Maschinelles Lernen 09

Prof. Dr. David Spieler – david.spieler@hm.edu

Hochschule München

1. Oktober 2019

Clustering

Clustering

Clustering

Ziel von Clustering Methoden ist es, den Datensatz in möglichst ähnlich Partitionen zu zerteilen, d.h.

- Datenpunkte *innerhalb* einer solchen Partition, auch Cluster genannt unterscheiden sich *wenig*,
- Datenpunkte *verschiedener* Cluster unterscheiden sich *mehr*.

Clustering Methoden sind auf ein Maß von (Un-)Ähnlichkeit angewiesen, was meist nur für reellwertige Daten wohldefiniert ist. Wir beschränken uns daher auf Probleme mit

$$\mathcal{D} \subseteq \mathbb{R}^d.$$

Clustering

Unterschied zwischen PCA und Clustering:

- PCA versucht eine Darstellung der Datenpunkte in einem Raum geringerer Dimension zu finden. Sie ermöglicht potentiell eine bessere Trennung der Daten, trennt sie jedoch nicht direkt.
- Clustering versucht tatsächliche Trennungen zu finden.

K-Means Clustering

K-Means Clustering

K-Means Clustering

Beim K-Means Clustering wird zunächst die Anzahl der Cluster $K \in \mathbb{N}$ als *Hyperparameter* vorgegeben. Anschließend wird automatisiert versucht eine möglichst *gute* Partitionierung der n Datenpunkte zu finden, d.h. unter den Bedingungen

- $\mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_K = \{1, \ldots, n\}$ und
- $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ für alle $i \neq j$,

minimieren wir ein Maß $W(\mathcal{C})$ für die Varianz innerhalb der Cluster, d.h.

$$\arg \min_{\mathcal{C}_1, \ldots, \mathcal{C}_K} \sum_{k=1}^{K} W(\mathcal{C}_k).$$

K-Means Clustering

Eine übliche Wahl für das Varianzmaß ist die quadrierte euklidische Distanz

$$W(\mathcal{C}_k) = \frac{1}{|\mathcal{C}_k|} \sum_{i,j \in \mathcal{C}_k} ||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}||_2^2$$

welches umgeformt werden kann zu

$$W(\mathcal{C}_k) = 2 \sum_{i \in \mathcal{C}_k} ||\mathbf{x}^{(i)} - \bar{\mathbf{x}}^{(k)}||_2^2$$

mit

$$\bar{\mathbf{x}}^{(k)} = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbf{x}^{(i)},$$

dem Cluster-Schwerpunkt von Cluster $\mathcal{C}_k$.

K-Means Clustering

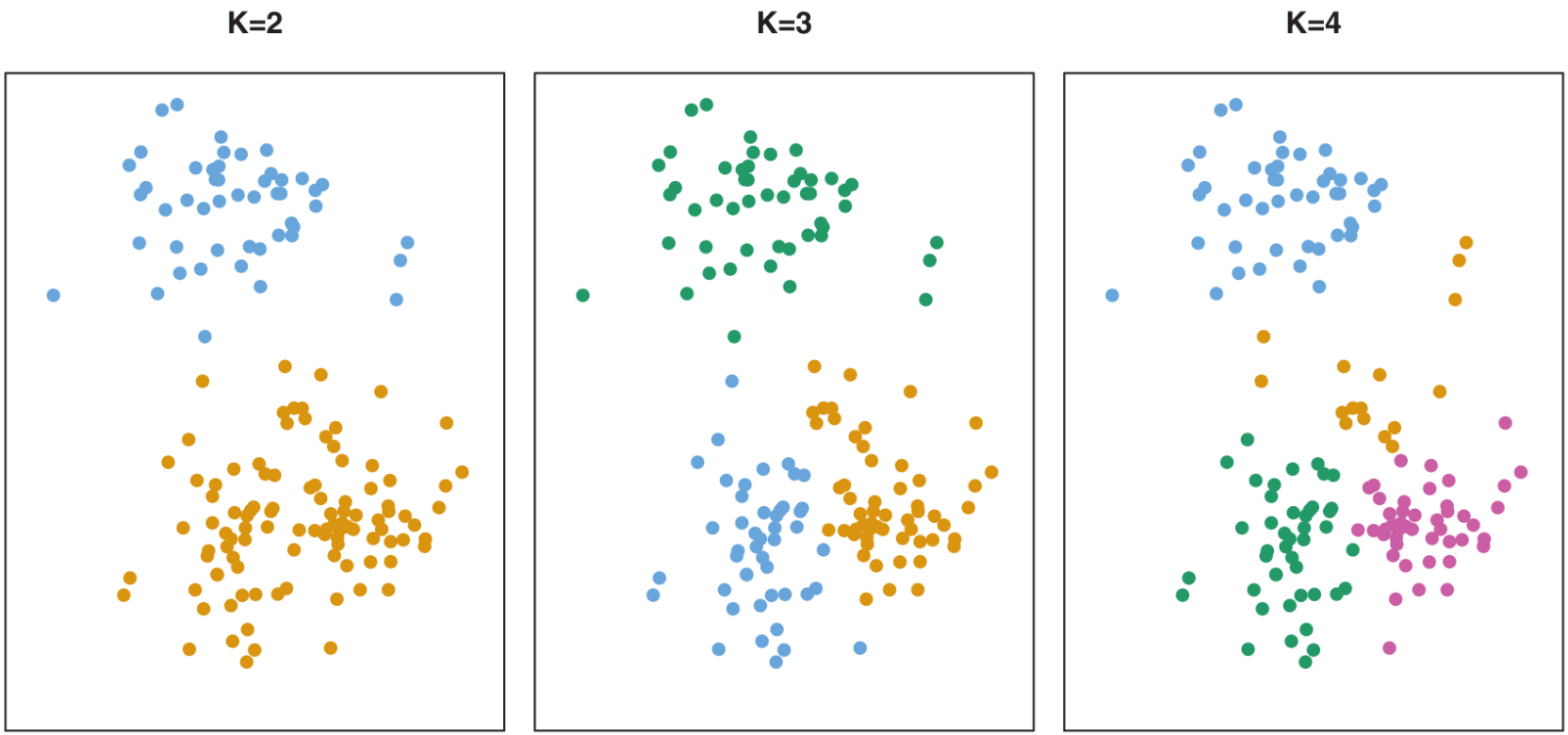

Abbildung 1: Beispielsdatensatz mit 150 Datenpunkten geclustered in $k = 2, 3, 4$ Cluster mit Hilfe von K-Means. Abbildung entnommen aus [JWHT14].

K-Means Clustering

Für n Datenpunkte und K Cluster gibt es annähernd K^n verschiedene Möglichkeiten der Partitionierung. Die Suche nach dem globalen Optimum durch Probieren aller Möglichkeiten (Brute-Force) ist daher meist nicht möglich bzw. sinnvoll.
Jedoch bieten bereits lokale Optima meist gute Cluster, für deren Suche ein effizienter Algorithmus existiert.

K-Means Clustering

Algorithm 1 kmeans_cluster($\mathcal{D}$, K)

for $i \in \{1, \ldots, n\}$ **do**
  cluster(i) = random(1,K)
end for
while any cluster(i) has changed **do**
  for $k \in \{1, \ldots, K\}$ **do**
    $\mathcal{C}_k = \{i \mid \text{cluster}(i) = k\}$
    $\bar{\mathbf{x}}^{(k)} = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbf{x}^{(i)}$
  end for
  for $i \in \{1, \ldots, n\}$ **do**
    $\text{cluster}(i) = \arg\min_{1,\ldots,K} ||\mathbf{x}^{(i)} - \bar{\mathbf{x}}^{(k)}||_2^2$
  end for
end while
return cluster

K-Means Clustering

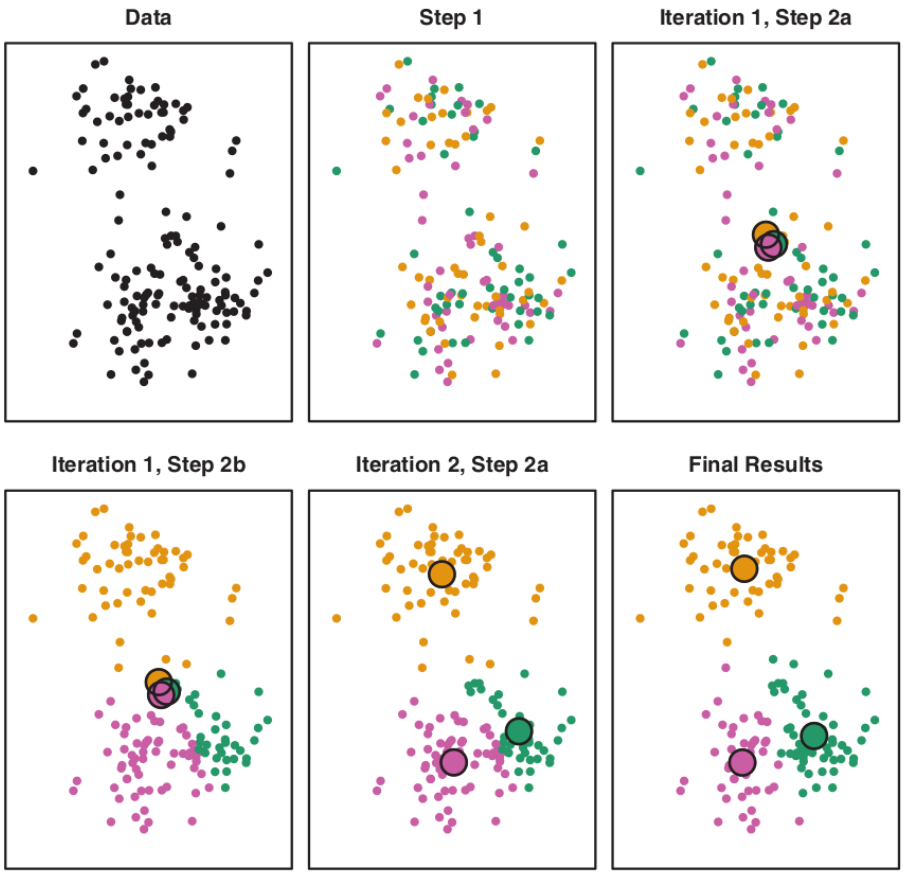

Abbildung 2: Schrittweise Ausführung des K-Means Clustering Algorithmus auf das Beispiel ($K = 3$). Abbildung entnommen aus [JWHT14].

K-Means Clustering

Da der K-Means Clustering Algorithmus lediglich ein lokales Optimum ausgehend von einer zufälligen initialen Clusterzuweisung liefert, sollte der Algorithmus mehrmals auf die Daten angewendet werden, um ein möglichst gutes Clustering zu finden.

K-Means Clustering

Abbildung 3: Wiederholte Ausführung des K-Means Clustering Algorithmus auf das Beispiel ($K = 3$) kann zu unterschiedlich guten Ergebnissen führen. Abbildung entnommen aus [JWHT14].

Hierarchische Clusteranalyse

Hierarchische Clusteranalyse

Hierarchische Clusteranalyse

Ein Nachteil von K-Means ist, dass die Anzahl der Cluster fest vorgegeben ist und vor Anwendung gewählt werden muss.
Methoden aus der Hierarchischen Clusteranalyse erstellen ein Dendrogramm, eine baumartige Repräsentation des Clusterings der Datenpunkte mit variierender Clusteranzahl.
Hierbei gibt es zwei Möglichkeiten der Erstellung:

- bottom-up (agglomerativ): Cluster starten als einzelne Datenpunkte und verschmelzen sukzessiv zu größeren Clustern
- top-down (divisiv): Ein Cluster mit allen Datenpunkten wird sukzessive geteilt

Wir beschäftigen uns hier lediglich mit einer bottom-up Methode.

Hierarchische Clusteranalyse

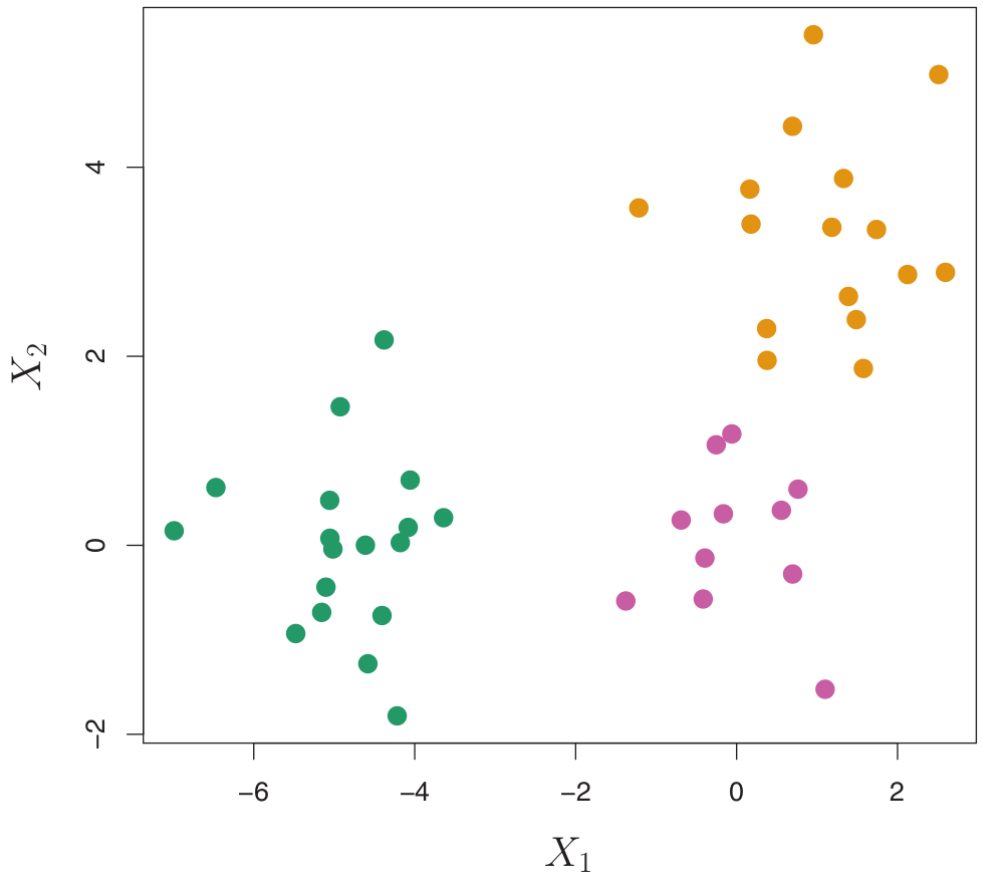

Abbildung 4: Beispielsdatensatz mit $K = 3$ echten Clustern. Abbildung entnommen aus [JWHT14].

Hierarchische Clusteranalyse

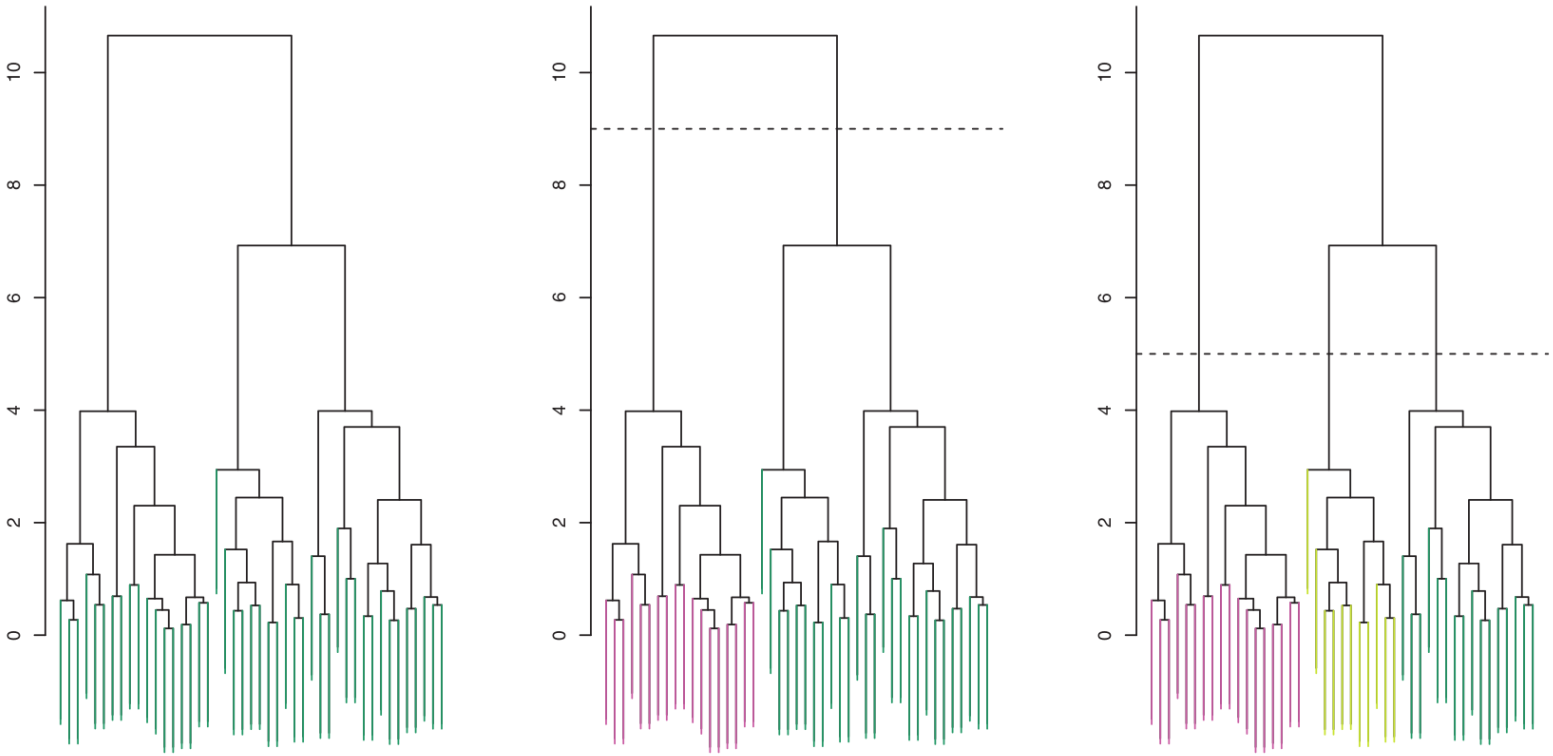

Abbildung 5: Agglomeratives Dendrogramm des Beispielsdatensatzes. Blätter entsprechen einzelnen Datenpunkten, Knoten (Datenpunkte oder Cluster) werden auf Höhe des Abstandsmaßes durch Linien zu einem neues Cluster verschmolzen. Abbildung entnommen aus [JWHT14].

Hierarchische Clusteranalyse

Hinweise:

- Ähnlichkeitsaussagen können nur vertikal aber nicht horizontal getroffen werden.
- In einem Dendrogramm kann durch Sichtbetrachtung ein Clustering für $K \in \{1, \dots, n\}]$ Cluster gewählt werden.
- Hierarchische Clustering-Verfahren nehmen an, dass die Cluster von Tiefe $m-1$ in den Clustern von Tiefe m (bei agglomerativen Verfahren von unten gesehen) enthalten sind. Das führt nicht unbedingt zum besten Ergebnis und K-Means z.B. könnte besser abschneiden.

Hierarchische Clusteranalyse

Auch hier benötigen wir ein Maß für die Varianz (Unähnlichkeit) zwischen zwei Datenpunkten. Wir wählen wieder den quadrierten euklidischen Abstand

$$d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = ||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}||_2^2.$$

Dieses Maß zwischen Datenpunkten müssen wir auf ein Varianz-Maß zwischen Clustern, also Mengen von Datenpunkten, auch Linkage genannt, heben. Hier gibt es viele Möglichkeiten.

Hierarchische Clusteranalyse

Linkage	Definition $d(\mathcal{C}_a, \mathcal{C}_b)$	Hinweis
Complete	$\max_{i \in \mathcal{C}_a, j \in \mathcal{C}_b} d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$	Balancierte Cluster
Single	$\min_{i \in \mathcal{C}_a, j \in \mathcal{C}_b} d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$	Häufige Fusion Datenpunkt mit Cluster
Average	$\frac{1}{\lvert\mathcal{C}_a\rvert \lvert\mathcal{C}_b\rvert} \sum_{i \in \mathcal{C}_a, j \in \mathcal{C}_b} d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$	Balancierte Cluster
Centroid	$d(\frac{1}{\lvert\mathcal{C}_a\rvert} \sum_{i \in \mathcal{C}_a} \mathbf{x}^{(i)}, \frac{1}{\lvert\mathcal{C}_b\rvert} \sum_{j \in \mathcal{C}_b} \mathbf{x}^{(j)})$	Inversion (sinkende Distanz) möglich

Tabelle 1: Die vier häufigsten Linkage Definitionen.

Hierarchische Clusteranalyse

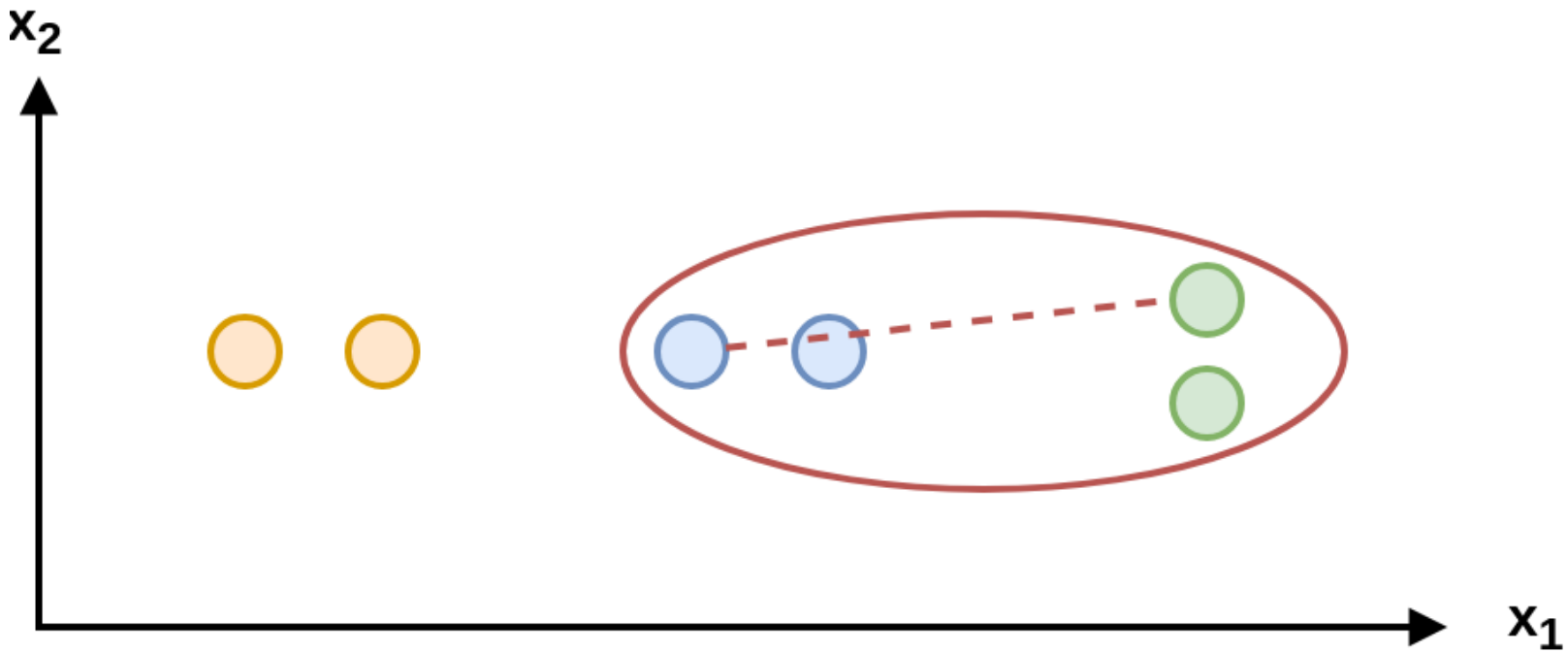

Abbildung 6: Beispiel – Complete Linkage.

Hierarchische Clusteranalyse

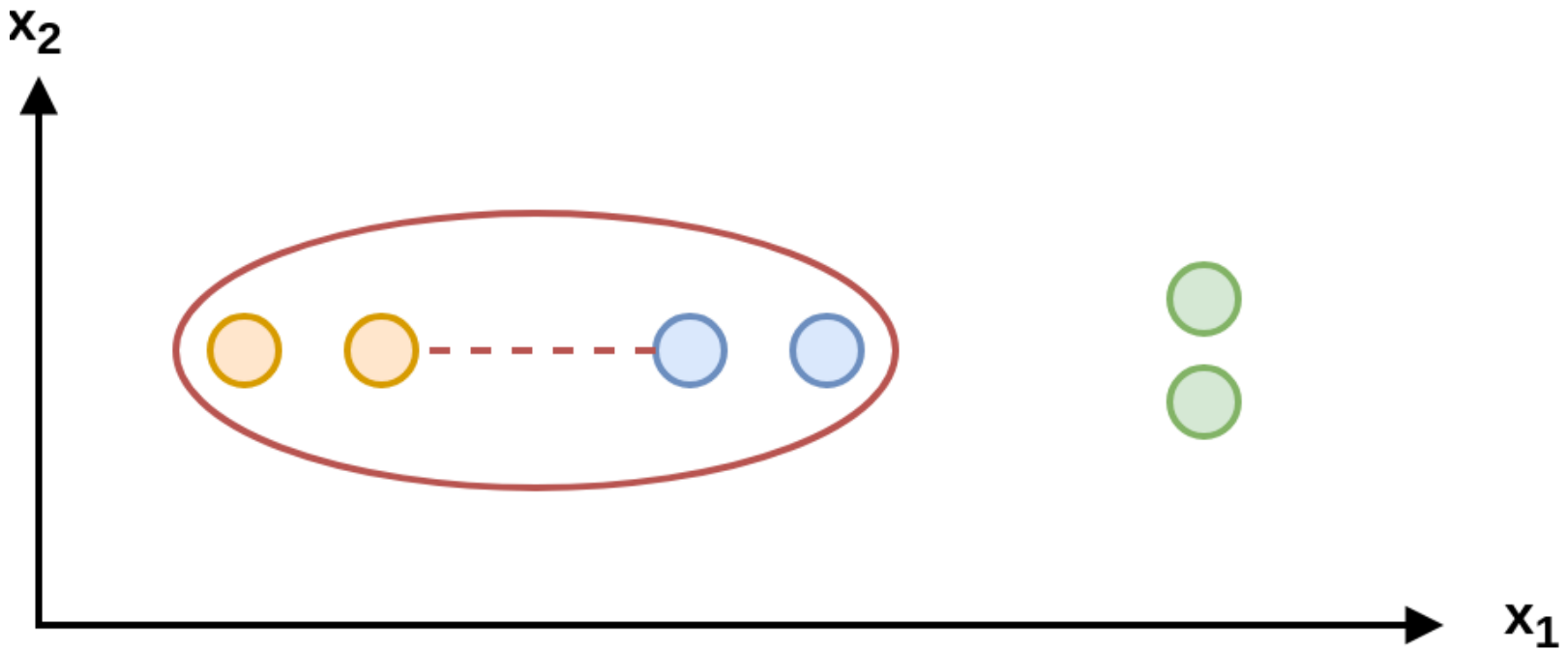

Abbildung 7: Beispiel – Single Linkage.

Hierarchische Clusteranalyse

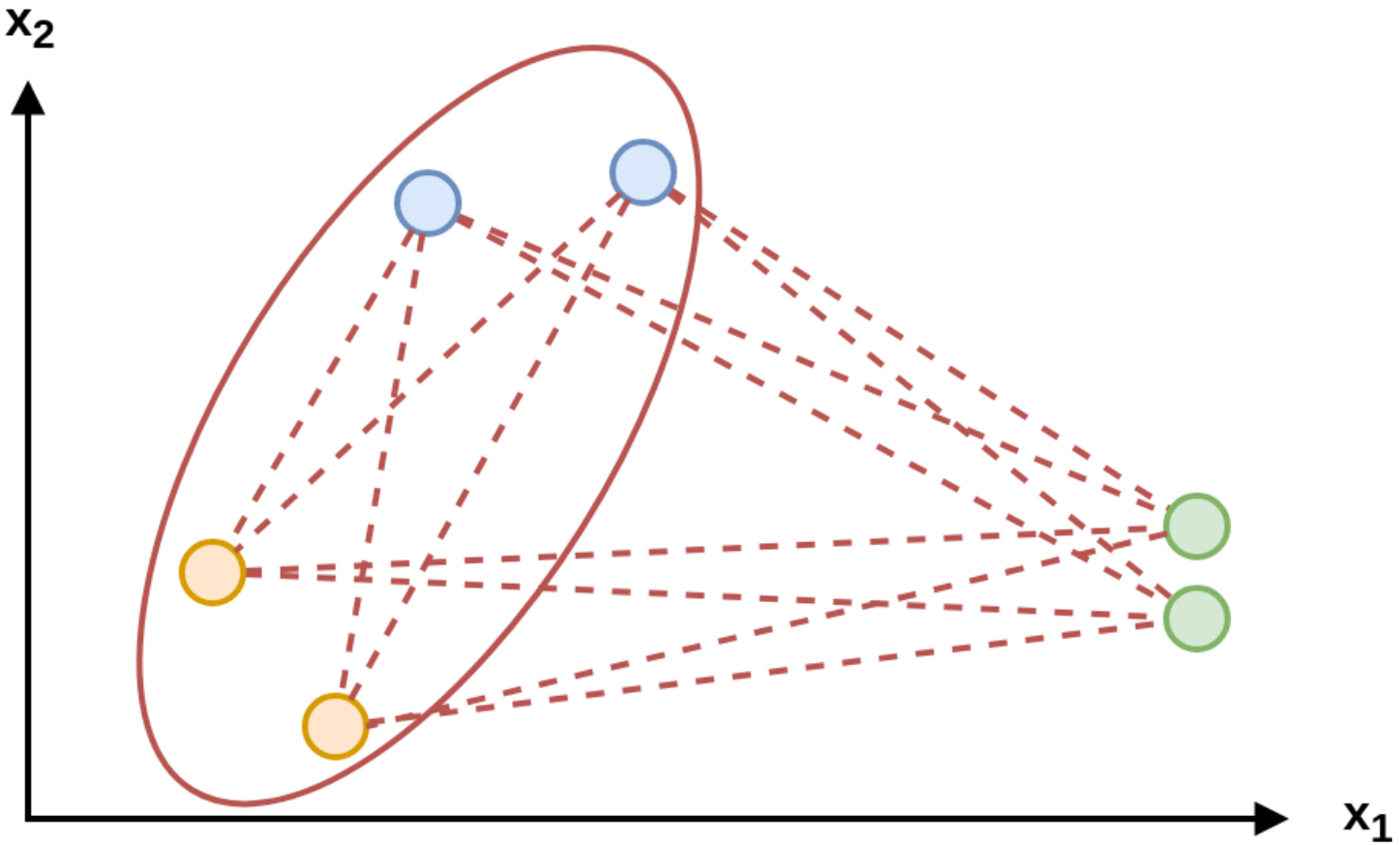

Abbildung 8: Beispiel – Average Linkage.

Hierarchische Clusteranalyse

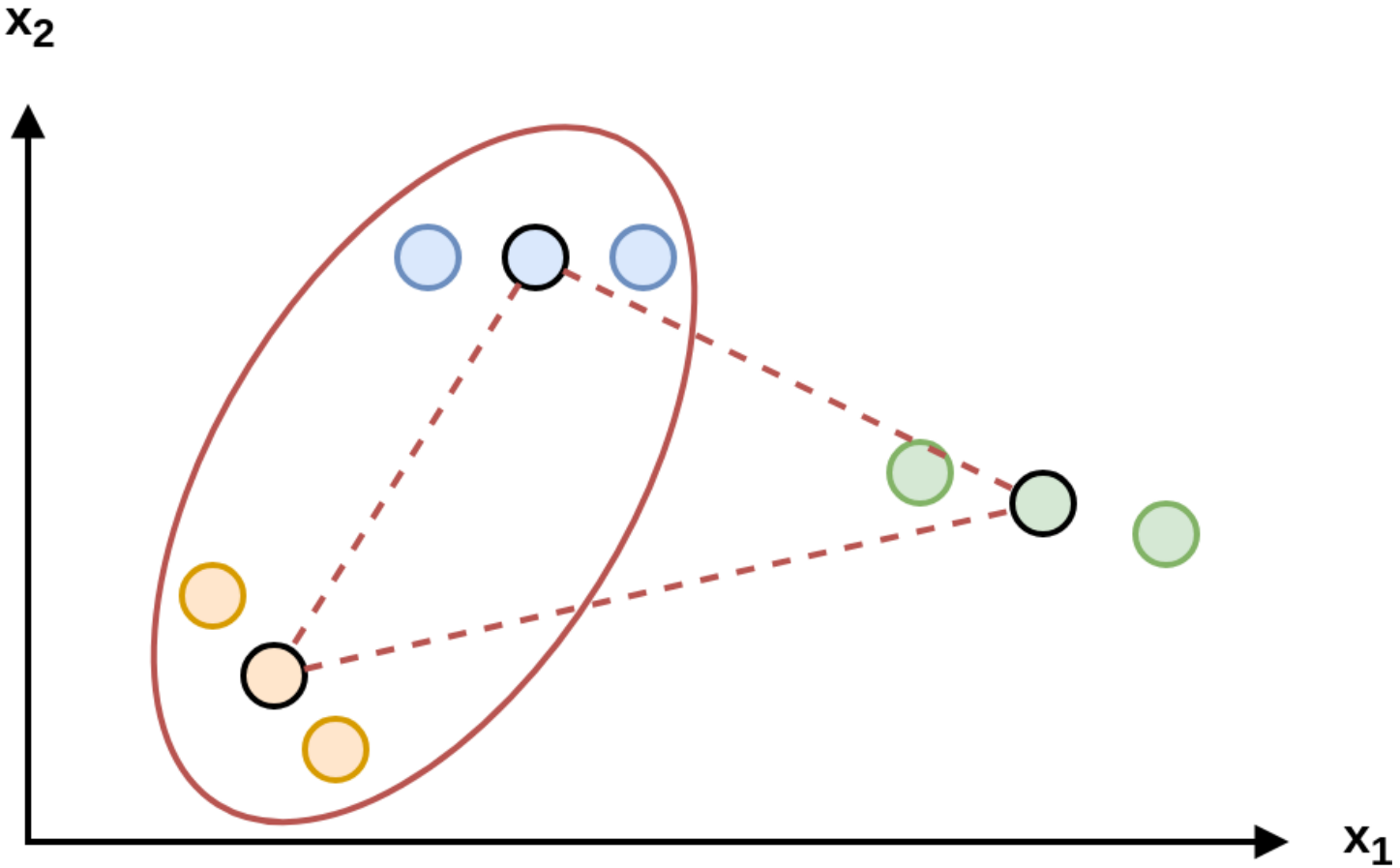

Abbildung 9: Beispiel – Centroid Linkage.

Hierarchische Clusteranalyse

Der Hierarchische Clustering Algorithmus kann nun beschrieben werden durch:

Algorithm 2 hierarchic_cluster($\mathcal{D}$, d)

1: $\mathcal{N} = \{\text{LEAF}(i) \mid 1 \leq i \leq n\}$
2: **while** $|\mathcal{N}| > 1$ **do**
3: $\quad a, b = \arg\min_{i,j \in \mathcal{N}, i \neq j} d(i,j)$
4: $\quad \mathcal{N} = \mathcal{N} \setminus \{a, b\} \cup \text{NODE}(a, b, d(a,b))$
5: **end while**
6: **return** root_node mit $\mathcal{N} = \{\text{root\_node}(a, b, d)\}$

Hierarchische Clusteranalyse

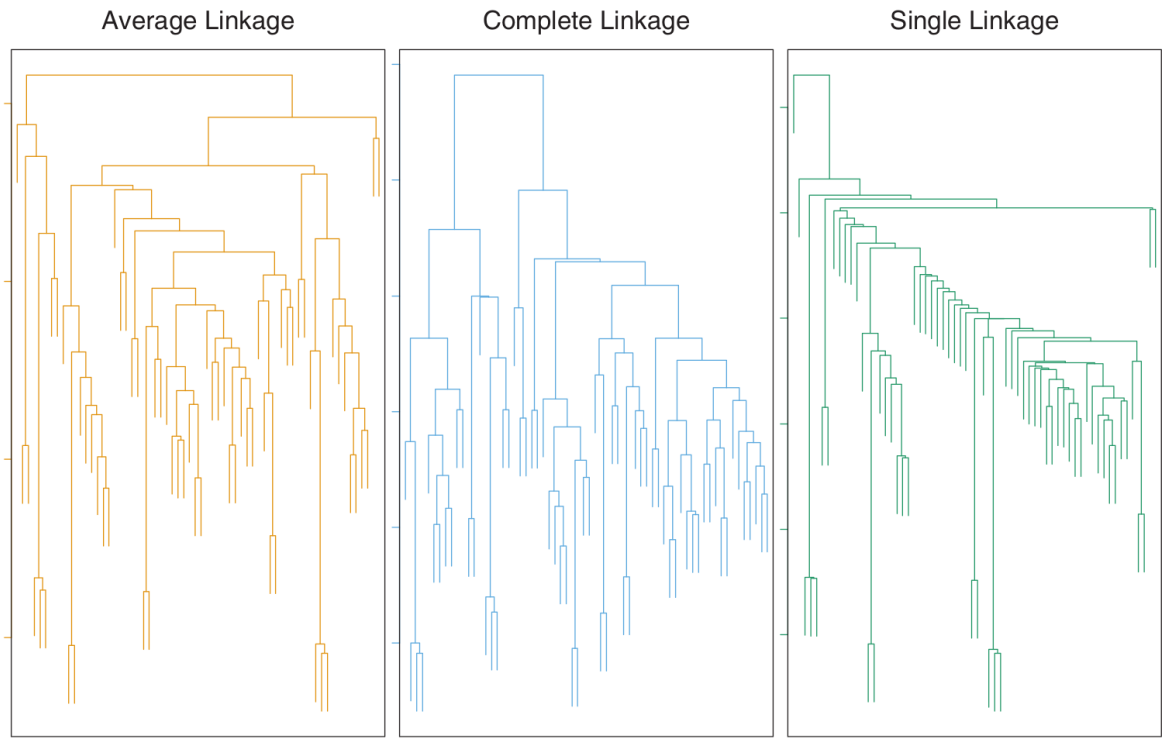

Abbildung 10: Hierarchisches Clustering mit drei unterschiedlichen Linkage Definitionen. Average und Complete Linkage führen meist zu balancierteren Clustern als Single Linkage. Abbildung entnommen aus [JWHT14].

Hierarchische Clusteranalyse

Hinweise:

- Das Varianzmaß sollte wohlüberlegt gewählt werden, so muss geprüft werden, ob die euklidische Distanz Sinn macht. Gegenbeispiel: Clustering von Käufertypen, da hierbei generell seltene Einkäufer gruppiert werden würden.
- Es sollte geprüft werden, ob eine Normalisierung sinnvoll ist. Beispiel: Oft gekaufte Gegenstände würden Clusterbildung bestimmen.
- Generell sollte das Ergebnis des Clusterings unabhängig von der Methode überprüft werden. Besitzen die Cluster eine sinnvolle Interpretation oder sind sie lediglich zufällige Gruppierungen von Rauschen?
- Clustering ist meist erst der Startpunkt einer explorativen Datenanalyse.

Hierarchische Clusteranalyse

References

G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: With applications in r*, Springer Publishing Company, Incorporated, 2014.