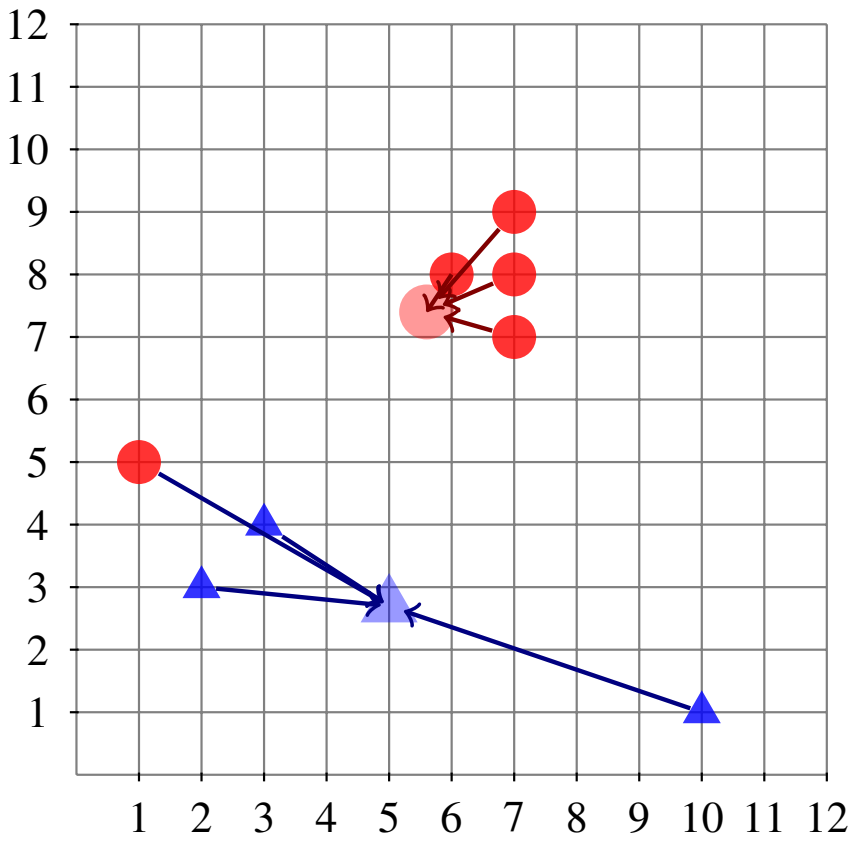


Zentroide neu berechnen:

$\mu \approx (5.0, 2.7)$

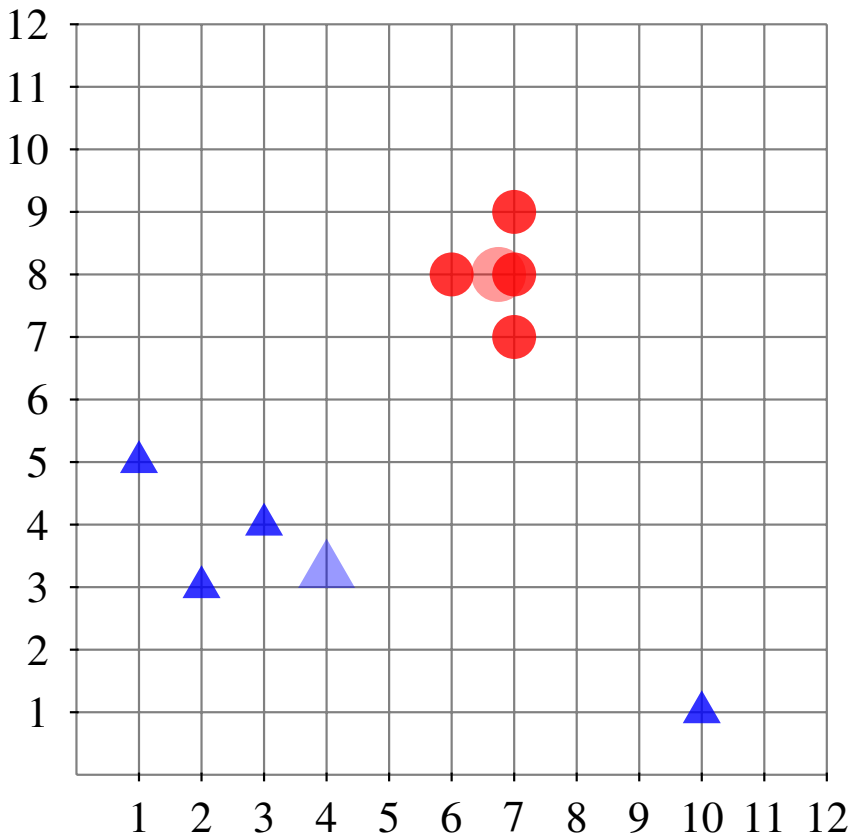
$\mu \approx (5.6, 7.4)$

k-Means Clustering – Lloyd/Forgy Algorithmus



Punkte neu zuordnen

k-Means Clustering – Lloyd/Forgy Algorithmus

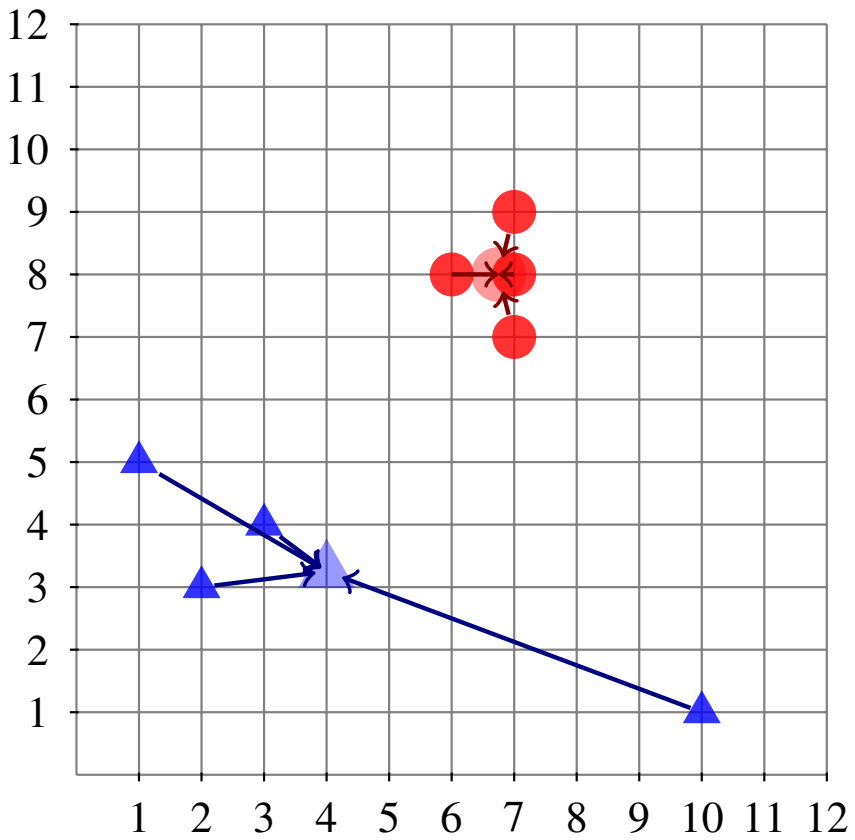


Zentroide neu berechnen:

$\mu \approx (4.0, 3.25)$

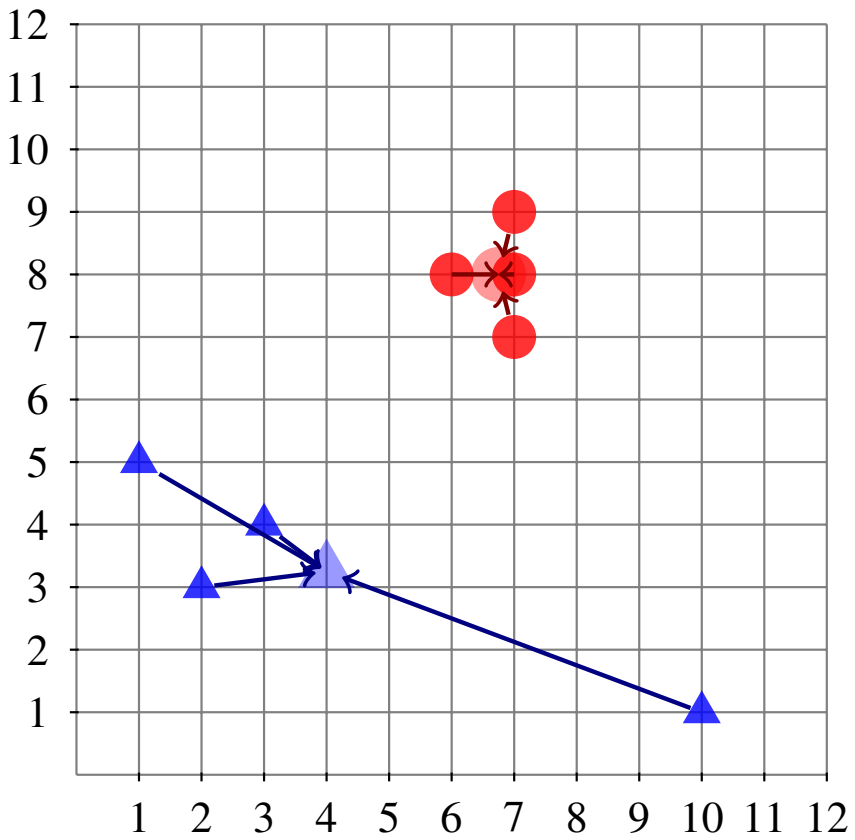
$\mu \approx (6.75, 8.0)$

k-Means Clustering – Lloyd/Forgy Algorithmus



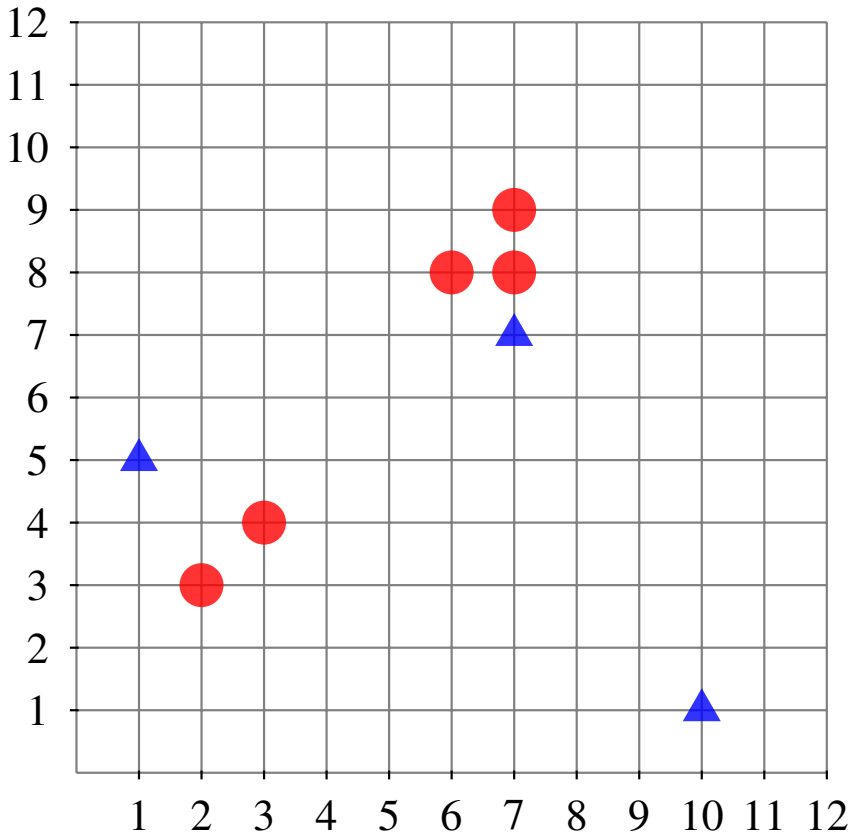
Punkte neu zuordnen

k-Means Clustering – Lloyd/Forgy Algorithmus

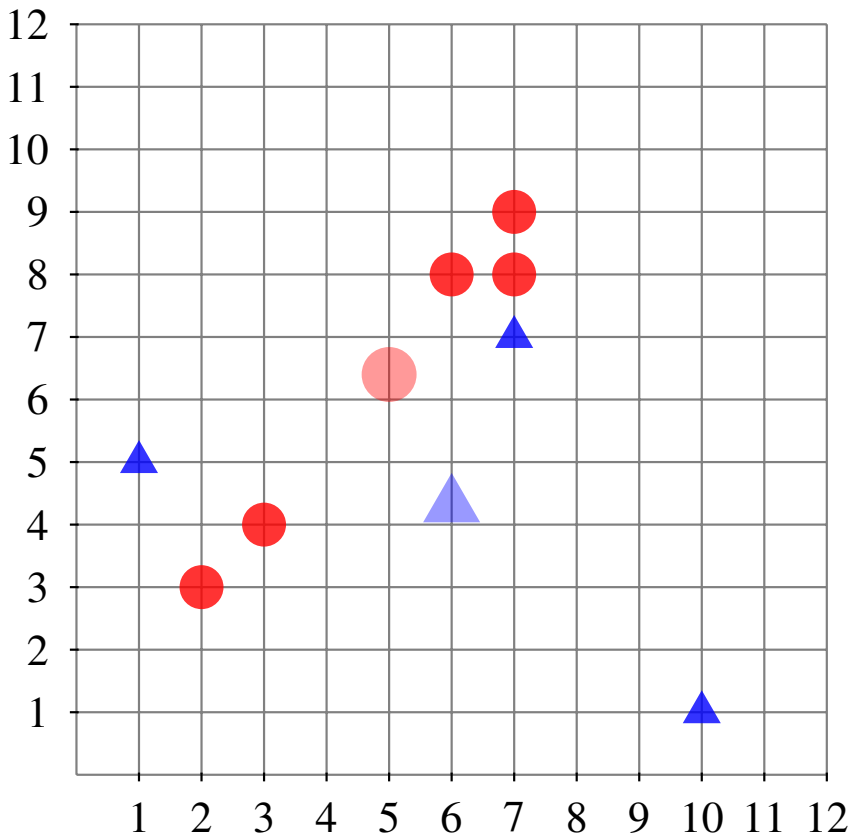


Punkte neu zuordnen
Keine Änderung
Konvergenz!

k-Means Clustering – MacQueen Algorithmus



k-Means Clustering – MacQueen Algorithmus

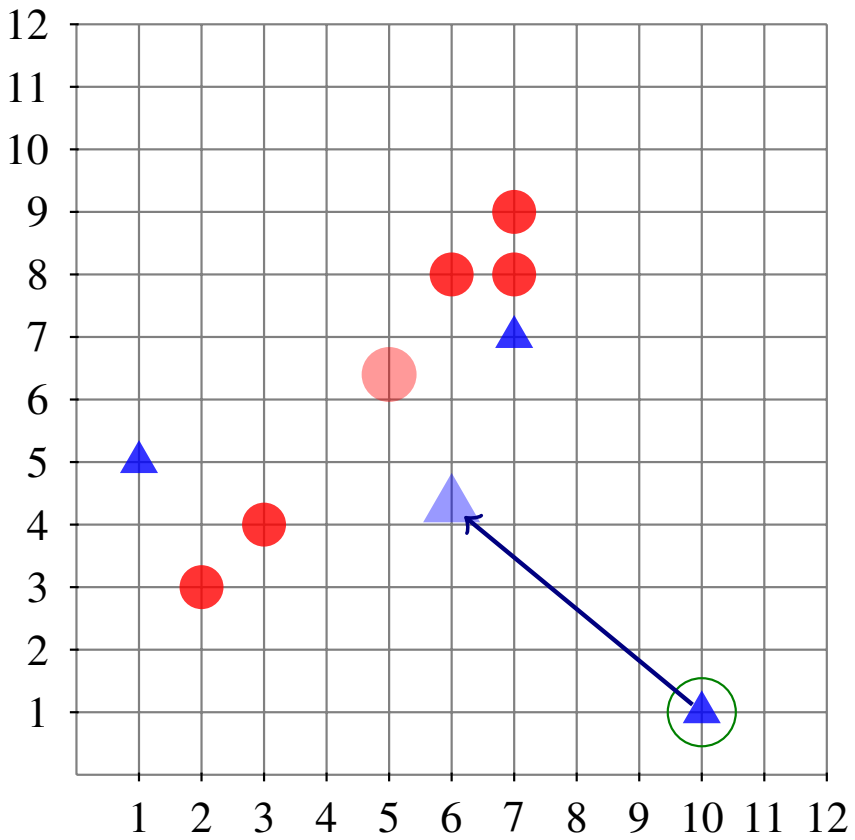


Zentroide (z.B.: aus vorheriger Iteration):

$\mu \approx (6.0, 4.3)$

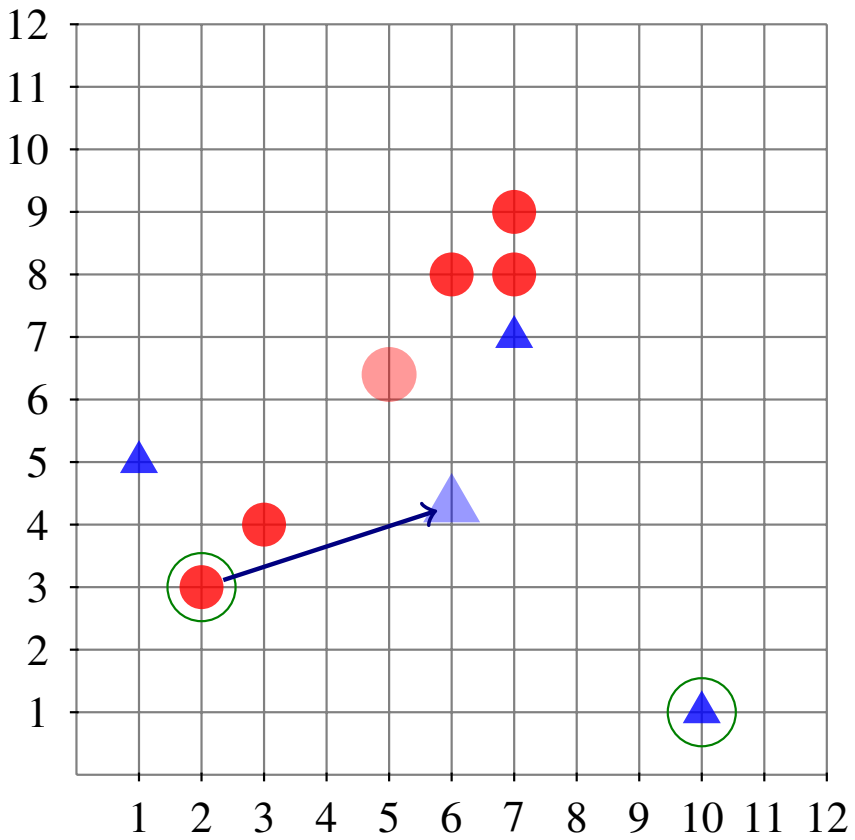
$\mu \approx (5.0, 6.4)$

k-Means Clustering – MacQueen Algorithmus



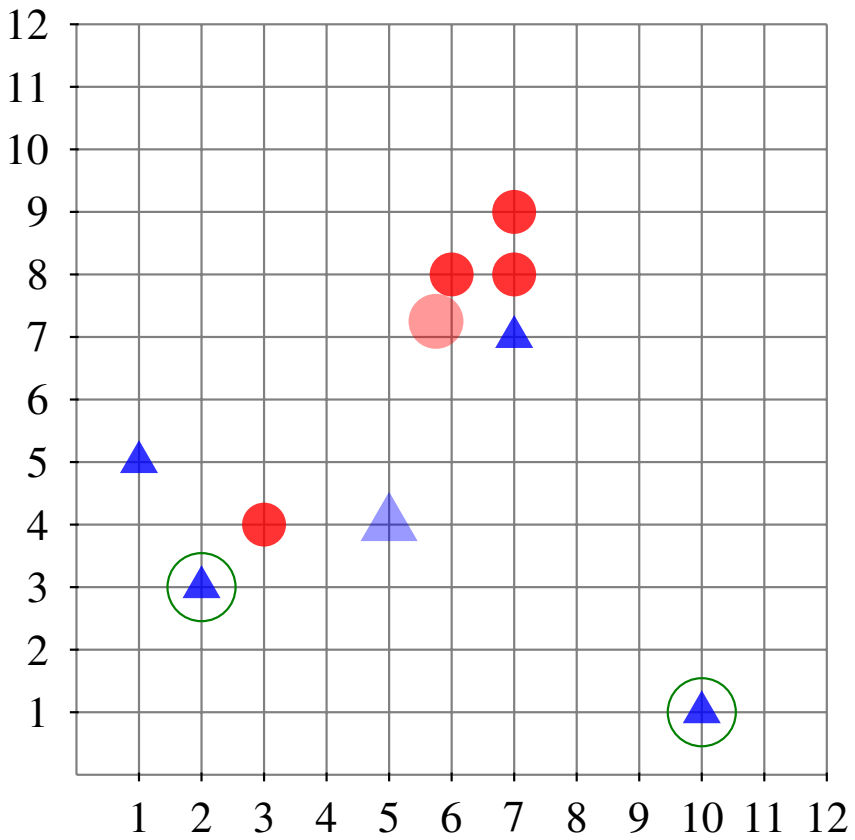
Ersten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus



Zweiten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

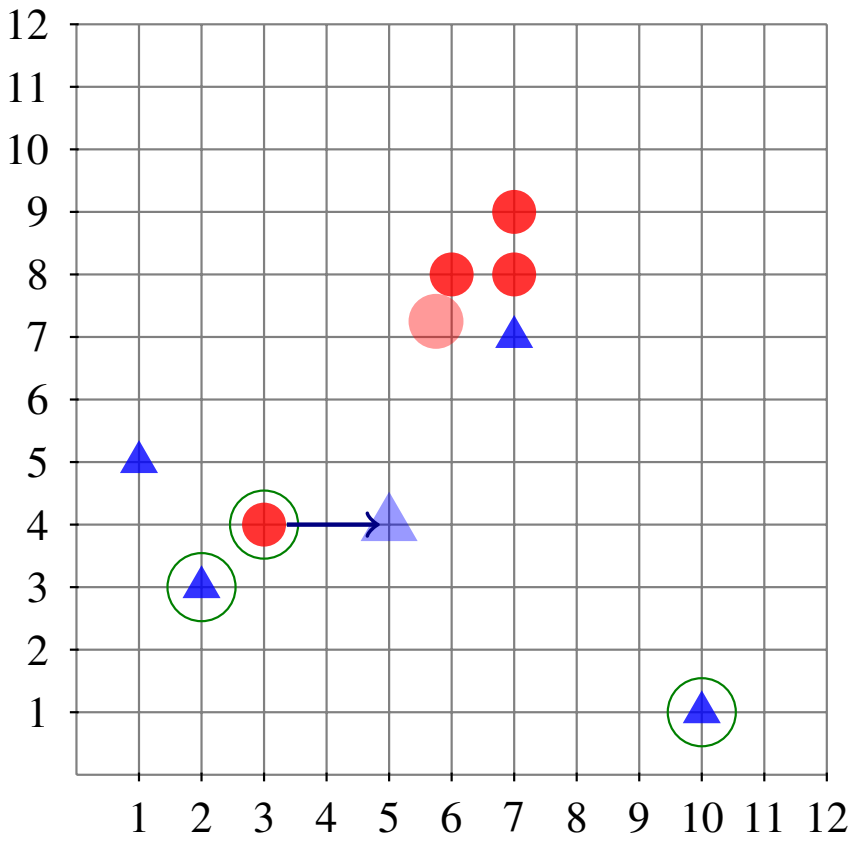


Zentroide aktualisieren:

$$\mu \approx (5.0, 4.0)$$

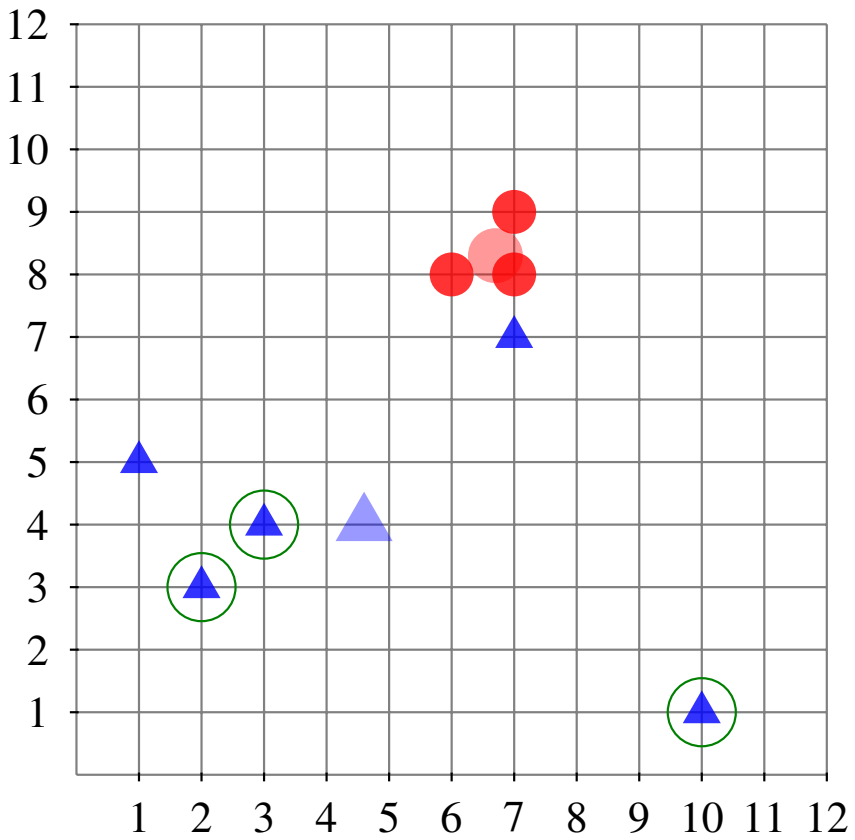
$$\mu \approx (5.75, 7.25)$$

k-Means Clustering – MacQueen Algorithmus



Dritten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

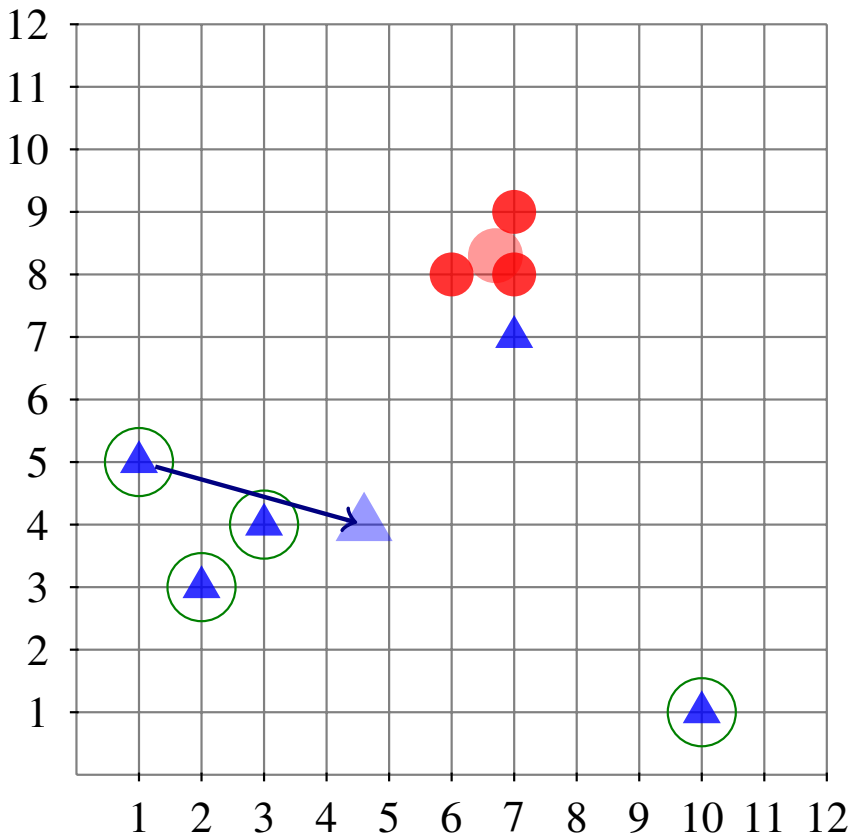


Zentroide aktualisieren:

$\mu \approx (4.6, 4.0)$

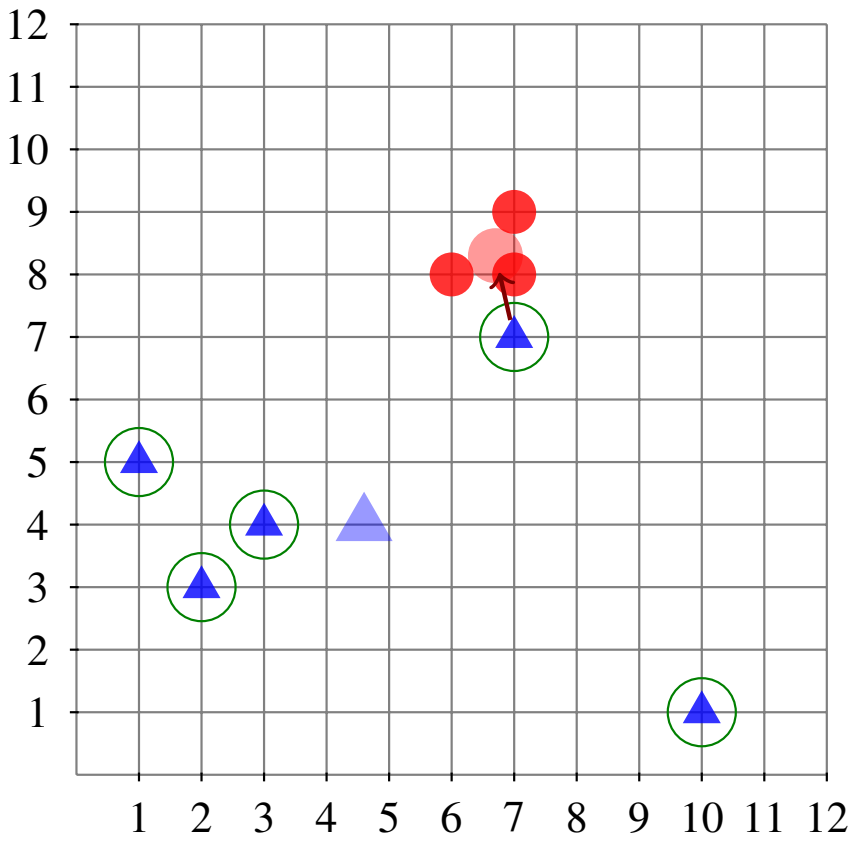
$\mu \approx (6.7, 8.3)$

k-Means Clustering – MacQueen Algorithmus



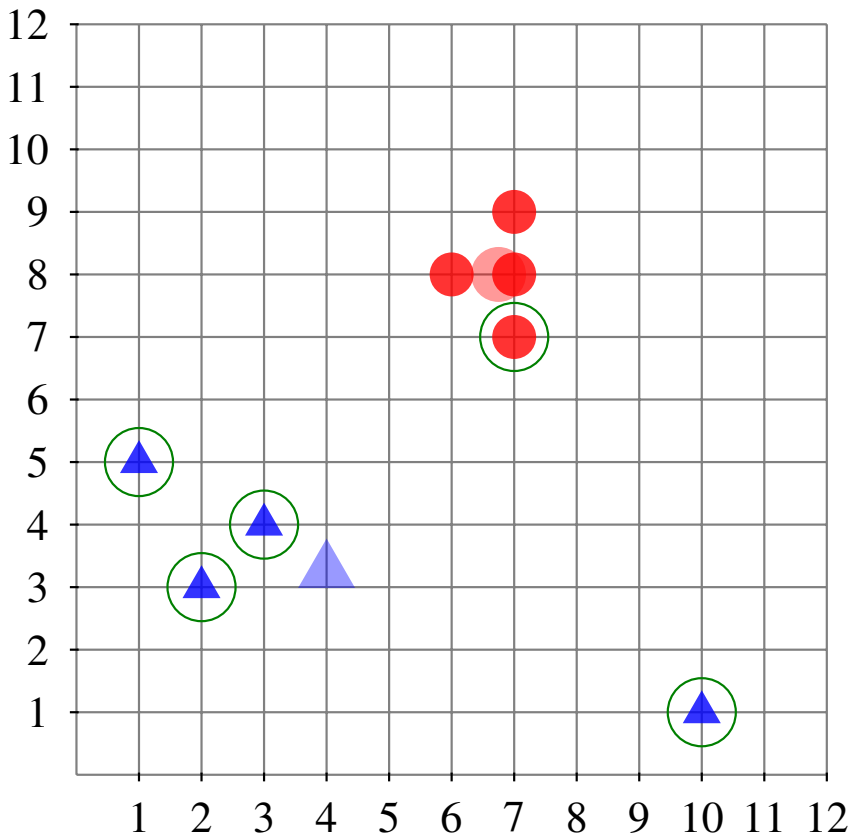
Vierten Punkt neu zuordnen

k-Means Clustering – MacQueen Algorithmus



Fünften Punkt neu zuordnen

k-Means Clustering – MacQueen Algorithmus

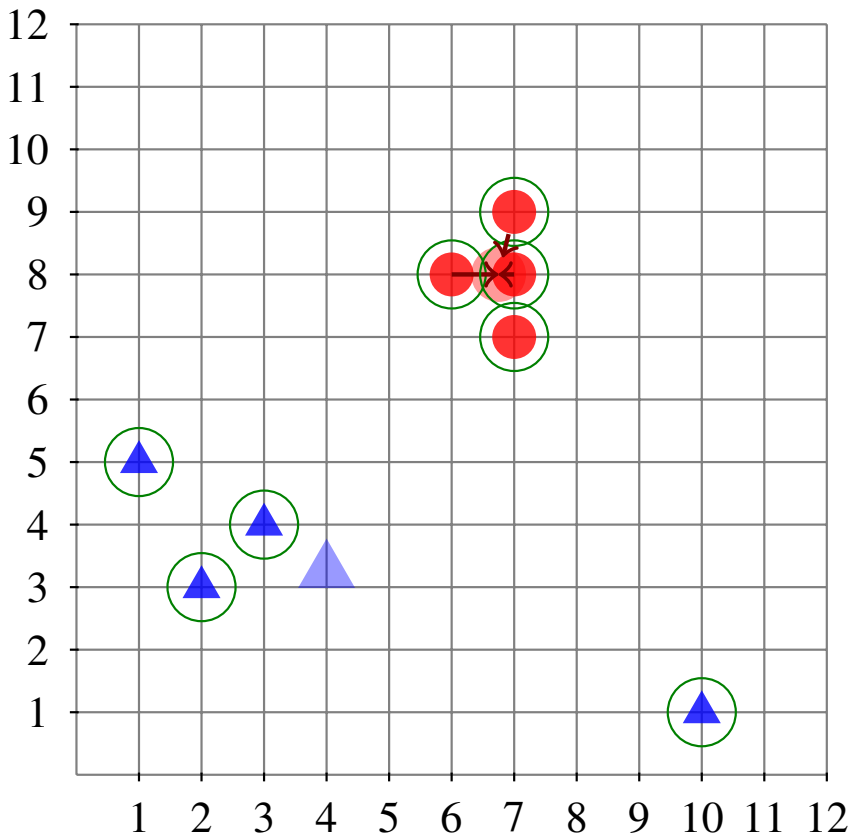


Zentroide aktualisieren:

$\mu \approx (4.0, 3.25)$

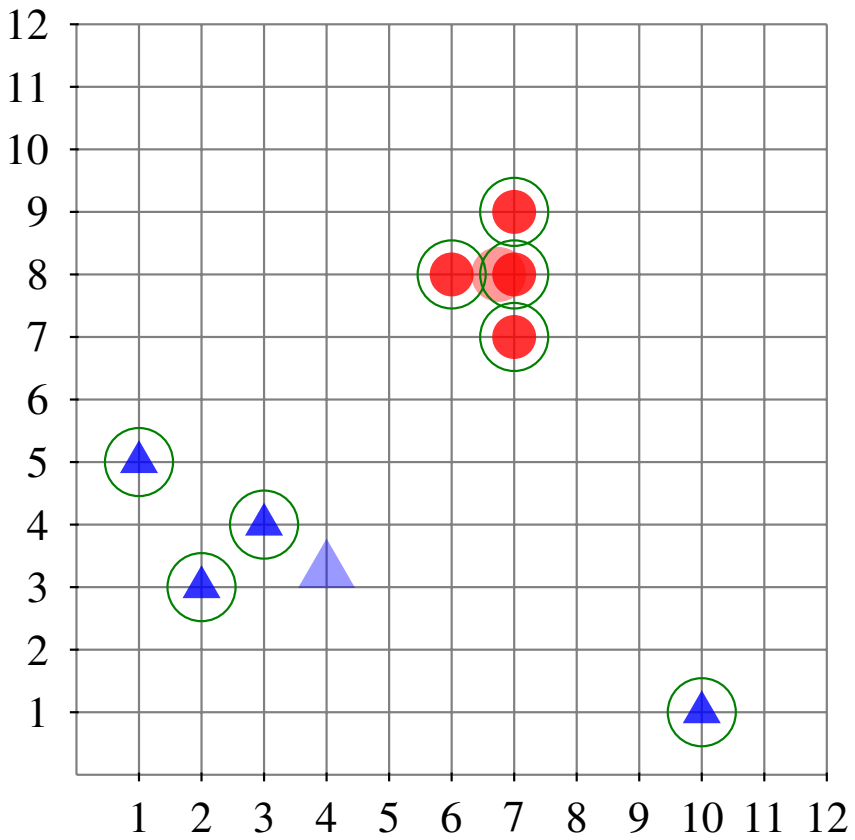
$\mu \approx (6.75, 8.0)$

k-Means Clustering – MacQueen Algorithmus



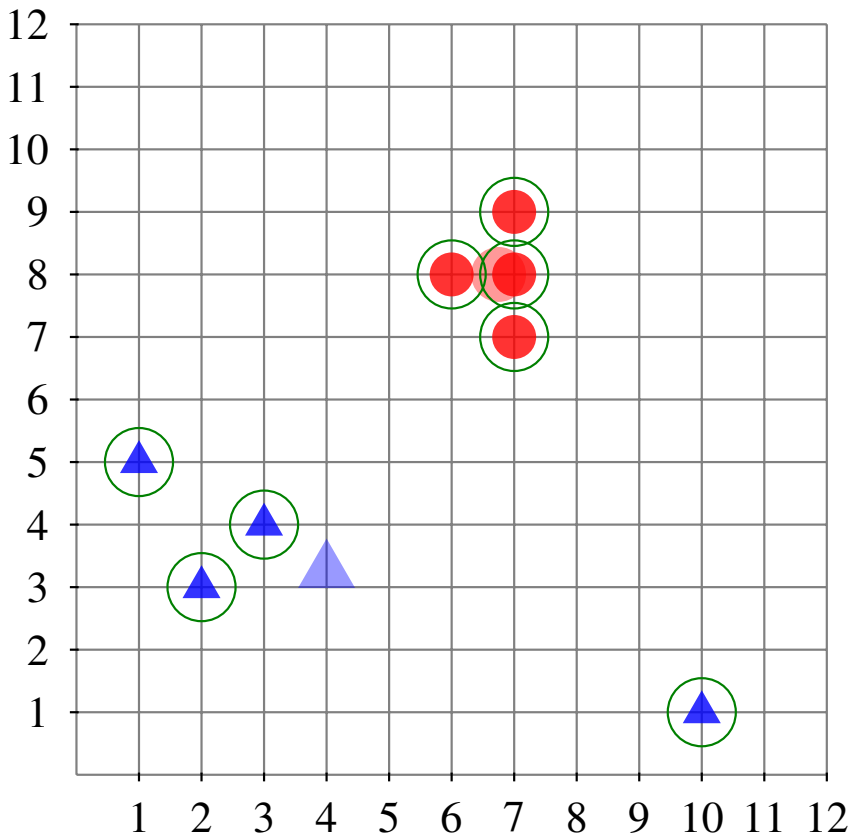
Weitere Punkte neu zuordnen

k-Means Clustering – MacQueen Algorithmus



Weitere Punkte neu zuordnen
ggf. Weitere Iterationen

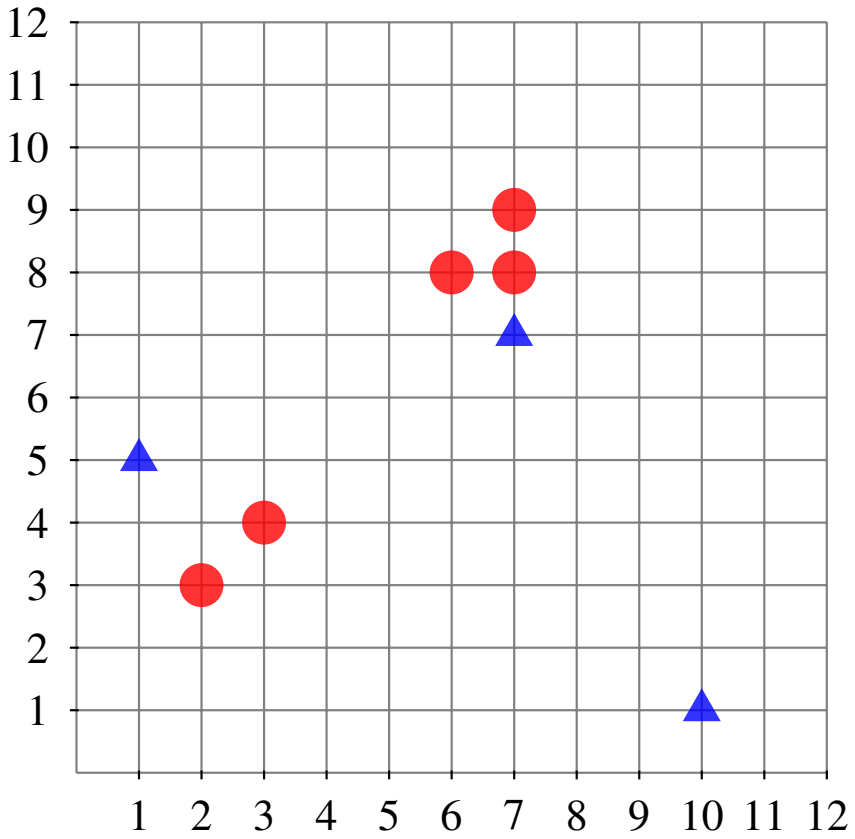
k-Means Clustering – MacQueen Algorithmus



Weitere Punkte neu zuordnen
ggf. Weitere Iterationen
Konvergenz

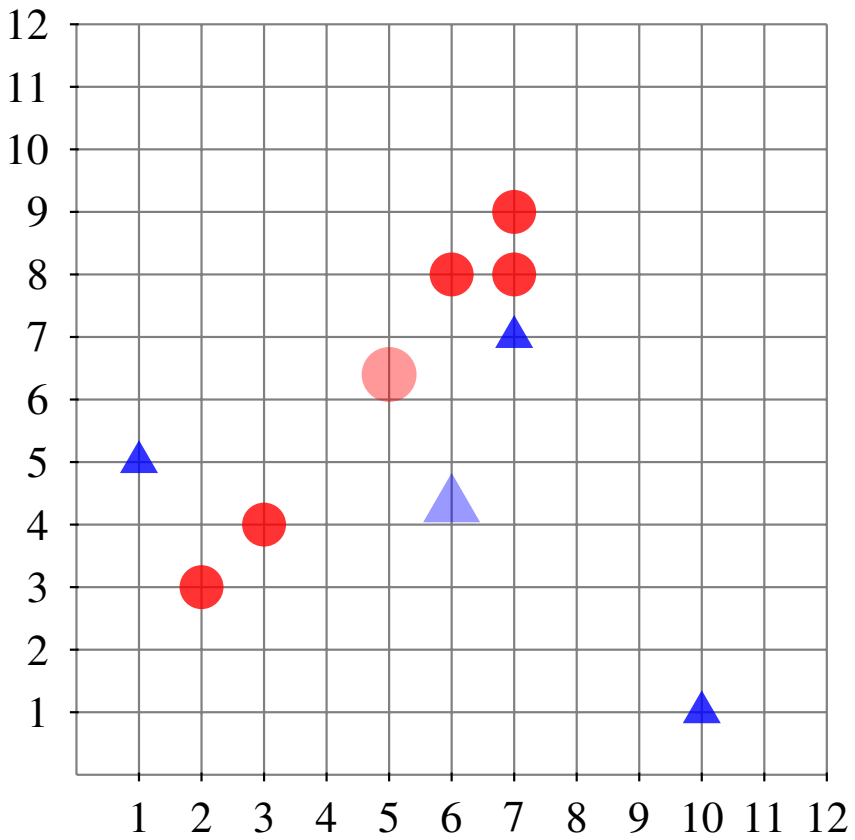
k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



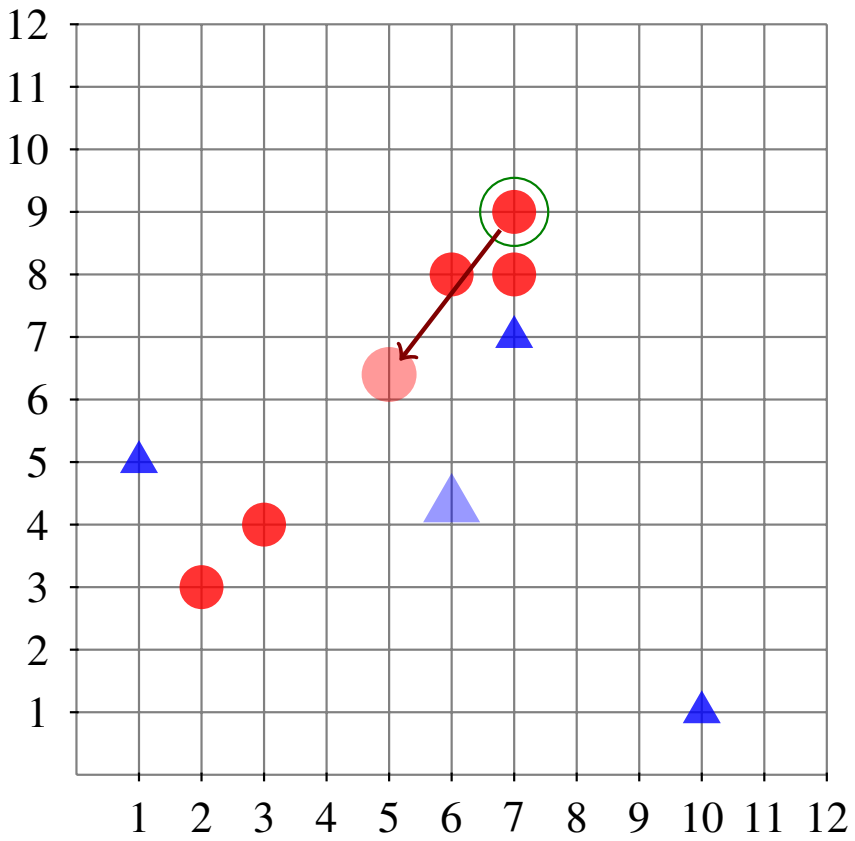
Zentroide (z.B.: aus vorheriger Iteration):

$\mu \approx (6.0, 4.3)$

$\mu \approx (5.0, 6.4)$

k-Means Clustering – MacQueen Algorithmus

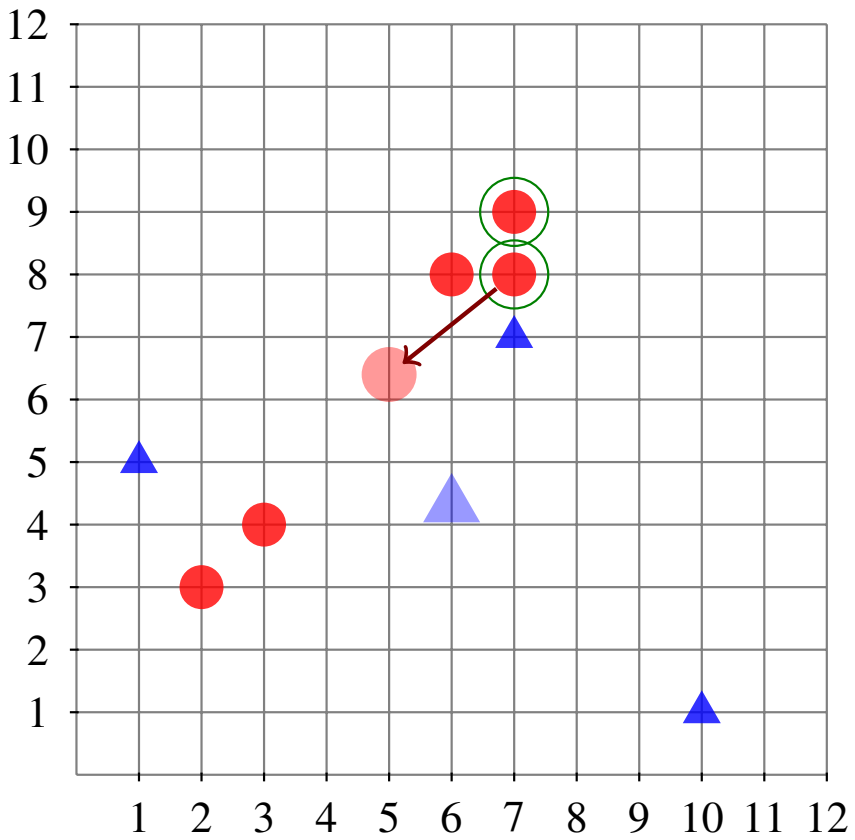
Alternativer Ablauf – andere Reihenfolge



Ersten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

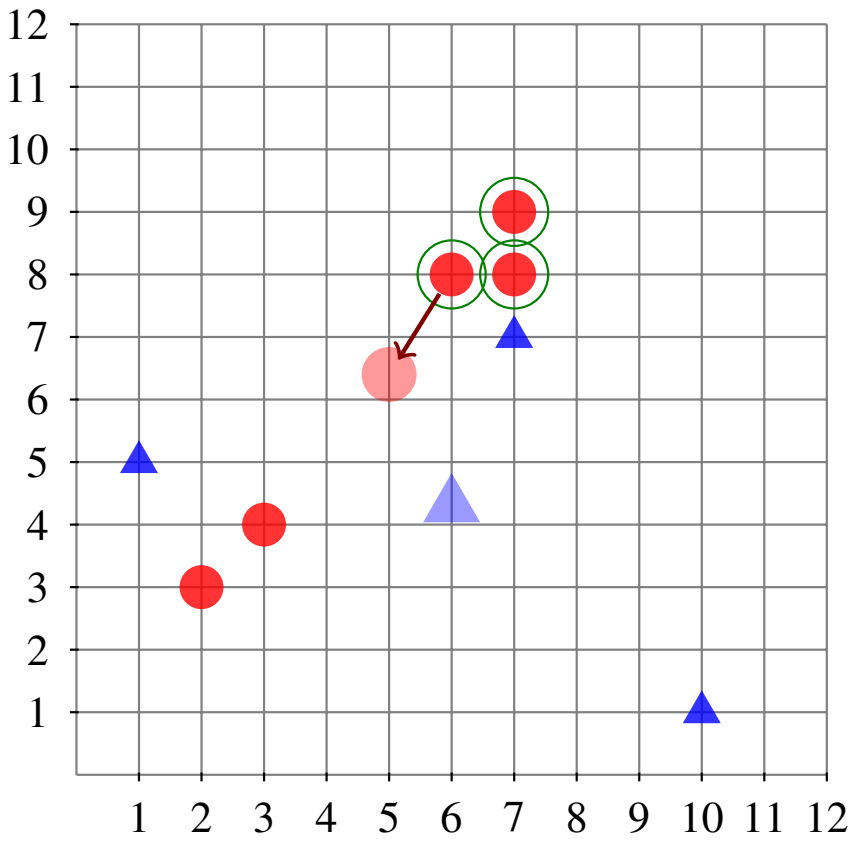
Alternativer Ablauf – andere Reihenfolge



Zweiten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

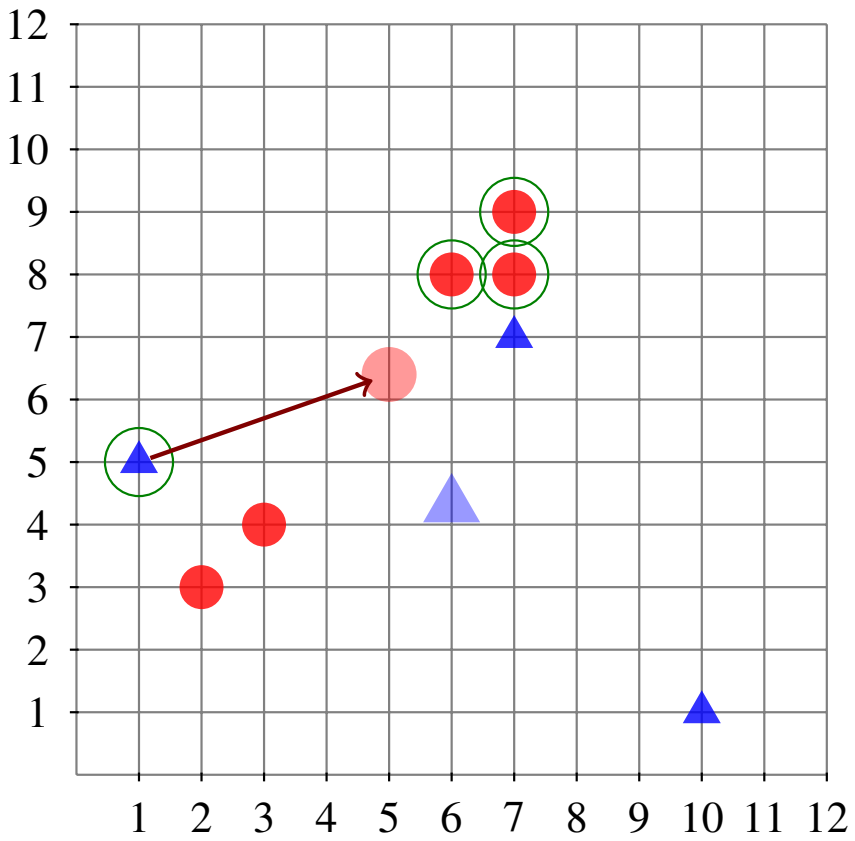
Alternativer Ablauf – andere Reihenfolge



Dritten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

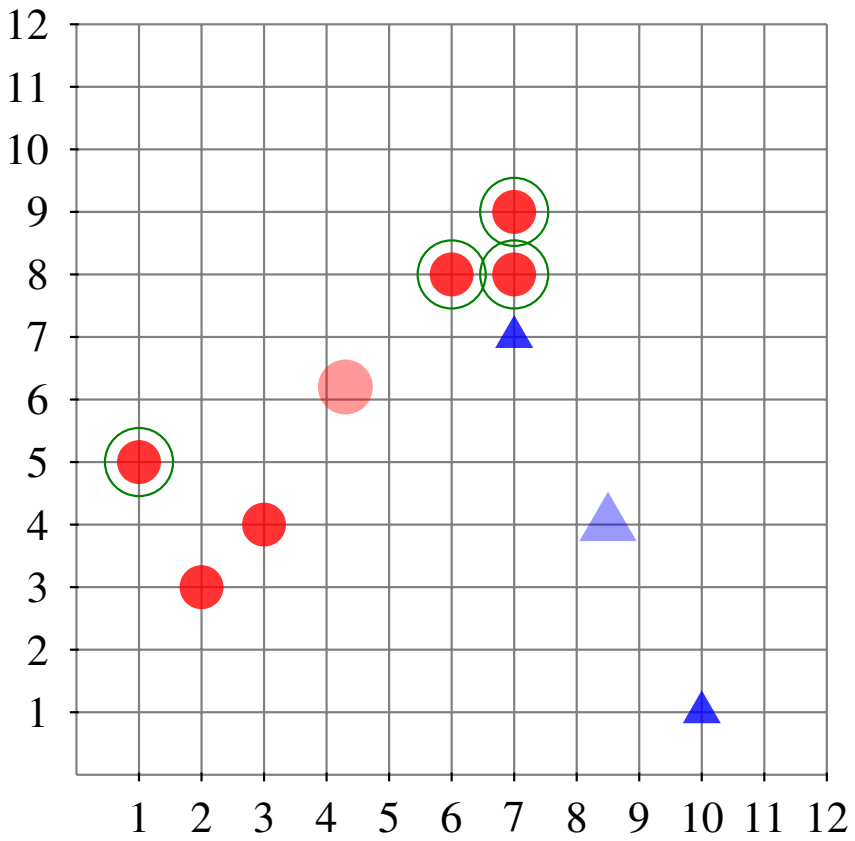
Alternativer Ablauf – andere Reihenfolge



Vierten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



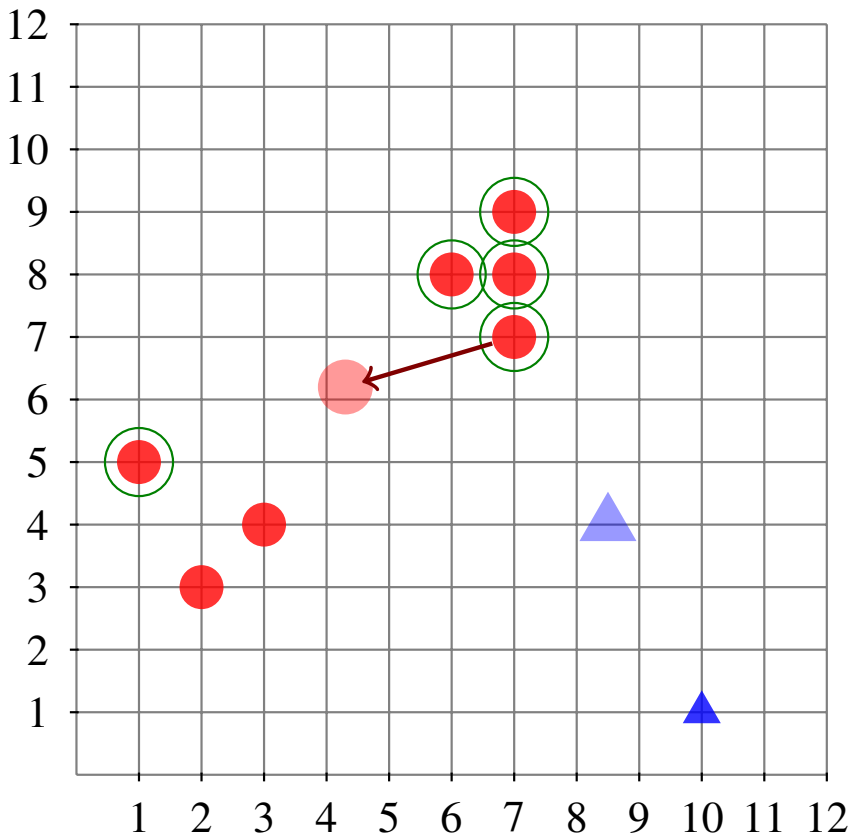
Zentroide aktualisieren:

$\mu \approx (4.0, 8.5)$

$\mu \approx (4.3, 6.2)$

k-Means Clustering – MacQueen Algorithmus

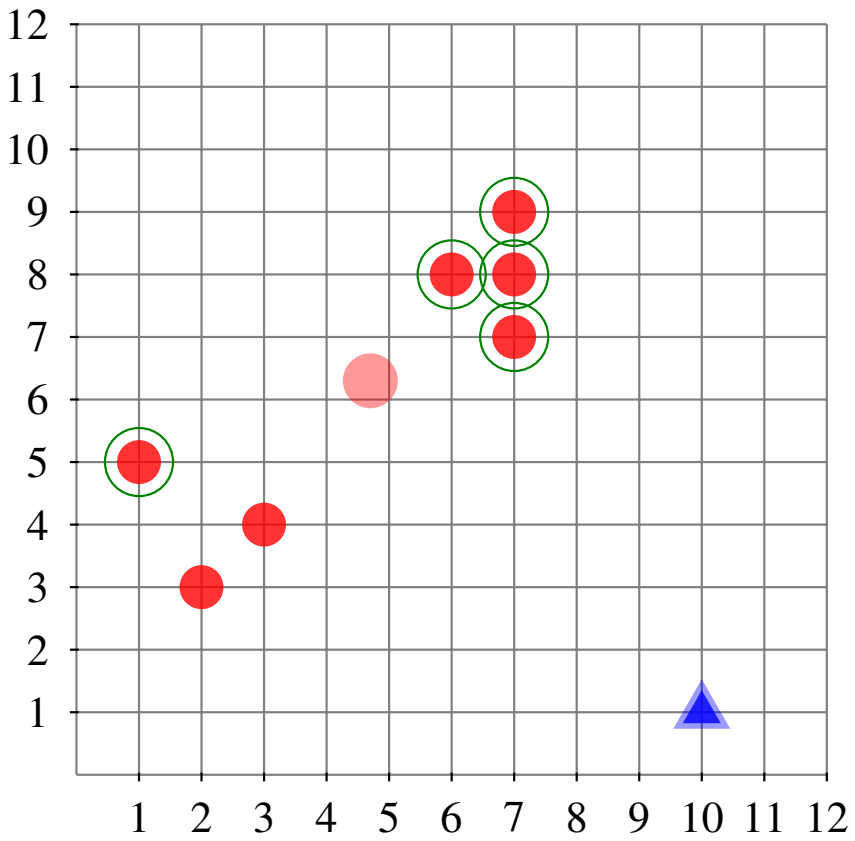
Alternativer Ablauf – andere Reihenfolge



Fünften Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



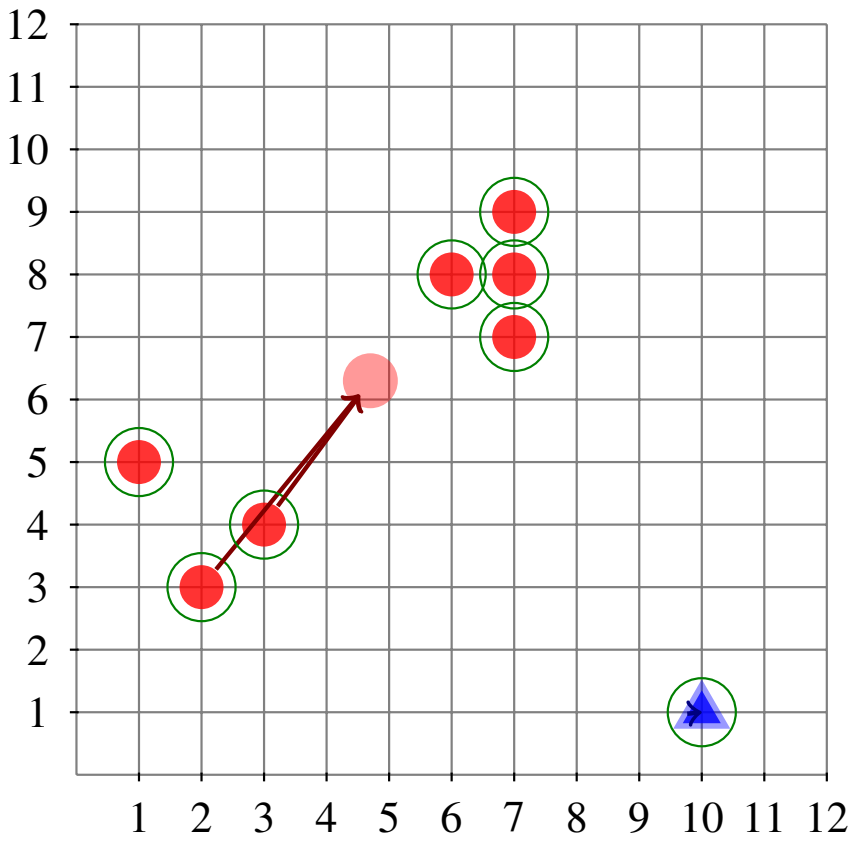
Zentroide aktualisieren:

$\mu \approx (10.0, 1.0)$

$\mu \approx (4.7, 6.3)$

k-Means Clustering – MacQueen Algorithmus

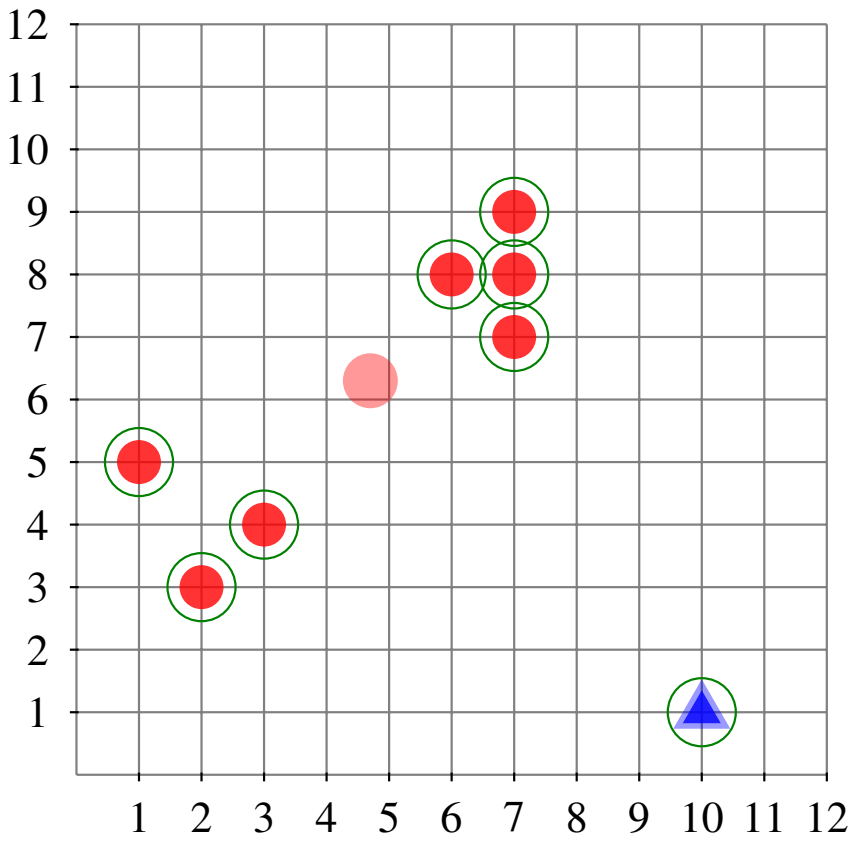
Alternativer Ablauf – andere Reihenfolge



Weitere Punkte neu zuordnen

k-Means Clustering – MacQueen Algorithmus

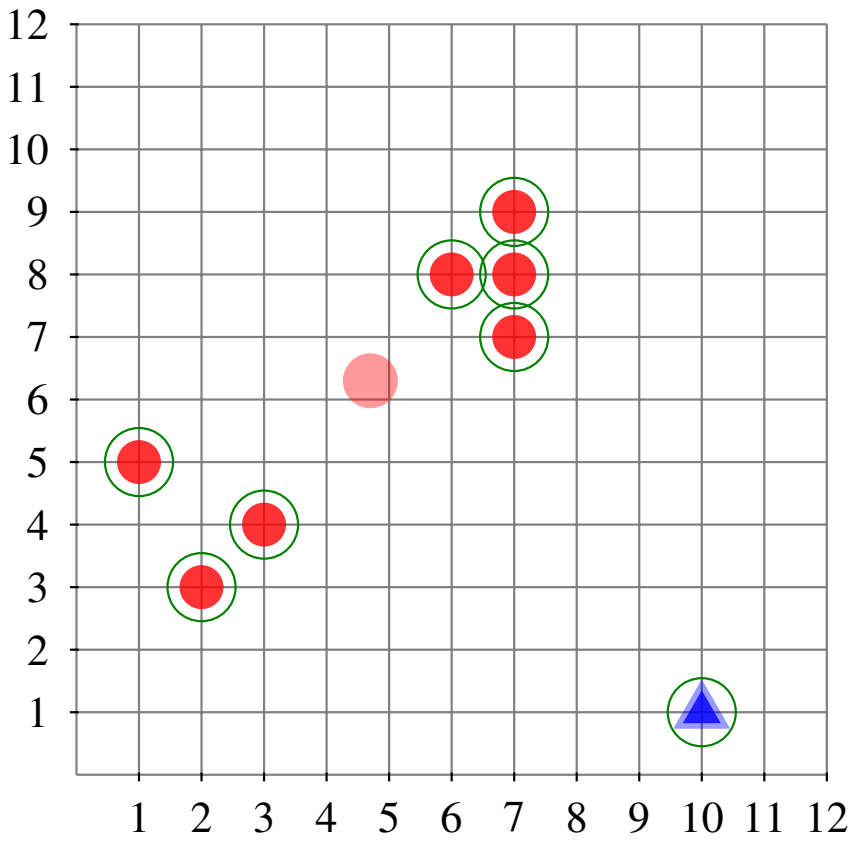
Alternativer Ablauf – andere Reihenfolge



Weitere Punkte neu zuordnen
ggf. Weitere Iterationen

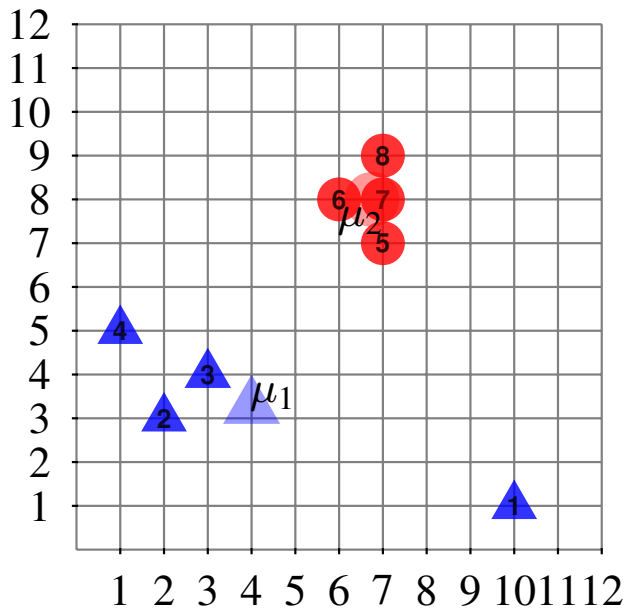
k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



Weitere Punkte neu zuordnen
ggf. Weitere Iterationen
Konvergenz

k-Means Clustering – Qualität



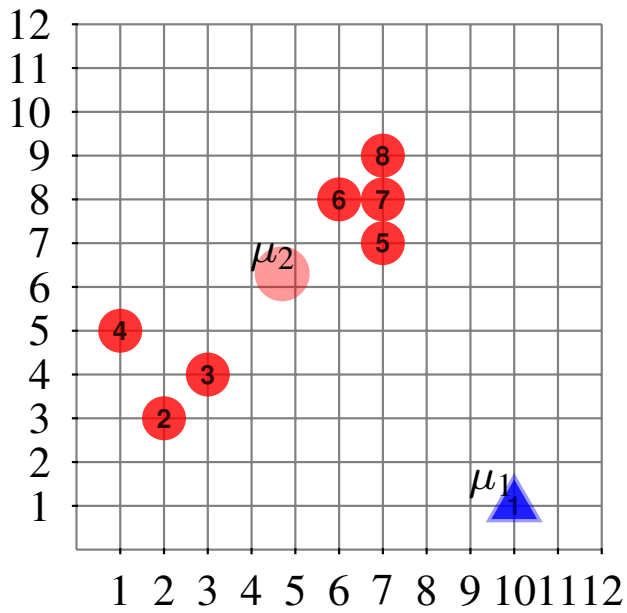
$$SSQ(\mu_1, p_1) = |4-10|^2 + |3.25-1|^2 = 36 + 5\tfrac{1}{16} = 41\tfrac{1}{16}$$
$$SSQ(\mu_1, p_2) = |4-2|^2 + |3.25-3|^2 = 4 + \tfrac{1}{16} = 4\tfrac{1}{16}$$
$$SSQ(\mu_1, p_3) = |4-3|^2 + |3.25-4|^2 = 1 + \tfrac{9}{16} = 1\tfrac{9}{16}$$
$$SSQ(\mu_1, p_4) = |4-1|^2 + |3.25-5|^2 = 9 + 3\tfrac{1}{16} = 12\tfrac{1}{16}$$
$$TD^2(C_1) = 58\tfrac{3}{4}$$

$$SSQ(\mu_2, p_5) = |6.75-7|^2 + |8-7|^2 = \tfrac{1}{16} + 1 = 1\tfrac{1}{16}$$
$$SSQ(\mu_2, p_6) = |6.75-6|^2 + |8-8|^2 = \tfrac{9}{16} + 0 = \tfrac{9}{16}$$
$$SSQ(\mu_2, p_7) = |6.75-7|^2 + |8-8|^2 = \tfrac{1}{16} + 0 = \tfrac{1}{16}$$
$$SSQ(\mu_2, p_8) = |6.75-7|^2 + |8-9|^2 = \tfrac{1}{16} + 1 = 1\tfrac{1}{16}$$
$$TD^2(C_2) = 2\tfrac{3}{4}$$

Erste Lösung: $TD^2 = 61\frac{1}{2}$

Beachte: $SSQ(\mu, p) = \mathit{Euclidean}(\mu, p)^2 = L_2^2(\mu, p)$.

k-Means Clustering – Qualität



$SSQ(\mu_1, p_1) = |10-10|^2 + |1-1|^2 = 0$
$TD^2(C_1) = 0$

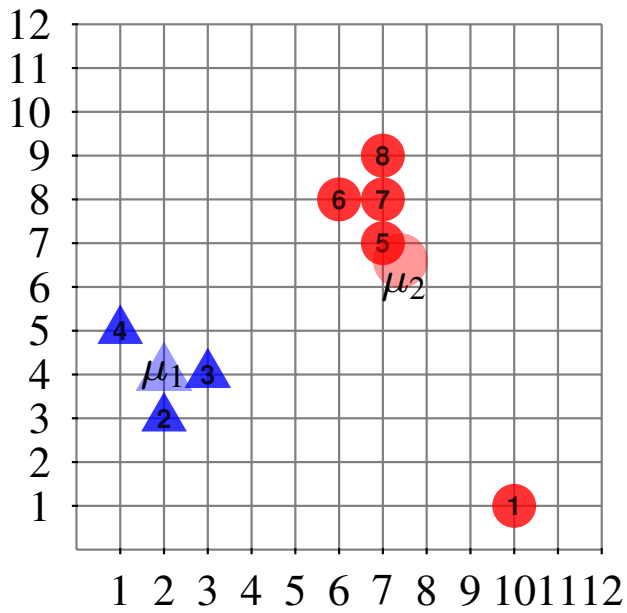
$SSQ(\mu_2, p_2) \approx |4.7-2|^2 + |6.3-3|^2 \approx 18.2$
$SSQ(\mu_2, p_3) \approx |4.7-3|^2 + |6.3-4|^2 \approx 8.2$
$SSQ(\mu_2, p_4) \approx |4.7-1|^2 + |6.3-5|^2 \approx 15.4$
$SSQ(\mu_2, p_5) \approx |4.7-7|^2 + |6.3-7|^2 \approx 5.7$
$SSQ(\mu_2, p_6) \approx |4.7-6|^2 + |6.3-8|^2 \approx 4.6$
$SSQ(\mu_2, p_7) \approx |4.7-7|^2 + |6.3-8|^2 \approx 8.2$
$SSQ(\mu_2, p_7) \approx |4.7-7|^2 + |6.3-9|^2 \approx 12.6$
$TD^2(C_2) \approx 72.86$

Erste Lösung: $TD^2 = 61\frac{1}{2}$

Zweite Lösung: $TD^2 \approx 72.68$

Beachte: $SSQ(\mu, p) = \mathit{Euclidean}(\mu, p)^2 = L_2^2(\mu, p)$.

k-Means Clustering – Qualität



$$SSQ(\mu_1, p_2) = |2-2|^2 + |4-3|^2 = 0 + 1 = 1$$
$$SSQ(\mu_1, p_3) = |2-3|^2 + |4-4|^2 = 1 + 0 = 1$$
$$SSQ(\mu_1, p_4) = |2-1|^2 + |4-5|^2 = 1 + 1 = 2$$
$$TD^2(C_1) = 4$$

$$SSQ(\mu_2, p_1) = |7.4-10|^2 + |6.6-1|^2 = 6\tfrac{19}{25} + 31\tfrac{9}{25} = 38\tfrac{3}{25}$$
$$SSQ(\mu_2, p_5) = |7.4-7|^2 + |6.6-7|^2 = \tfrac{4}{25} + \tfrac{4}{25} = \tfrac{8}{25}$$
$$SSQ(\mu_2, p_6) = |7.4-6|^2 + |6.6-8|^2 = 1\tfrac{24}{25} + 1\tfrac{24}{25} = 3\tfrac{23}{25}$$
$$SSQ(\mu_2, p_7) = |7.4-7|^2 + |6.6-8|^2 = \tfrac{4}{25} + 1\tfrac{24}{25} = 2\tfrac{3}{25}$$
$$SSQ(\mu_2, p_8) = |7.4-7|^2 + |6.6-9|^2 = \tfrac{4}{25} + 5\tfrac{19}{25} = 5\tfrac{23}{25}$$
$$TD^2(C_2) = 50\tfrac{2}{5}$$

Erste Lösung: $TD^2 = 61\frac{1}{2}$

Zweite Lösung: $TD^2 \approx 72.68$

Optimale Lösung: $TD^2 = 54\frac{2}{5}$

Beachte: $SSQ(\mu, p) = \mathit{Euclidean}(\mu, p)^2 = L_2^2(\mu, p)$.

Fazit k-Means Clustering

Merke:

- K-means konvergiert nur gegen ein lokales Minimum
- K-means ist abhängig von den Startparametern
- K-means nach MacQueen ist reihenfolgeabhängig
- K-means ist anfällig gegen Rauschen
  - Degenerierte 1-Element “Cluster”
  - Dadurch Reduktion von effektivem k
- K-means minimiert Varianzen, ist also eigentlich nur für Euklidische Distanz korrekt (oftmals aber auch Konvergenz bei anderen Distanzen)
- K-means (nach Lloyd) ist dennoch das beliebteste Verfahren, da es sehr einfach und schnell ist und mit geringem Aufwand implementiert werden kann!