Data Mining Tutorial

Clusteranalyse – Teil I

Erich Schubert, Arthur Zimek

Ludwig-Maximilians-Universität München

2013-05-03 — KDD Übung

Induzierte Metrik

Gegeben eine Pseudo-Metrik d auf der Menge A:
$d : A \times A \to \mathbb{R}_0^+$.

Definiere die Äquivalenzrelation $\sim$ so dass
$x \sim y \Leftrightarrow d(x, y) = 0$.

Sei $A^\sim$ die Menge der Äquivalenzklassen von A bzgl. $\sim$.

$$d^\sim : A^\sim \times A^\sim \to \mathbb{R}_0^+$$
$$\text{mit } d^\sim(x^\sim, y^\sim) := d(x, y)$$

Eigenschaften?

Induzierte Metrik

Gegeben eine Pseudo-Metrik d auf der Menge A:
$d : A \times A \to \mathbb{R}_0^+$.

Definiere die Äquivalenzrelation $\sim$ so dass
$x \sim y \Leftrightarrow d(x, y) = 0$.

Sei $A^\sim$ die Menge der Äquivalenzklassen von A bzgl. $\sim$.

$$d^\sim : A^\sim \times A^\sim \to \mathbb{R}_0^+$$
$$\text{mit } d^\sim(x^\sim, y^\sim) := d(x, y)$$

Eigenschaften? Wohldefiniert?

Induzierte Metrik

Zu zeigen: für alle $z \in x^{\sim}$, $w \in y^{\sim}$ gilt $d(z, w) = d(x, y)$.

Induzierte Metrik

Zu zeigen: für alle $z \in x^{\sim}$, $w \in y^{\sim}$ gilt $d(z,w) = d(x,y)$.

Da $z \in x^{\sim}$ und $w \in y^{\sim}$ haben wir per Definition:

$$z^{\sim} = x^{\sim} \text{ und } d(z,x) = 0$$
$$w^{\sim} = y^{\sim} \text{ und } d(w,y) = 0$$

Induzierte Metrik

Zu zeigen: für alle $z \in x^{\sim}$, $w \in y^{\sim}$ gilt $d(z,w) = d(x,y)$.

Da $z \in x^{\sim}$ und $w \in y^{\sim}$ haben wir per Definition:

$$z^{\sim} = x^{\sim} \text{ und } d(z,x) = 0$$
$$w^{\sim} = y^{\sim} \text{ und } d(w,y) = 0$$

Durch $1\times$ anwenden Dreiecksungleichung erhalten wir:

$$d(z,w) \leq d(z,x) + d(x,w)$$
$$d(x,y) \leq d(x,z) + d(z,y)$$

Induzierte Metrik

Zu zeigen: für alle $z \in x^{\sim}$, $w \in y^{\sim}$ gilt $d(z,w) = d(x,y)$.

Da $z \in x^{\sim}$ und $w \in y^{\sim}$ haben wir per Definition:

$$z^{\sim} = x^{\sim} \text{ und } d(z,x) = 0$$
$$w^{\sim} = y^{\sim} \text{ und } d(w,y) = 0$$

Durch $2\times$ anwenden Dreiecksungleichung erhalten wir:

$$d(z,w) \leq d(z,x) + d(x,y) + d(y,w)$$
$$d(x,y) \leq d(x,z) + d(z,w) + d(w,y)$$

Induzierte Metrik

Zu zeigen: für alle $z \in x^{\sim}$, $w \in y^{\sim}$ gilt $d(z,w) = d(x,y)$.

Da $z \in x^{\sim}$ und $w \in y^{\sim}$ haben wir per Definition:

$$z^{\sim} = x^{\sim} \text{ und } d(z,x) = 0$$
$$w^{\sim} = y^{\sim} \text{ und } d(w,y) = 0$$

Durch $2\times$ anwenden Dreiecksungleichung erhalten wir:

$$\begin{aligned} d(z,w) &\leq d(z,x) + d(x,y) + d(y,w) &&= d(x,y) \\ d(x,y) &\leq d(x,z) + d(z,w) + d(w,y) &&= d(z,w) \end{aligned}$$

Induzierte Metrik

Zu zeigen: für alle $z \in x^{\sim}$, $w \in y^{\sim}$ gilt $d(z,w) = d(x,y)$.

Da $z \in x^{\sim}$ und $w \in y^{\sim}$ haben wir per Definition:

$$z^{\sim} = x^{\sim} \text{ und } d(z,x) = 0$$
$$w^{\sim} = y^{\sim} \text{ und } d(w,y) = 0$$

Durch $2\times$ anwenden Dreiecksungleichung erhalten wir:

$$\begin{aligned} d(z,w) &\leq d(z,x) + d(x,y) + d(y,w) &&= d(x,y) \\ d(x,y) &\leq d(x,z) + d(z,w) + d(w,y) &&= d(z,w) \end{aligned}$$

$\Rightarrow$ Distanzberechnungen auf den Äquivalenzklassen wohldefiniert: egal welchen *Repräsentanten* wir wählen, es kommt das gleiche Ergebnis für $d^{\sim}$ heraus.

Induzierte Metrik

Zu zeigen (Striktheit): $d^{\sim}(a^{\sim}, b^{\sim}) = 0 \Rightarrow a^{\sim} = b^{\sim}$

Induzierte Metrik

Zu zeigen (Striktheit): $d^{\sim}(a^{\sim}, b^{\sim}) = 0 \Rightarrow a^{\sim} = b^{\sim}$

$$
\begin{aligned}
& d^{\sim}(a^{\sim}, b^{\sim}) = 0 \\
\Rightarrow & d(a, b) = 0 \\
\Rightarrow & a \sim b \\
\Rightarrow & a^{\sim} = b^{\sim}
\end{aligned}
$$

Induzierte Metrik

Zu zeigen (Striktheit): $d^{\sim}(a^{\sim}, b^{\sim}) = 0 \Rightarrow a^{\sim} = b^{\sim}$

$$
\begin{aligned}
& d^{\sim}(a^{\sim}, b^{\sim}) = 0 \\
\Rightarrow & d(a, b) = 0 \\
\Rightarrow & a \sim b \\
\Rightarrow & a^{\sim} = b^{\sim}
\end{aligned}
$$

Reflexivität, Symmetrie und Dreickecksungleichung folgen trivial aus den Eigenschaften von d!

Induzierte Metrik

$$\text{euclid}_{xy}((r_1, x_1, y_1), (r_2, x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Datensatz ID	x	y
1	0	1
2	1	1
3	0	1

Datensatz ID	x	y
4	1	1
5	2	2
6	3	3

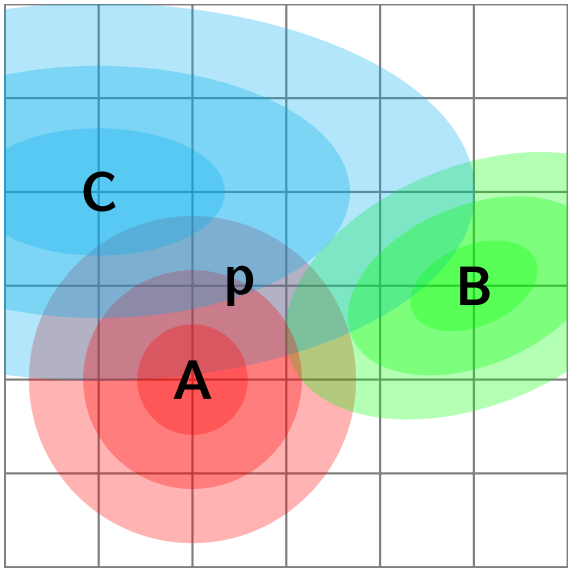
Induzierte Metrik

$$\text{euclid}_{xy}((r_1, x_1, y_1), (r_2, x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Datensatz ID	x	y	Datensatz ID	x	y
1	0	1	4	1	1
2	1	1	5	2	2
3	0	1	6	3	3

Euklidische Distanz auf $X \times Y$. Metrik auf $\mathbb{R}^2 \sim X \times Y$, aber nur eine Pseudo-metric auf Datensatz ID $\times X \times Y$. “Duplikate” haben eine Distanz von 0!

Multivariate Normalverteilung

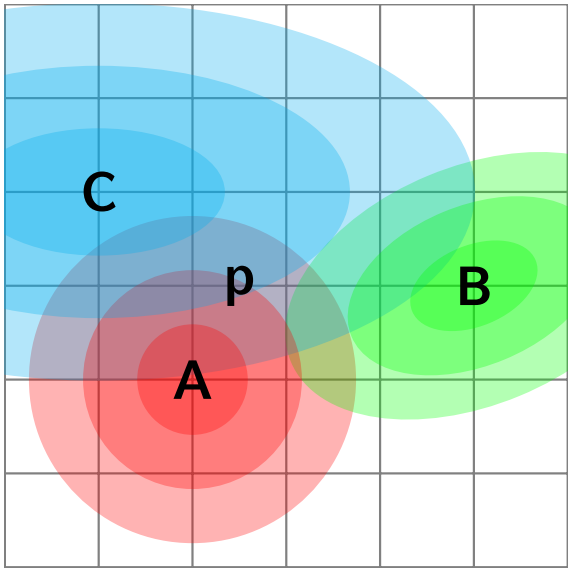


Cluster können sein:

- Kugelförmig (A)
- Ellipsoid (C)
- Rotierter Ellipsoid (B)

Aber: es ist immer die gleiche Formel!

Multivariate Normalverteilung

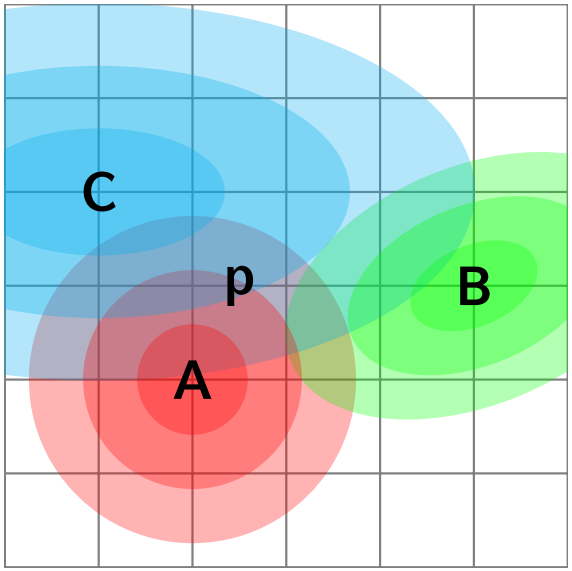


Cluster können sein:

- Kugelförmig (A)
- Ellipsoid (C)
- Rotierter Ellipsoid (B)

Aber: es ist immer die gleiche Formel!

Multivariate Normalverteilung

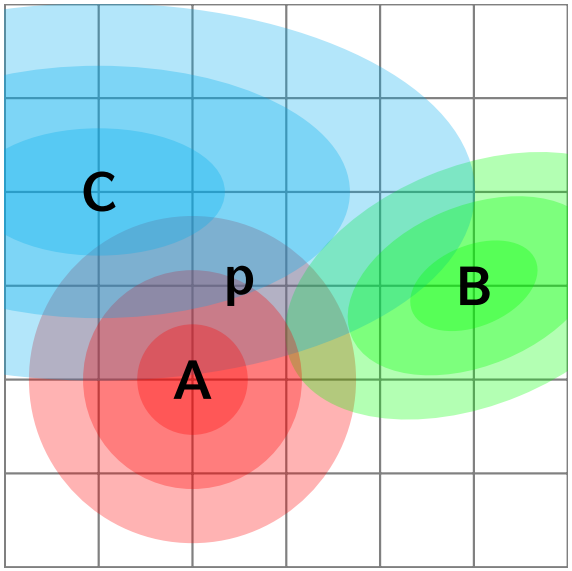


Wahrscheinlichkeitsdichte (PDF) der multivariaten Normalverteilung:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

Die wichtigste mehrdimensionale Verteilungsfunktion!

Multivariate Normalverteilung



Wahrscheinlichkeitsdichte (PDF) der multivariaten Normalverteilung:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

Die wichtigste mehrdimensionale Verteilungsfunktion!

Das sollten wir uns genauer anschauen!

Multivariate Normalverteilung:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

Multivariate Normalverteilung
Zerlegung der Formel

Multivariate Normalverteilung:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

1-dimensionale Normalverteilung

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)}\sigma} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

Multivariate Normalverteilung
Zerlegung der Formel

Multivariate Normalverteilung:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

1-dimensionale Normalverteilung

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma}} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

Multivariate Normalverteilung
Zerlegung der Formel

Multivariate Normalverteilung:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

1-dimensionale Normalverteilung

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma}} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

Normalisierung (auf Gesamtvolumen 1!) und quadrierte Abweichung vom Mittelwert

Multivariate Normalverteilung
Zerlegung der Formel

Multivariate Normalverteilung:

$$pdf(p,\mu,\Sigma) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T\Sigma^{-1}(p-\mu)\right)}$$

1-dimensionale Normalverteilung

$$pdf(x,\mu,\sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Normalisierung (auf Gesamtvolumen 1!) und quadrierte Abweichung vom Mittelwert

Multivariate Normalverteilung
Zerlegung der Formel

Multivariate Normalverteilung:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

1-dimensionale Normalverteilung

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Mahalanobis-Distanz:

$$d_{Mahalanobis}(x, \mu, \Sigma) := \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Multivariate Normalverteilung
Zerlegung der Formel

Multivariate Normalverteilung:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

1-dimensionale Normalverteilung

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Mahalanobis-Distanz:

$$d_{Mahalanobis}(x, \mu, \Sigma)^2 := (x-\mu)^T \Sigma^{-1} (x-\mu)$$

Multivariate Normalverteilung
Zerlegung der Formel

Multivariate Normalverteilung:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

1-dimensionale Normalverteilung

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Mahalanobis-Distanz:

$$d_{Mahalanobis}(x, \mu, \Sigma)^2 := (x-\mu)^T \Sigma^{-1} (x-\mu)$$

Was ist die Rolle von Σ^{-1}?

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Kovarianzmatrizen sind symmetrisch und auf der Diagonalen nicht negativ, und können daher normalerweise invertiert werden (es gibt degenerierte Fälle, auch die kann man handhaben!)

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Kovarianzmatrizen sind symmetrisch und auf der Diagonalen nicht negativ, und können daher normalerweise invertiert werden (es gibt degenerierte Fälle, auch die kann man handhaben!)
Die Matrix kann zerlegt werden:

$$V\Lambda V^{-1} = \Sigma \quad \equiv \quad V\Lambda^{-1}V^{-1} = \Sigma^{-1}$$

wobei V die Eigenvektoren und Λ die Eigenwerte enthält.

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Kovarianzmatrizen sind symmetrisch und auf der Diagonalen nicht negativ, und können daher normalerweise invertiert werden (es gibt degenerierte Fälle, auch die kann man handhaben!)
Die Matrix kann zerlegt werden:

$$V\Lambda V^{-1} = \Sigma \quad \equiv \quad V\Lambda^{-1}V^{-1} = \Sigma^{-1}$$

wobei V die Eigenvektoren und Λ die Eigenwerte enthält.
$V \cong$ Drehung, $\Lambda \cong$ Skalierung^2!
(Das ist die Kernidee der Hauptachsentransformation=PCA)

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Konstruiere Ω als $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Dann gilt $\Omega\Omega = \Lambda^{-1}$.

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Konstruiere Ω als $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Dann gilt $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Konstruiere Ω als $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Dann gilt $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$d^2_{\text{Mahalanobis}} = (x-\mu)^T\Sigma^{-1}(x-\mu)$$

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Konstruiere Ω als $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Dann gilt $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$d^2_{\text{Mahalanobis}} = (x-\mu)^T V\Omega(V\Omega)^T(x-\mu)$$

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Konstruiere Ω als $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Dann gilt $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)^T V\Omega(V\Omega)^T(x-\mu) \\ &= \left\langle (V\Omega)^T(x-\mu), (V\Omega)^T(x-\mu) \right\rangle \end{aligned}$$

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Konstruiere Ω als $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Dann gilt $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)^T V\Omega(V\Omega)^T(x-\mu) \\ &= \left\langle (V\Omega)^T(x-\mu), (V\Omega)^T(x-\mu) \right\rangle \\ &= L_2((V\Omega)^T(x-\mu))^2 \end{aligned}$$

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Konstruiere Ω als $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Dann gilt $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)^T V\Omega(V\Omega)^T(x-\mu) \\ &= \left\langle (V\Omega)^T(x-\mu), (V\Omega)^T(x-\mu) \right\rangle \\ &= L_2((V\Omega)^T(x-\mu))^2 \end{aligned}$$

L_2 ist die L_2-Norm (Euclidische Distanz $d(x,y) = L_2(x-y)$!)

Multivariate Normalverteilung

Inverse Kovarianzmatrix – Σ^{-1}

Konstruiere Ω als $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Dann gilt $\Omega\Omega = \Lambda^{-1}$.

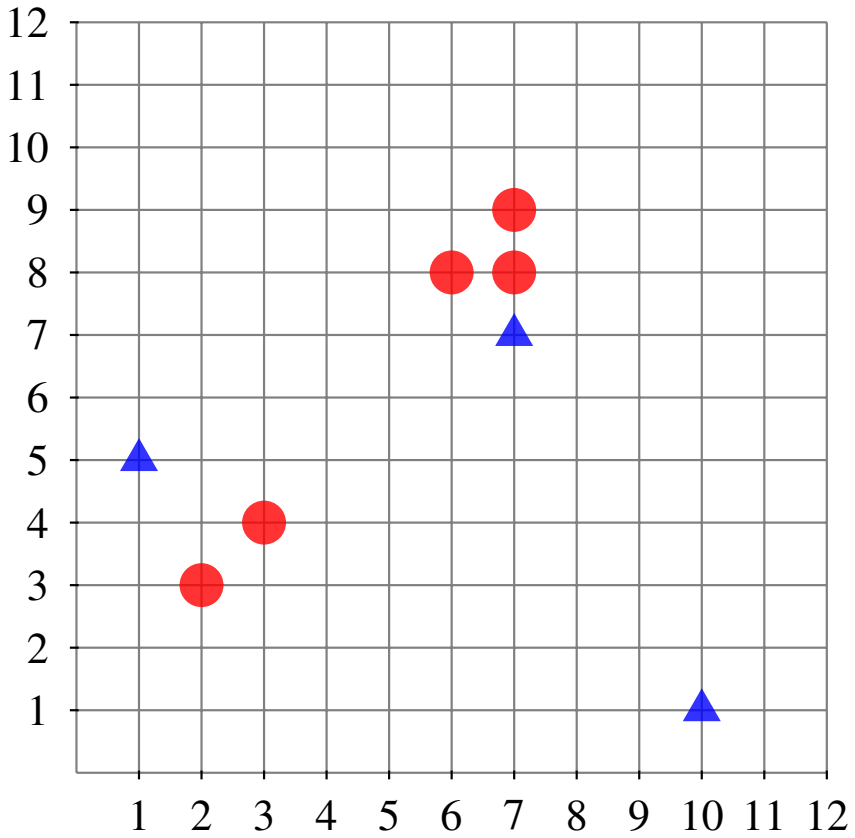
$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)^T V\Omega(V\Omega)^T(x-\mu) \\ &= \left\langle (V\Omega)^T(x-\mu), (V\Omega)^T(x-\mu) \right\rangle \\ &= L_2((V\Omega)^T(x-\mu))^2 \end{aligned}$$

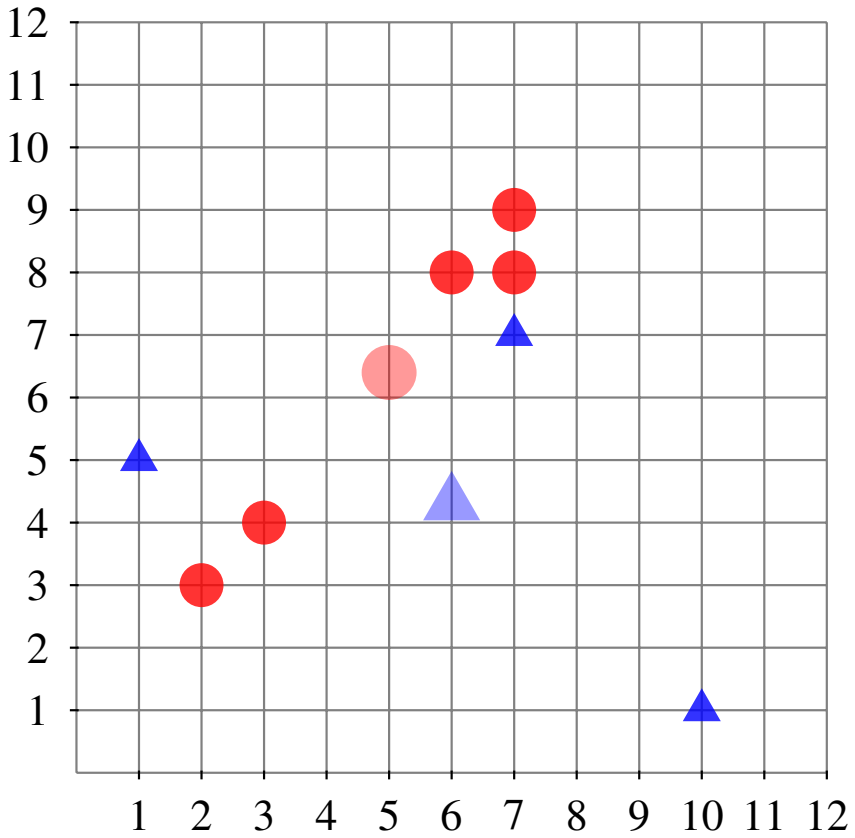
L_2 ist die L_2-Norm (Euclidische Distanz $d(x,y) = L_2(x-y)$!)

Mahalanobis $\approx$ Euclidische Distanz nach PCA!

k-Means Clustering – Lloyd/Forgy Algorithmus



k-Means Clustering – Lloyd/Forgy Algorithmus

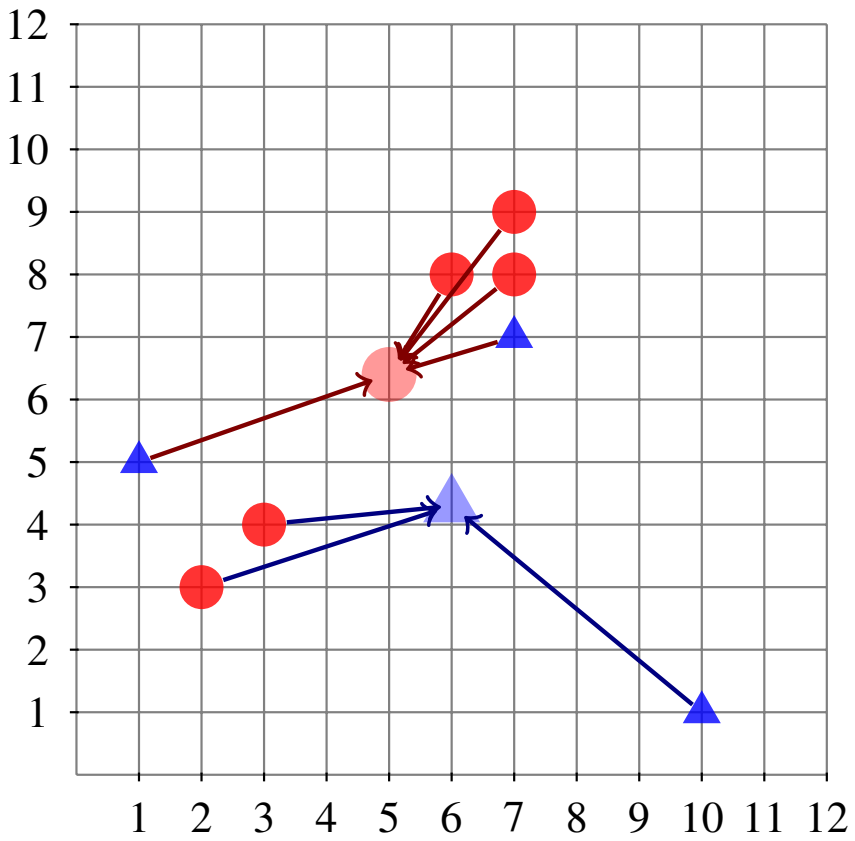


Zentroide neu berechnen:

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

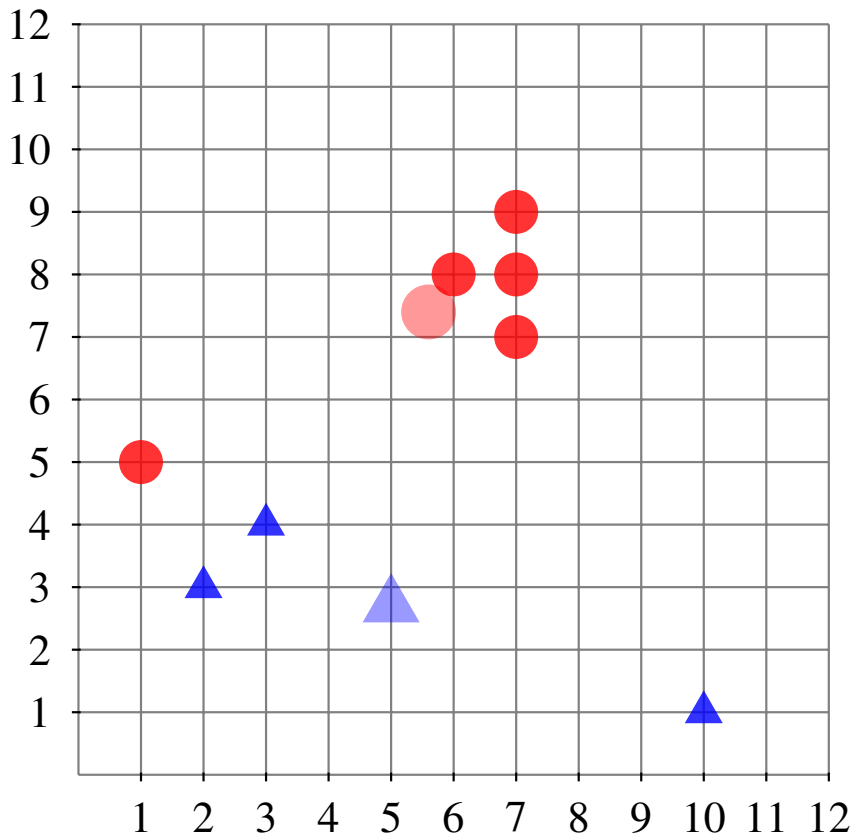
$$\mu \approx (5.0, 6.4)$$

k-Means Clustering – Lloyd/Forgy Algorithmus



Punkte neu zuordnen

k-Means Clustering – Lloyd/Forgy Algorithmus

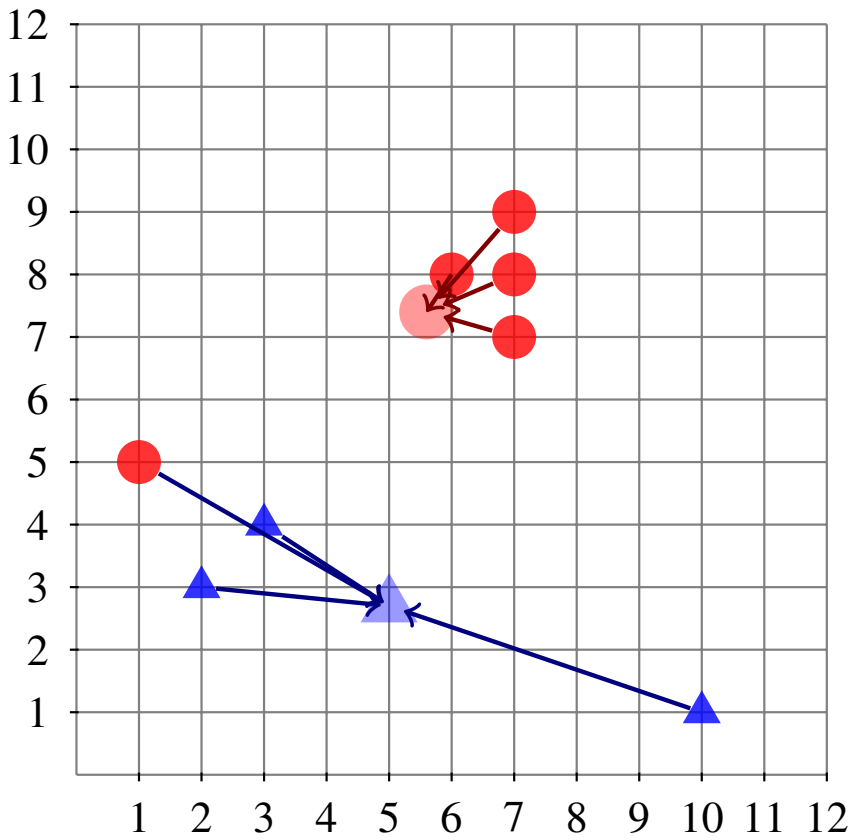


Zentroide neu berechnen:

$\mu \approx (5.0, 2.7)$

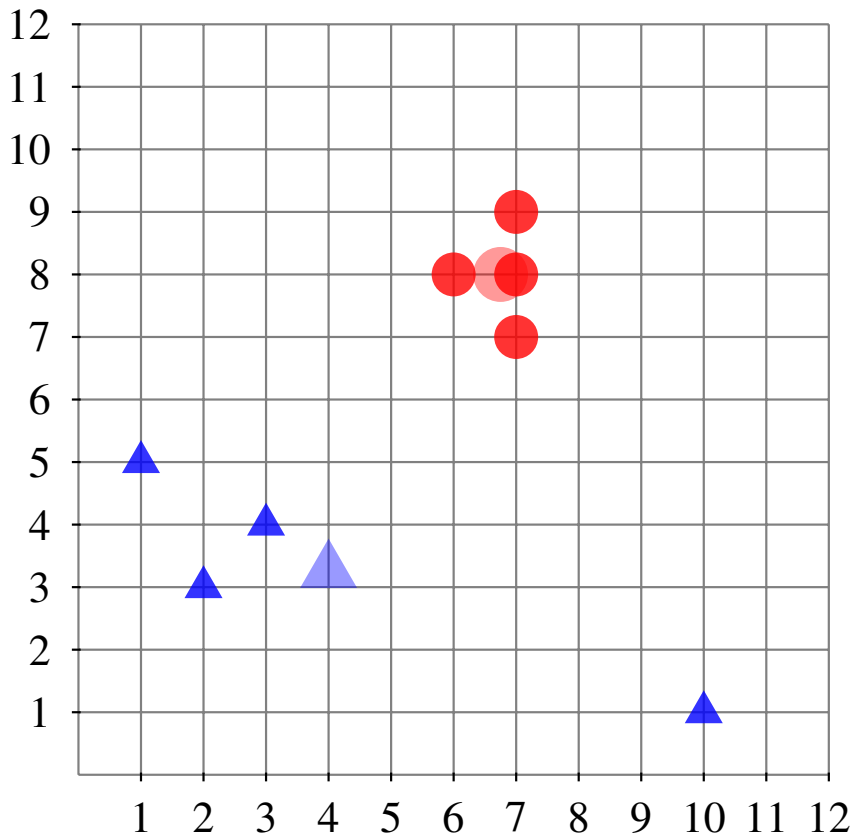
$\mu \approx (5.6, 7.4)$

k-Means Clustering – Lloyd/Forgy Algorithmus



Punkte neu zuordnen

k-Means Clustering – Lloyd/Forgy Algorithmus

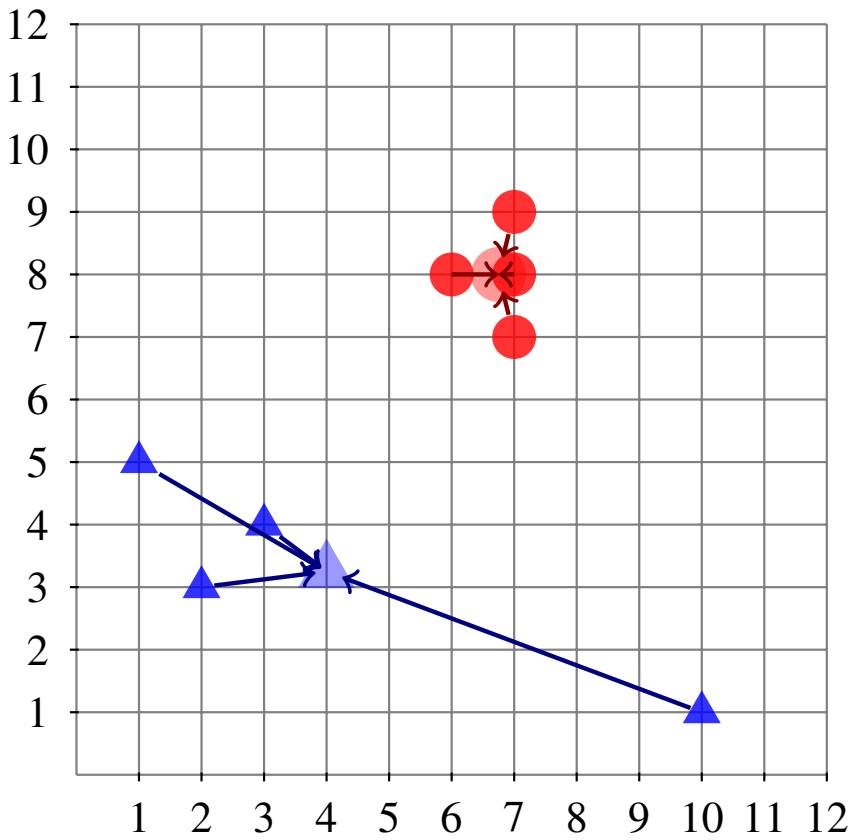


Zentroide neu berechnen:

$$\mu \approx (4.0, 3.25)$$

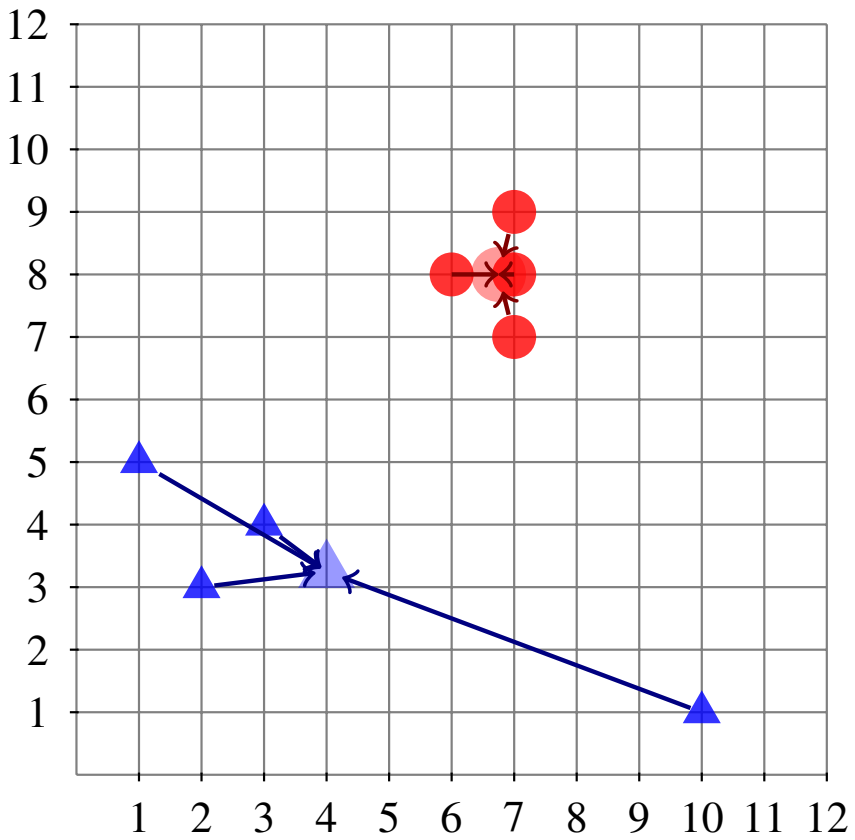
$$\mu \approx (6.75, 8.0)$$

k-Means Clustering – Lloyd/Forgy Algorithmus



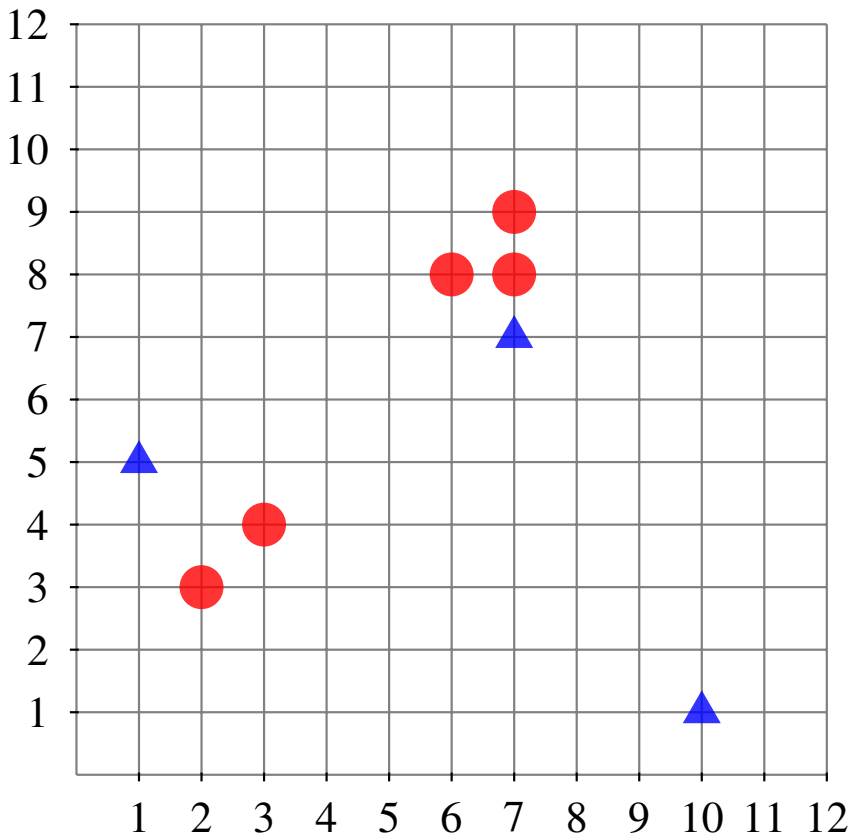
Punkte neu zuordnen

k-Means Clustering – Lloyd/Forgy Algorithmus

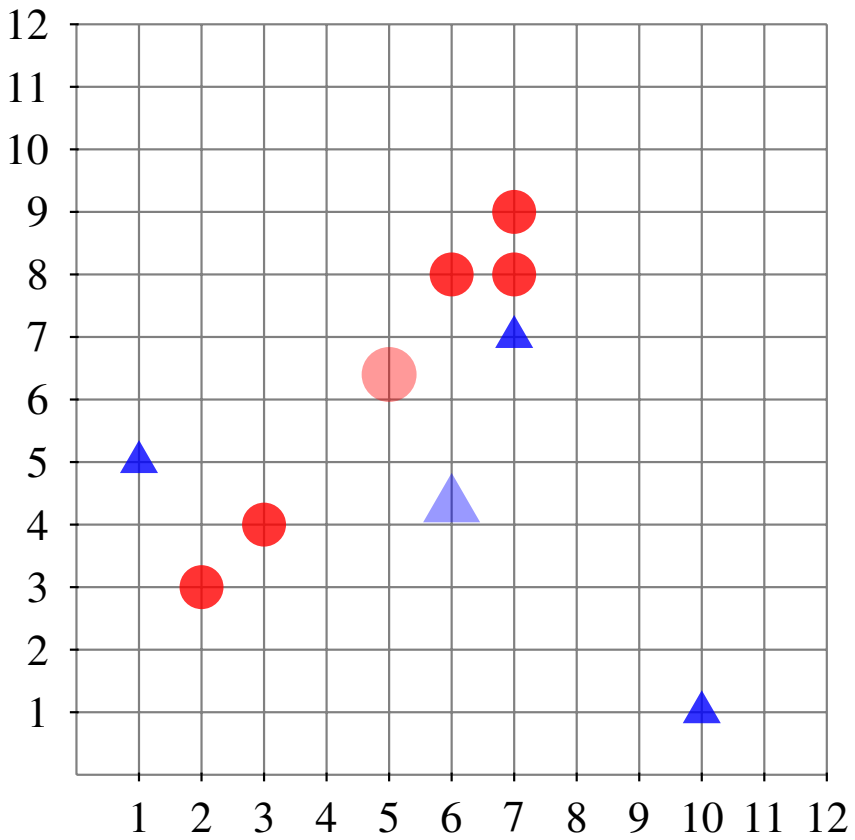


Punkte neu zuordnen
Keine Änderung
Konvergenz!

k-Means Clustering – MacQueen Algorithmus



k-Means Clustering – MacQueen Algorithmus

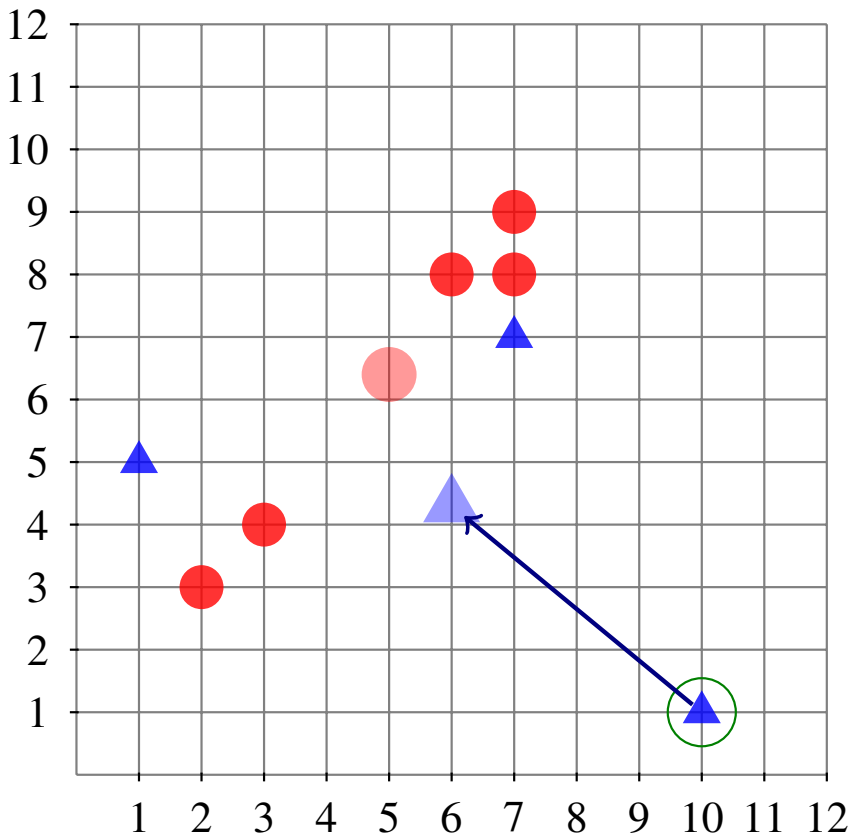


Zentroide (z.B.: aus vorheriger Iteration):

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

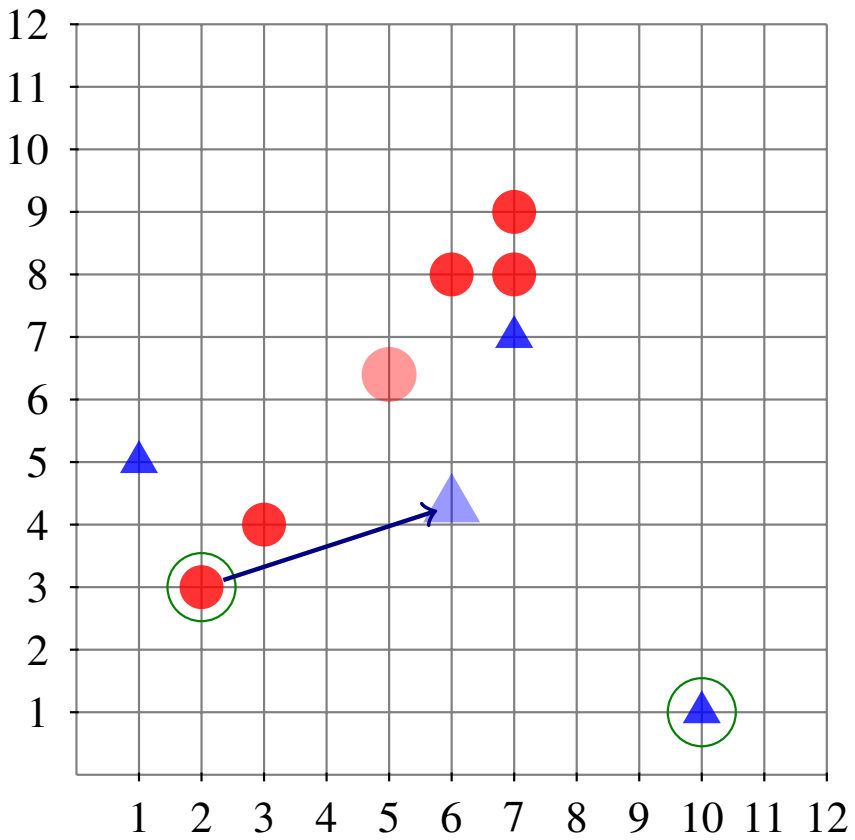
$$\mu \approx (5.0, 6.4)$$

k-Means Clustering – MacQueen Algorithmus



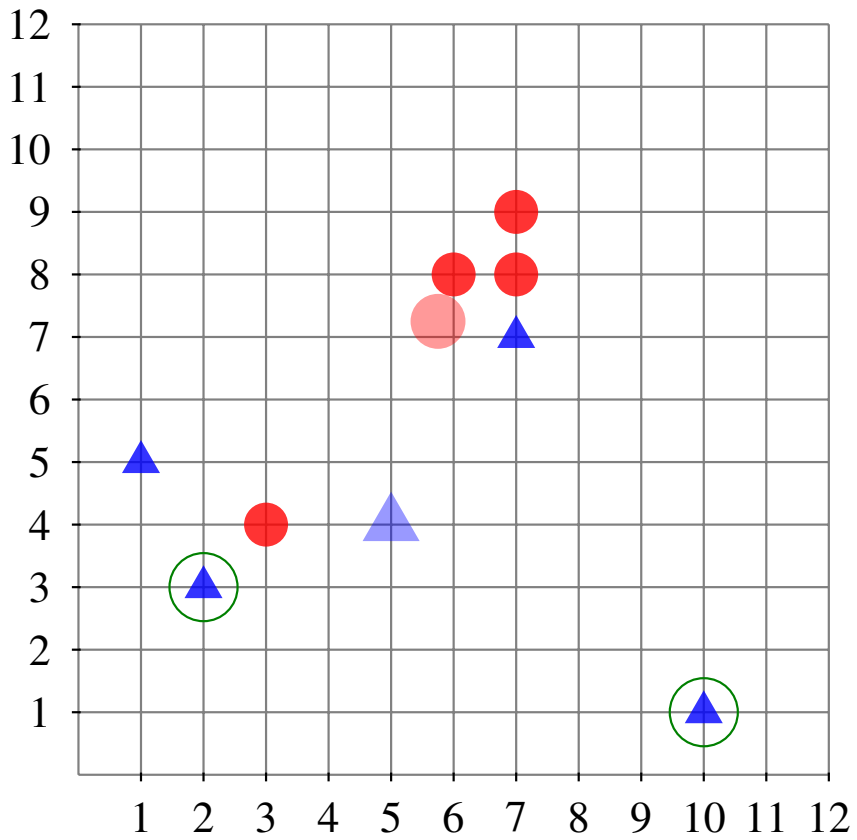
Ersten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus



Zweiten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

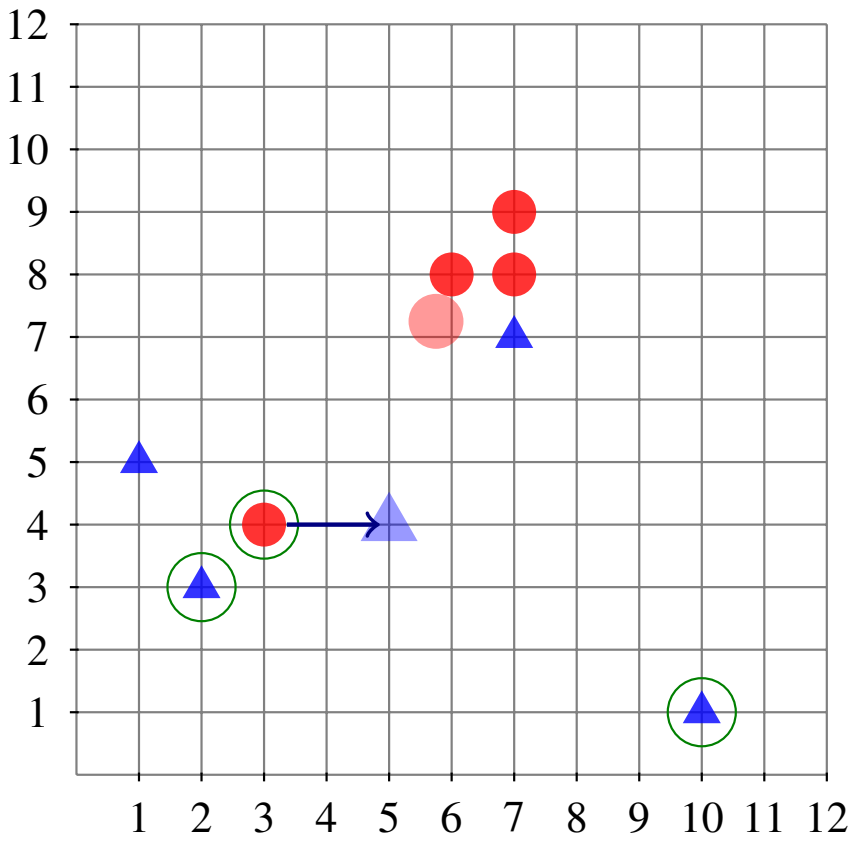


Zentroide aktualisieren:

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

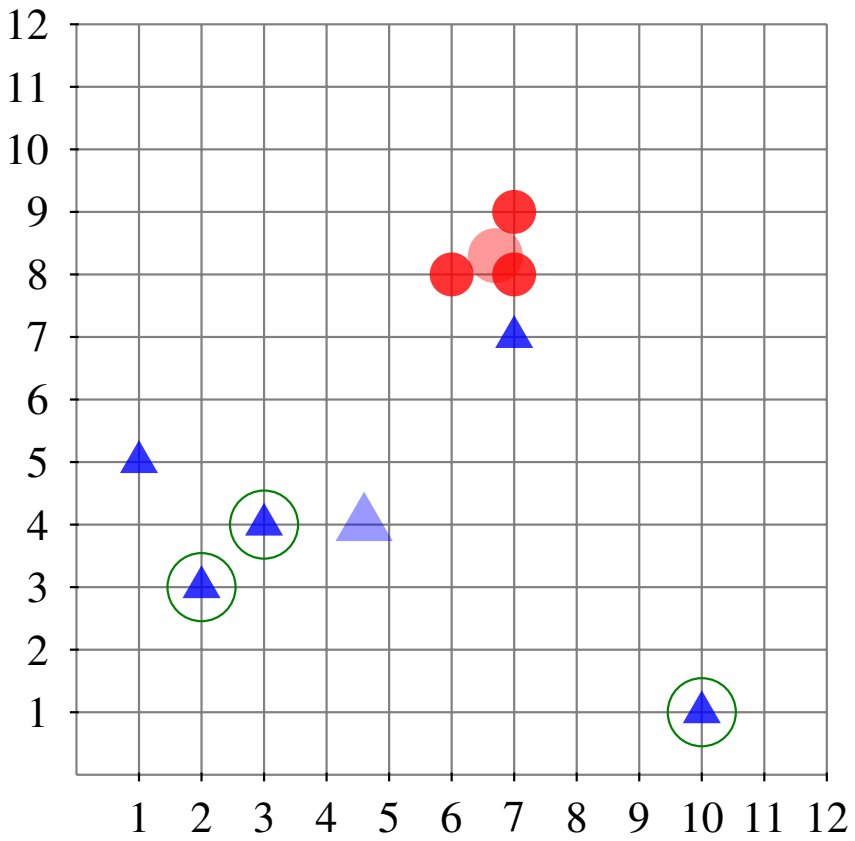
$$\mu \approx (5.75, 7.25)$$

k-Means Clustering – MacQueen Algorithmus



Dritten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

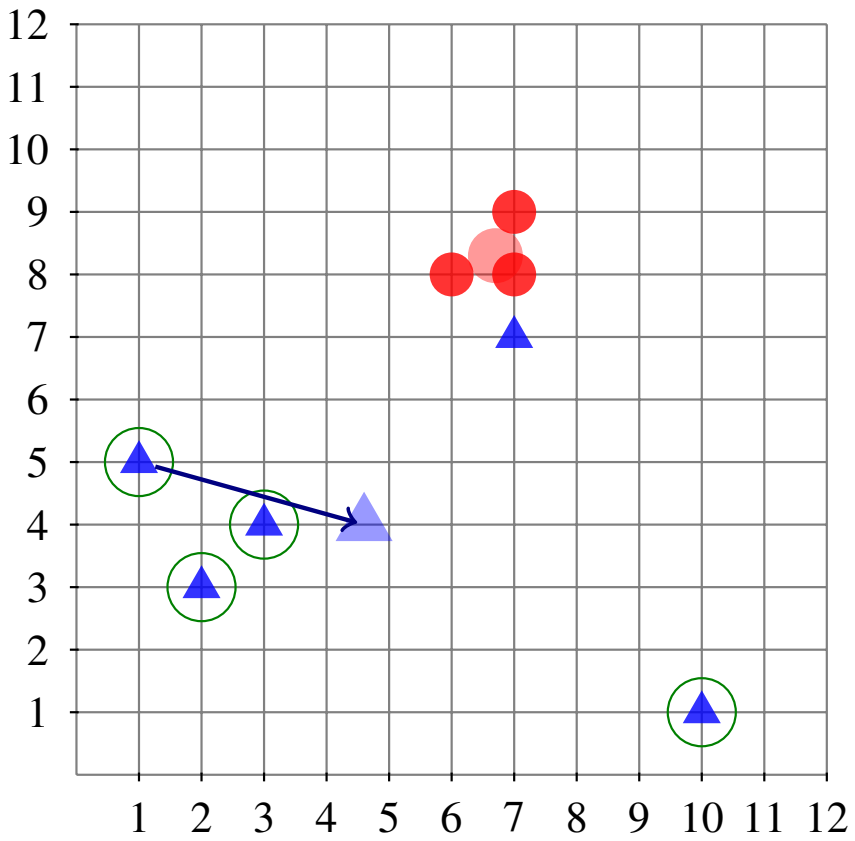


Zentroide aktualisieren:

$$\mu \approx (4.6, 4.0)$$

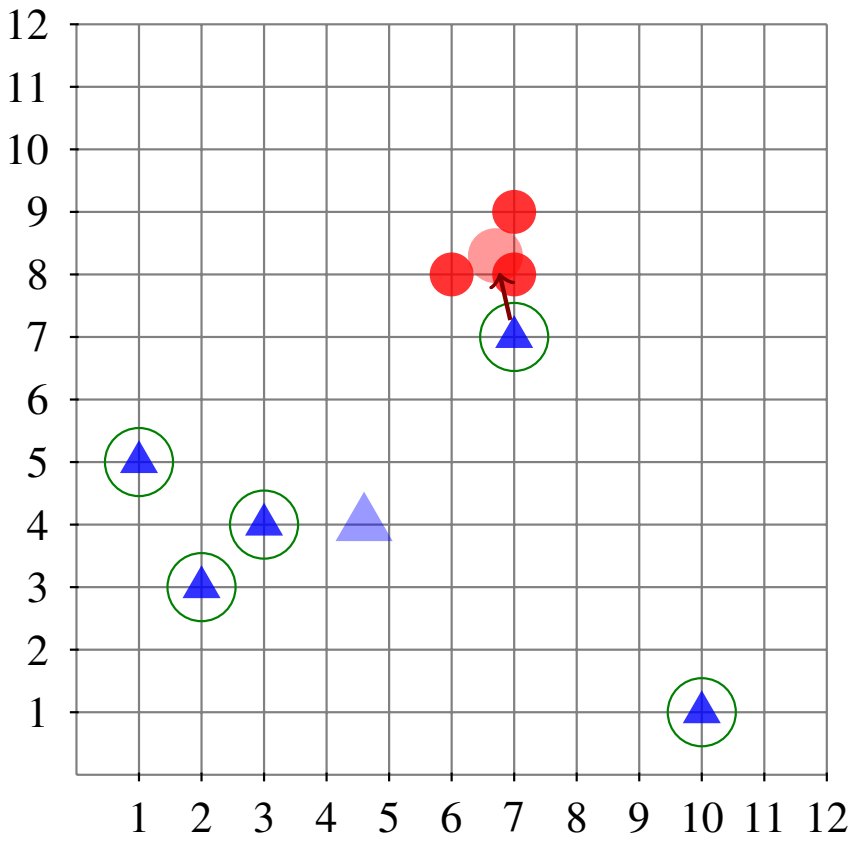
$$\mu \approx (6.7, 8.3)$$

k-Means Clustering – MacQueen Algorithmus



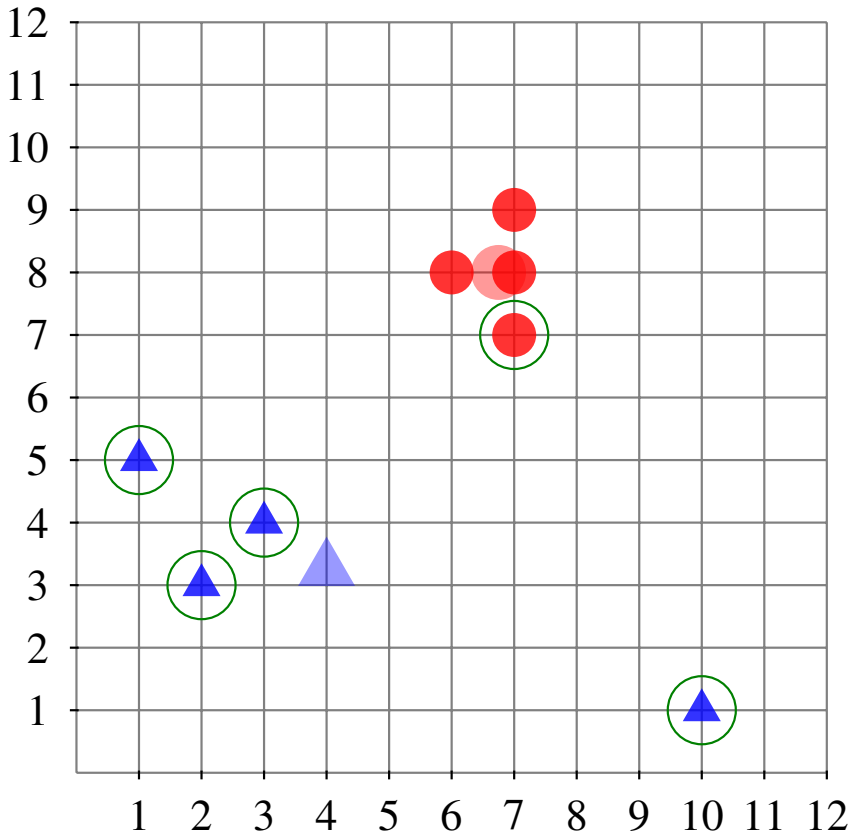
Vierten Punkt neu zuordnen

k-Means Clustering – MacQueen Algorithmus



Fünften Punkt neu zuordnen

k-Means Clustering – MacQueen Algorithmus

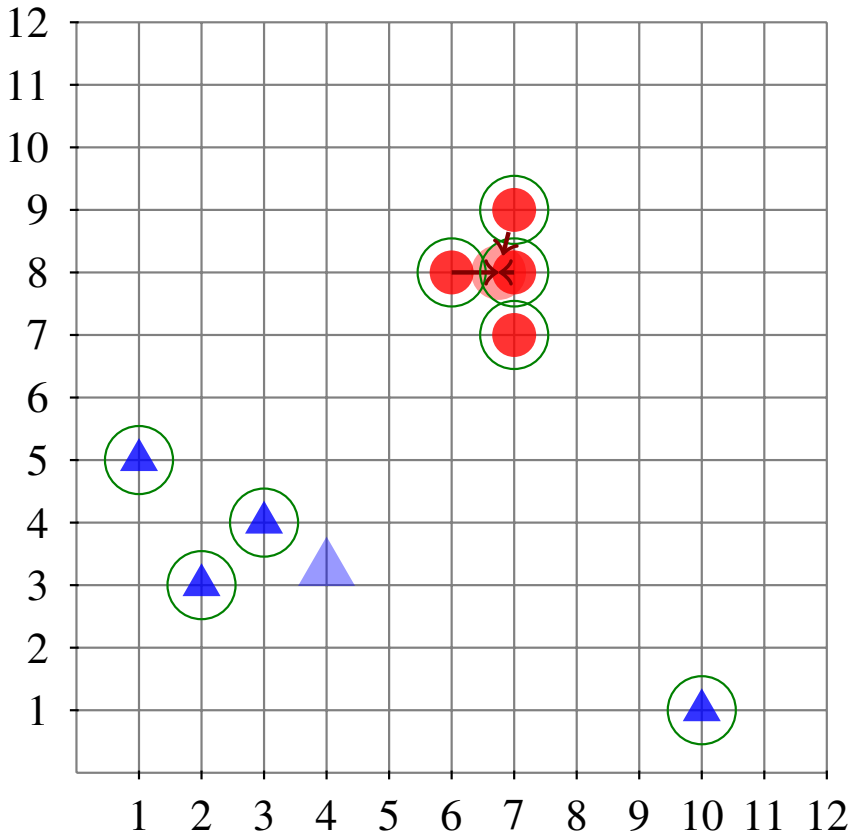


Zentroide aktualisieren:

$$\mu \approx (4.0, 3.25)$$

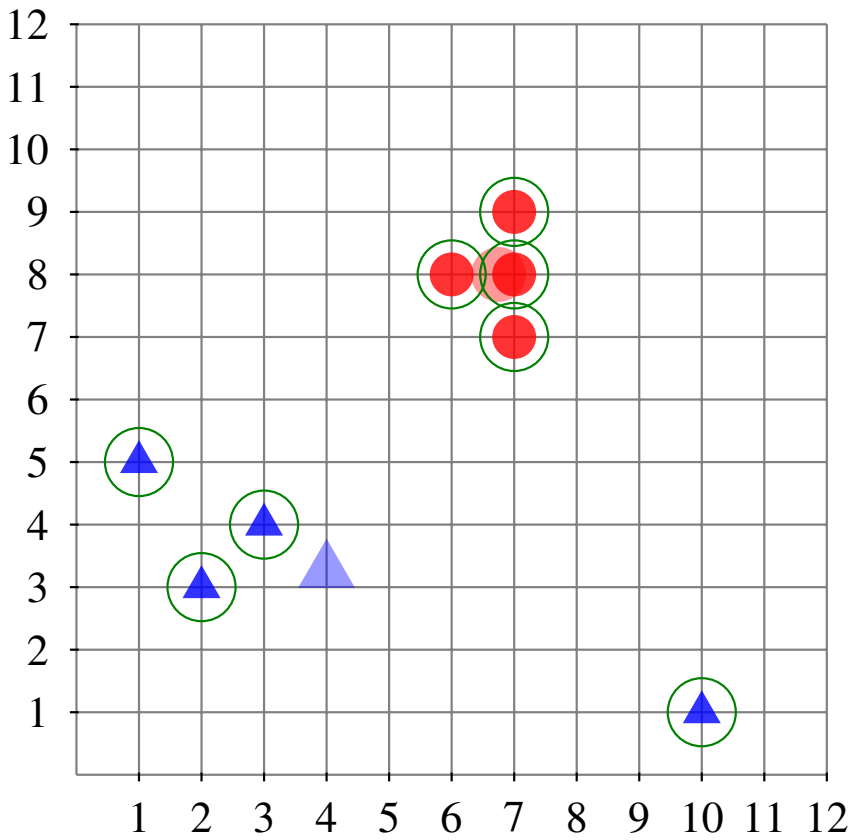
$$\mu \approx (6.75, 8.0)$$

k-Means Clustering – MacQueen Algorithmus



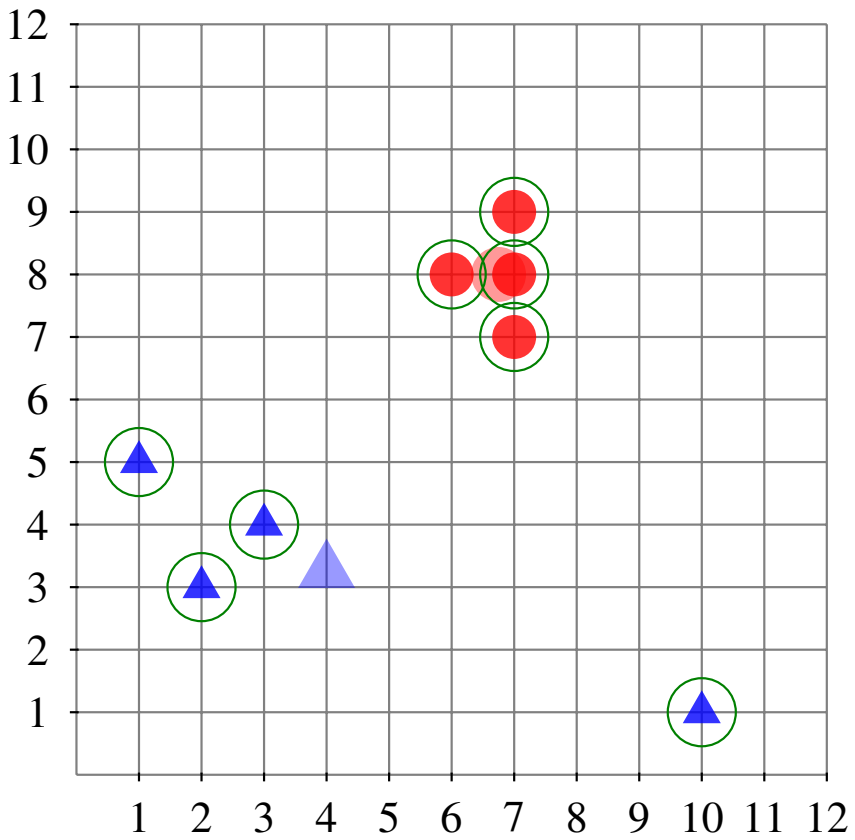
Weitere Punkte neu zuordnen

k-Means Clustering – MacQueen Algorithmus



Weitere Punkte neu zuordnen
ggf. Weitere Iterationen

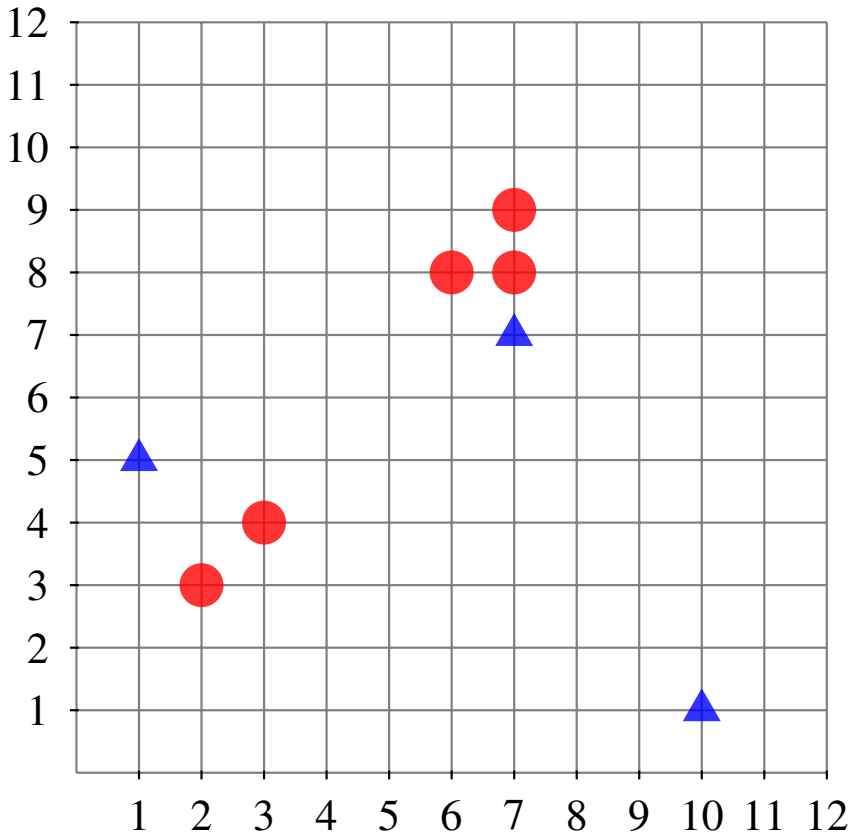
k-Means Clustering – MacQueen Algorithmus



Weitere Punkte neu zuordnen
ggf. Weitere Iterationen
Konvergenz

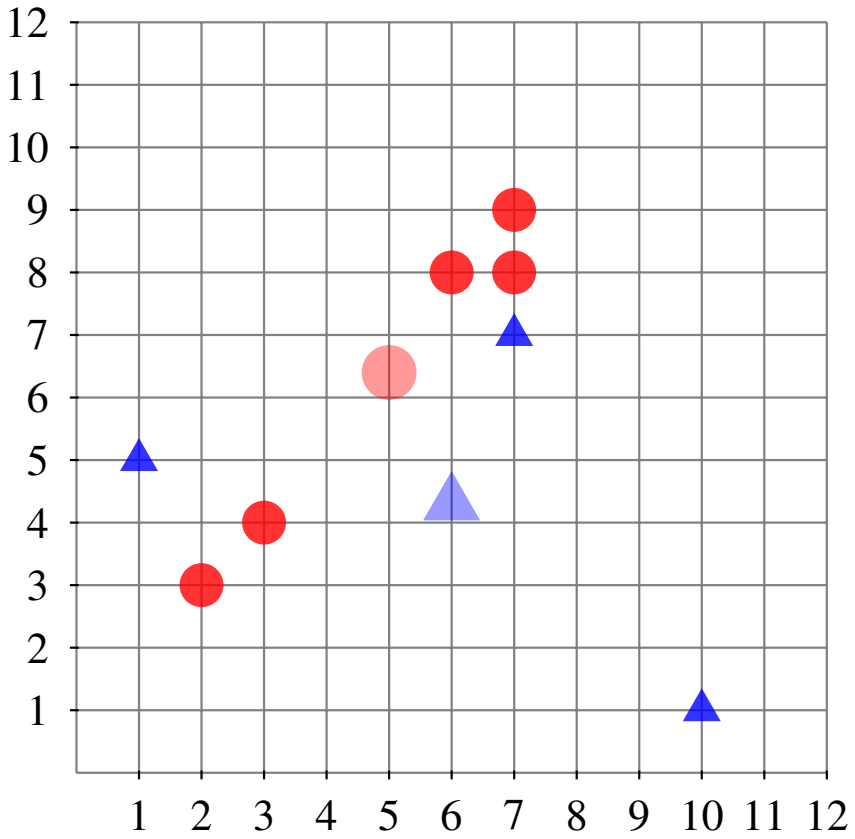
k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



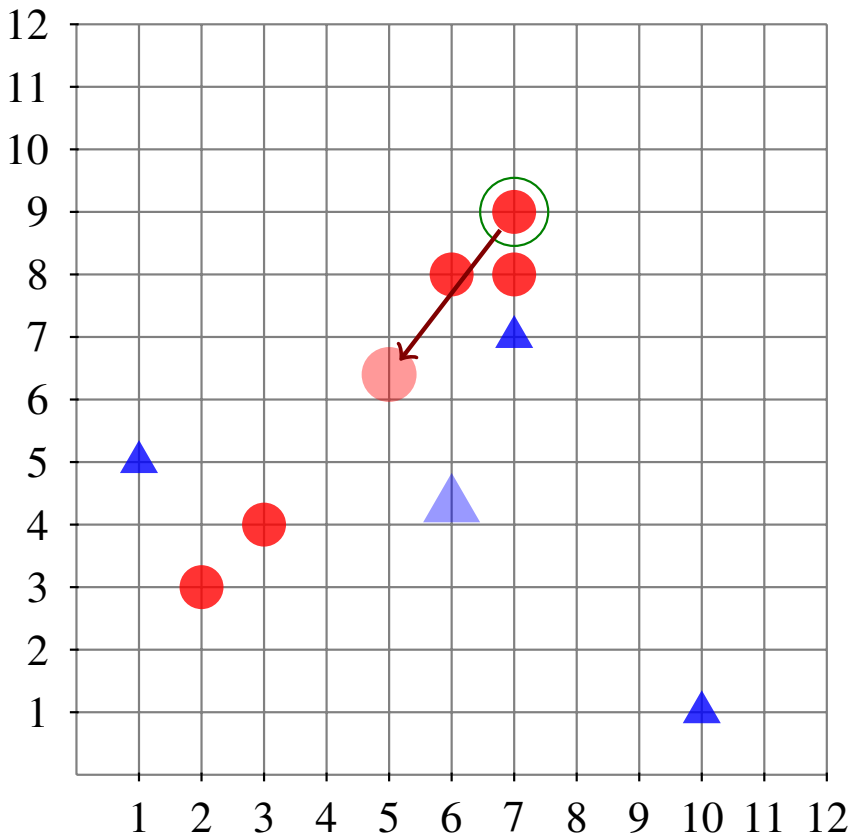
Zentroide (z.B.: aus vorheriger Iteration):

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

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

k-Means Clustering – MacQueen Algorithmus

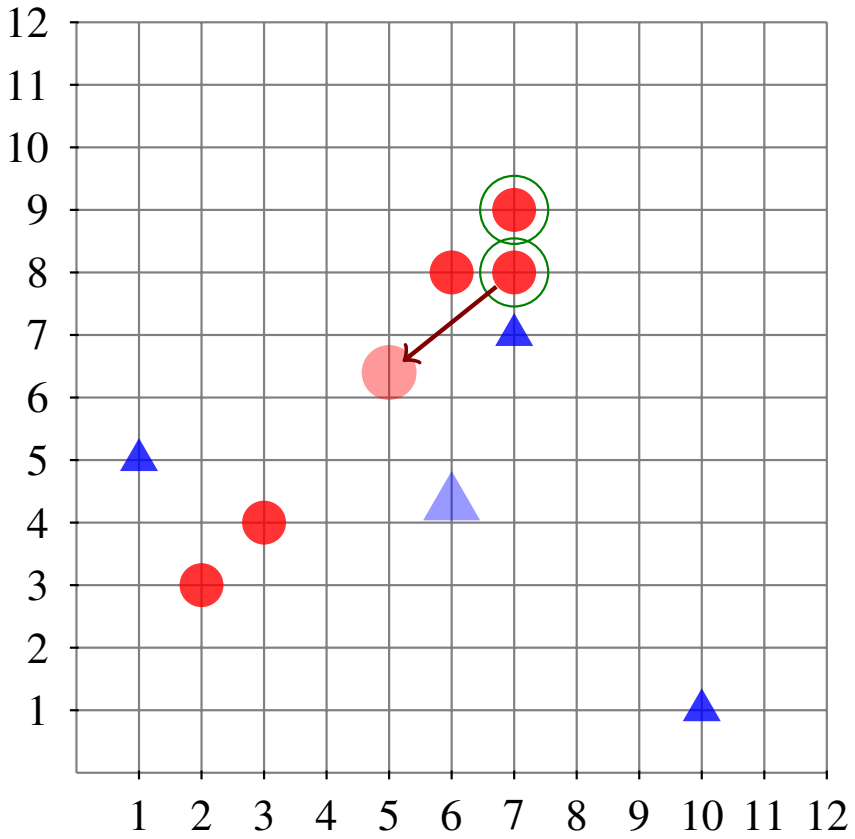
Alternativer Ablauf – andere Reihenfolge



Ersten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

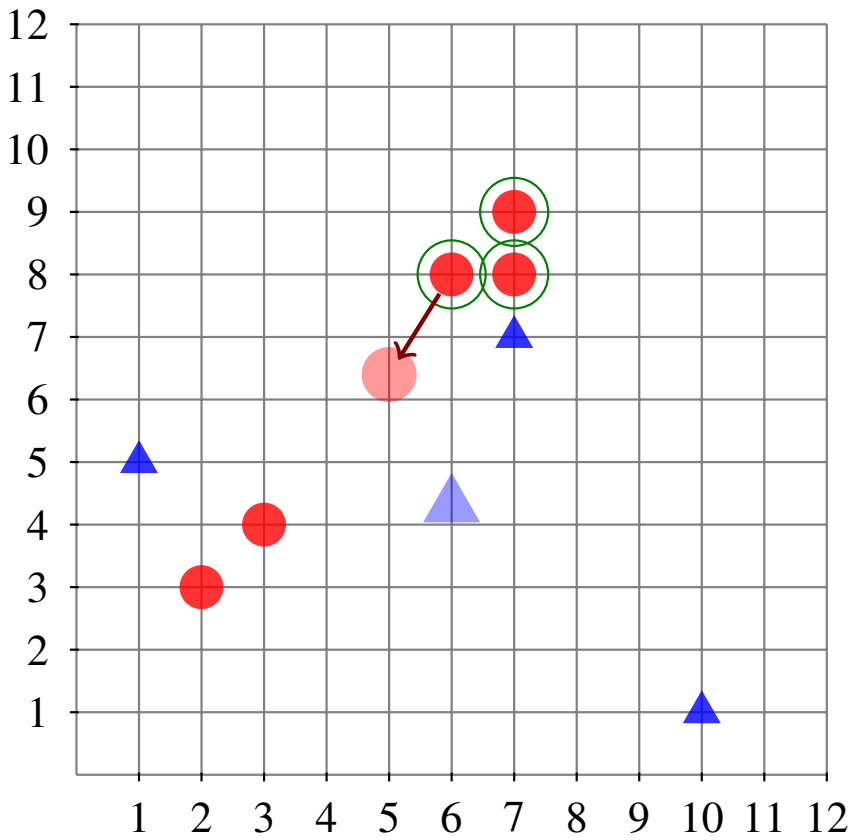
Alternativer Ablauf – andere Reihenfolge



Zweiten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

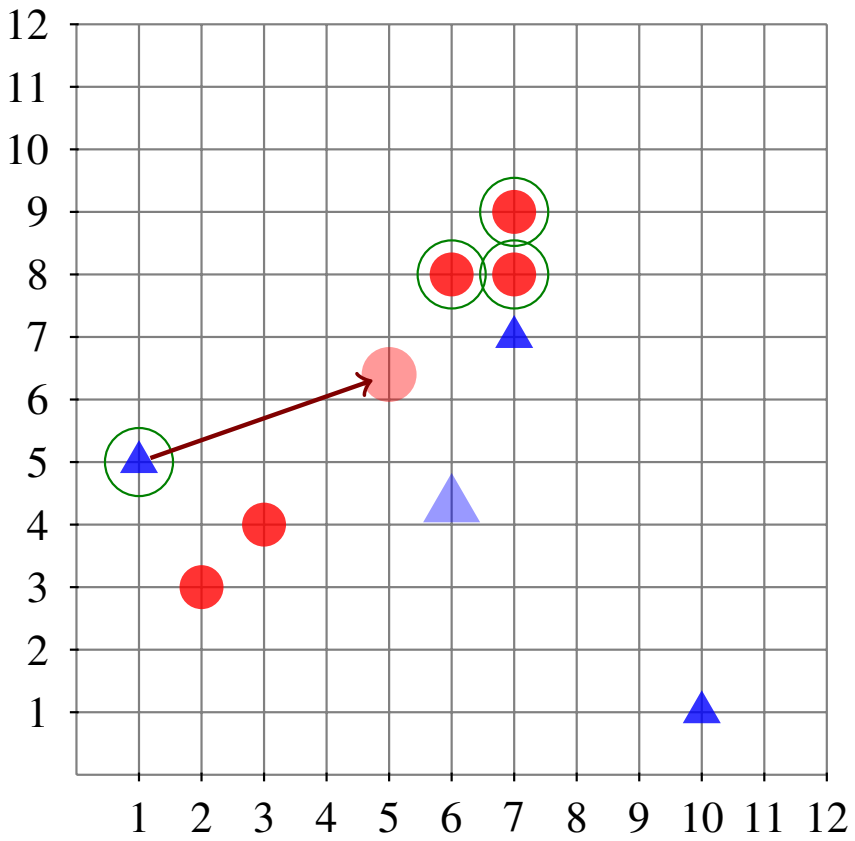
Alternativer Ablauf – andere Reihenfolge



Dritten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

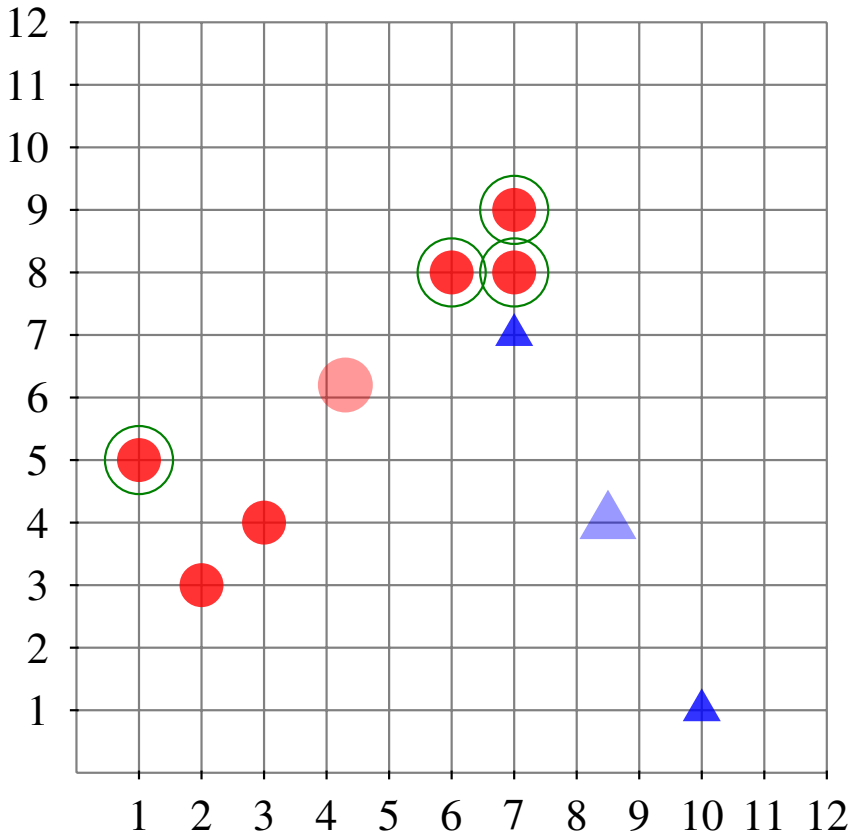
Alternativer Ablauf – andere Reihenfolge



Vierten Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



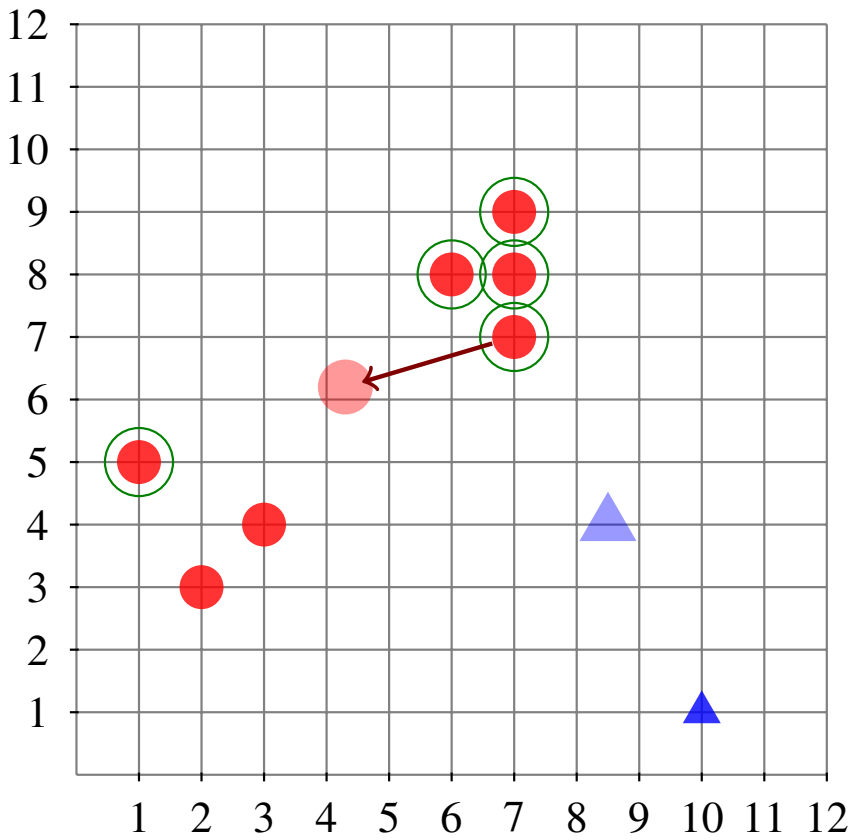
Zentroide aktualisieren:

$$\mu \approx (4.0, 8.5)$$

$$\mu \approx (4.3, 6.2)$$

k-Means Clustering – MacQueen Algorithmus

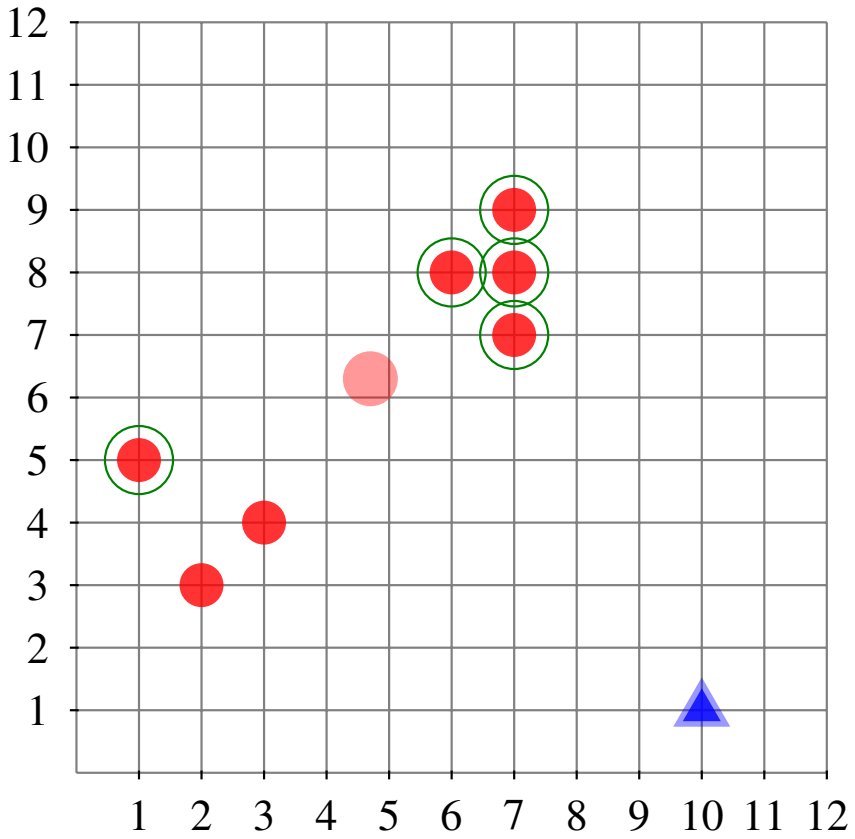
Alternativer Ablauf – andere Reihenfolge



Fünften Punkt zuordnen

k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



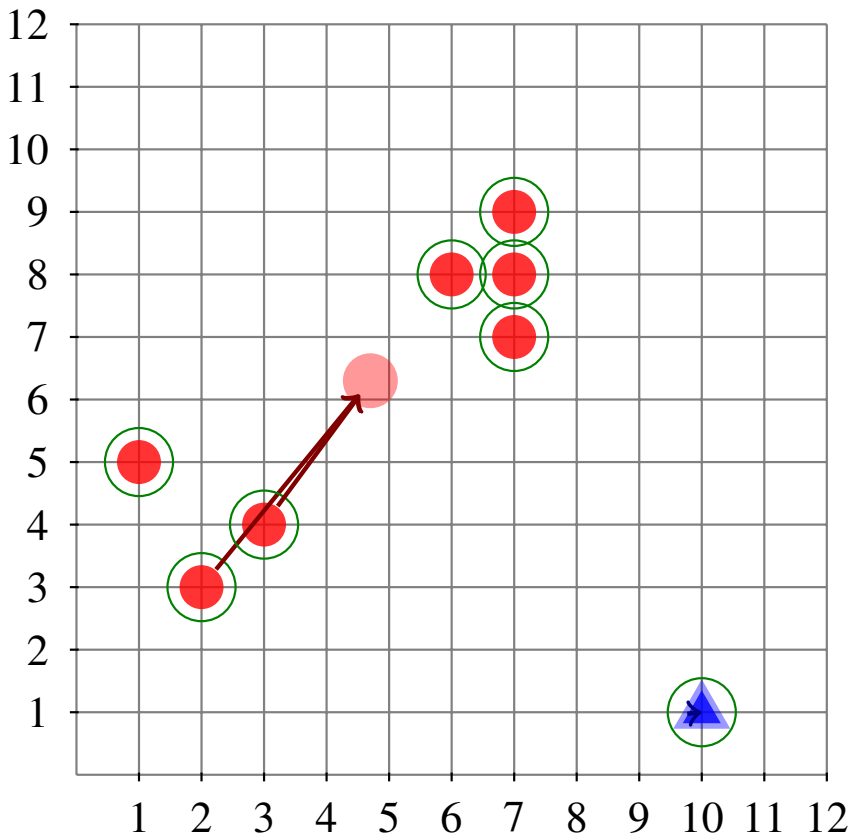
Zentroide aktualisieren:

$$\mu \approx (10.0, 1.0)$$

$$\mu \approx (4.7, 6.3)$$

k-Means Clustering – MacQueen Algorithmus

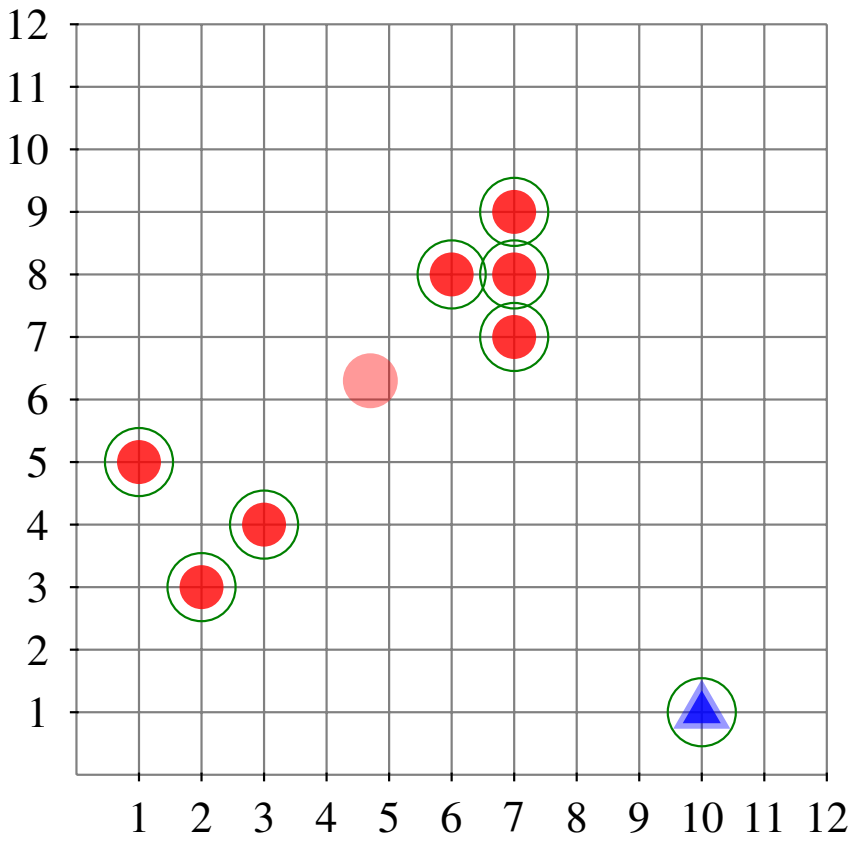
Alternativer Ablauf – andere Reihenfolge



Weitere Punkte neu zuordnen

k-Means Clustering – MacQueen Algorithmus

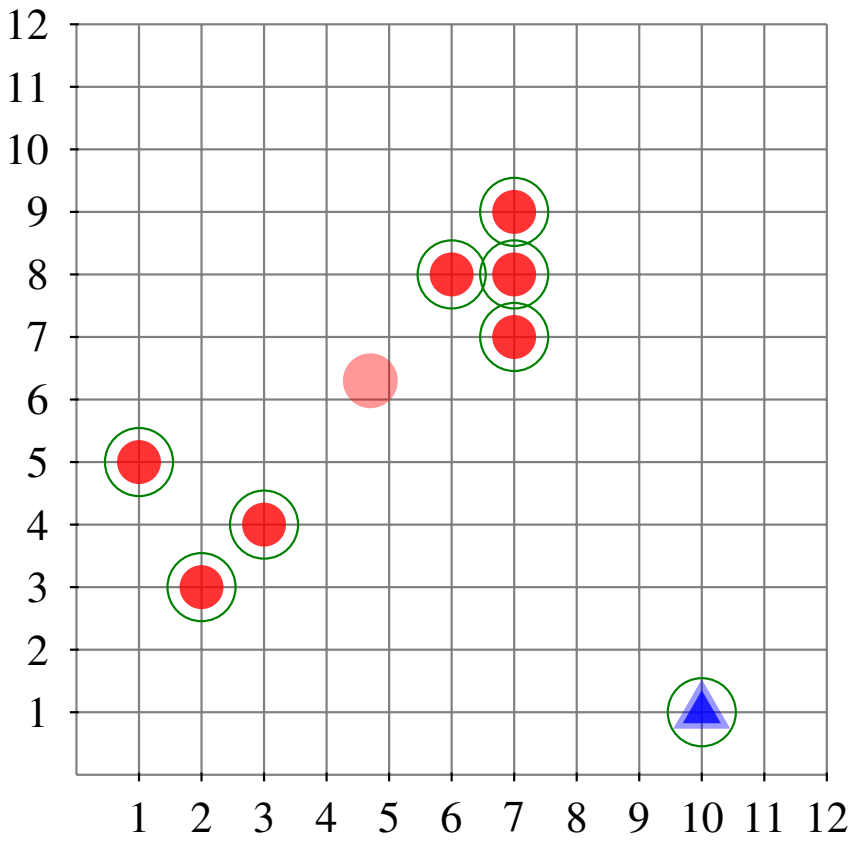
Alternativer Ablauf – andere Reihenfolge



Weitere Punkte neu zuordnen
ggf. Weitere Iterationen

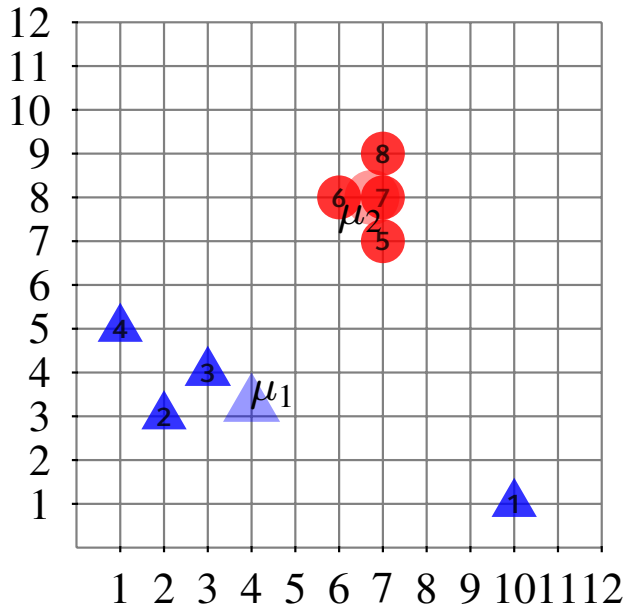
k-Means Clustering – MacQueen Algorithmus

Alternativer Ablauf – andere Reihenfolge



Weitere Punkte neu zuordnen
ggf. Weitere Iterationen
Konvergenz

k-Means Clustering – Qualität



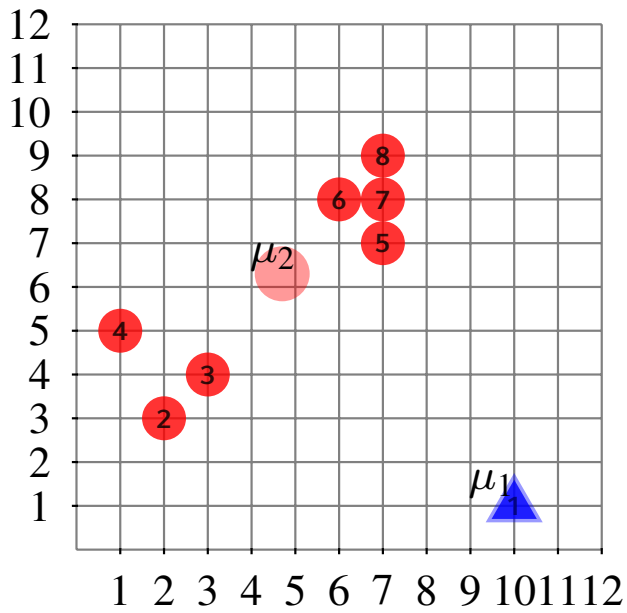
Erste Lösung: $TD^2 = 61\frac{1}{2}$

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

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

(Hier: sum-of-squares $\equiv$ quadrierte Euklidische Distanz – mit Manhattan^2 kommen andere aber ähnliche Zahlen heraus)

k-Means Clustering – Qualität



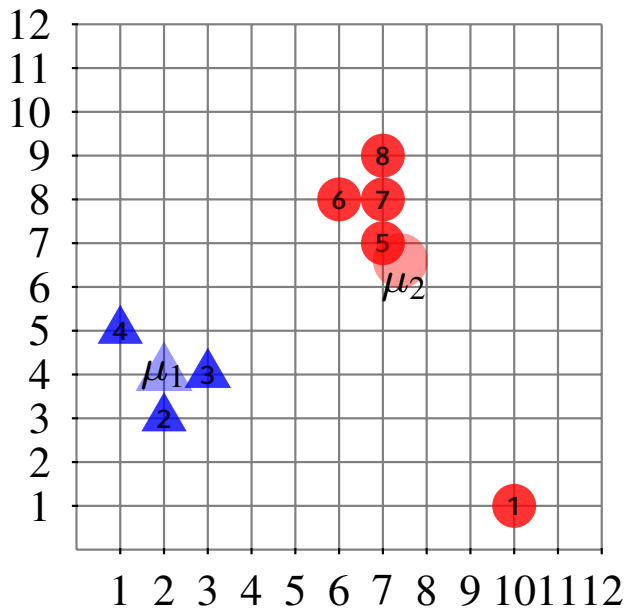
$$d^2(\mu_1, p_1) = |10-10|^2 + |1-1|^2 = 0$$
$$TD^2(C_1) = 0$$

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

Erste Lösung: $TD^2 = 61\frac{1}{2}$
Zweite Lösung: $TD^2 \approx 72.68$

(Hier: sum-of-squares $\equiv$ quadrierte Euklidische Distanz – mit Manhattan^2 kommen andere aber ähnliche Zahlen heraus)

k-Means Clustering – Qualität



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

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

Erste Lösung: $TD^2 = 61\frac{1}{2}$
Zweite Lösung: $TD^2 \approx 72.68$
Optimale Lösung: $TD^2 = 54\frac{2}{5}$

(Hier: sum-of-squares ≡ quadrierte Euklidische Distanz – mit Manhattan^2 kommen andere aber ähnliche Zahlen heraus)

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 (ggf. aber auch garantierte 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!