KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Clustering: vom Sortieren bis zum Explorieren:
06_04Modellauswahl_Modellauswahl01

Verfahren des Clustering

Erarbeitet von
Dr. Katja Theune

Lernziele

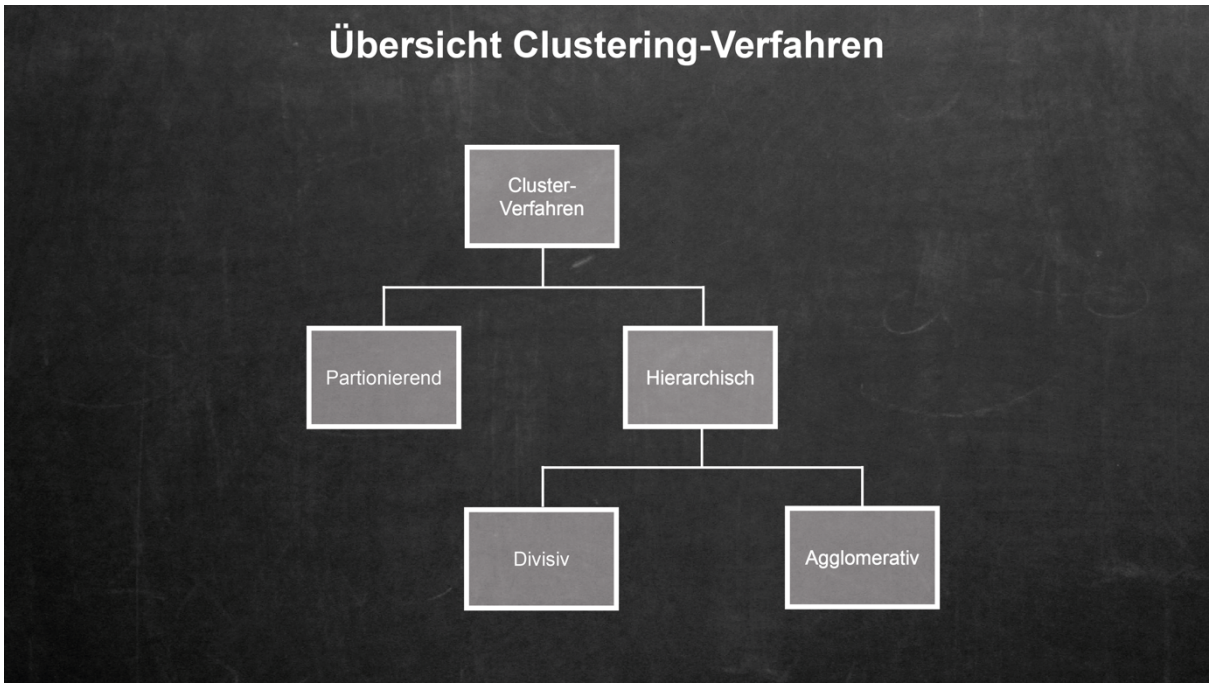
- Du kannst die Problematik bei der Wahl von k beim k-means Verfahren erläutern
- Du kannst die Idee und Vorgehensweise hierarchischer Verfahren (agglomerativ und divisiv) erläutern
- Du kannst ein Dendrogramm erklären
- Du kannst verschiedene Fusionierungsvorschriften erklären
- Du kannst Vor- und Nachteile der vorgestellten Cluster-Verfahren erläutern

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Wenn wir Beobachtungen aufgrund ähnlicher Eigenschaften in vorher unbekannte Gruppen einteilen wollen, können wir sogenannte Clustering-Verfahren verwenden. Sie gehören damit zum unsupervised bzw. unüberwachten Lernen. Das Ziel dieser Verfahren ist, dass sich Beobachtungen in den Gruppen, oder hier Clustern, sehr ähnlich und zwischen den Clustern sehr unähnlich sind.

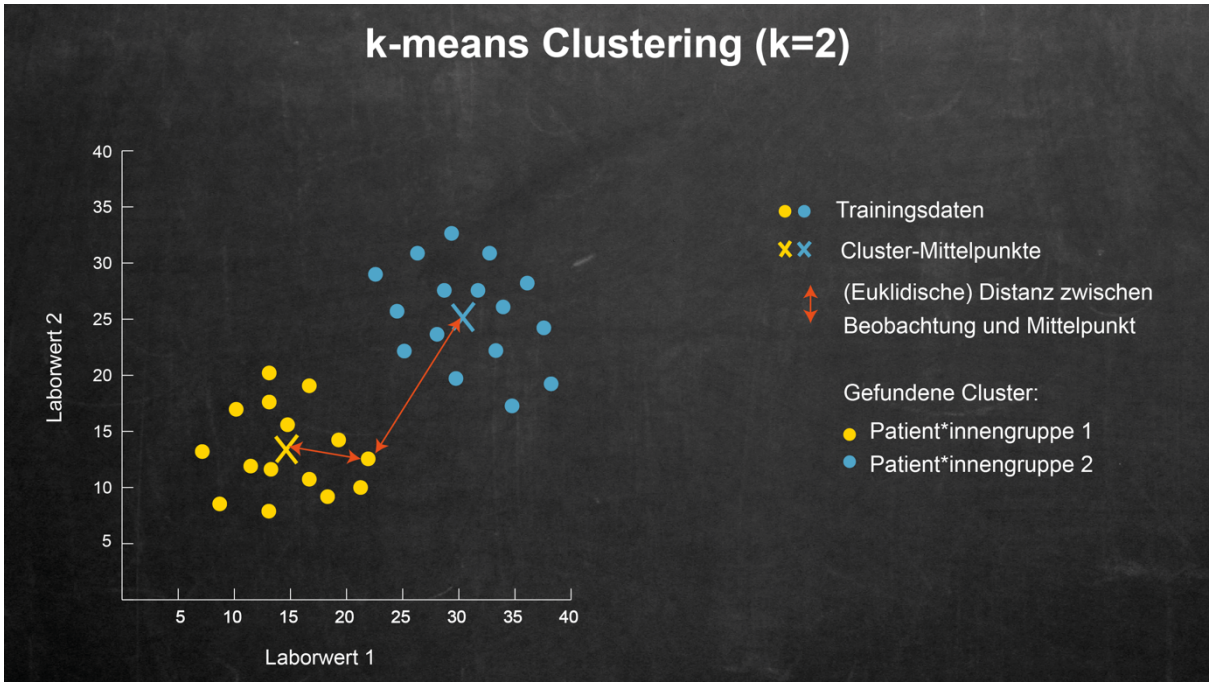


Es gibt sehr viele verschiedene Clustering-Verfahren. Wir werden uns hier mit partionierenden und insbesondere hierarchischen Verfahren beschäftigen. Bei hierarchischen Verfahren unterscheiden wir nochmal zwischen divisiven und agglomerativen Verfahren. Dazu kommen wir aber später noch.

Partionierende Verfahren: k-means

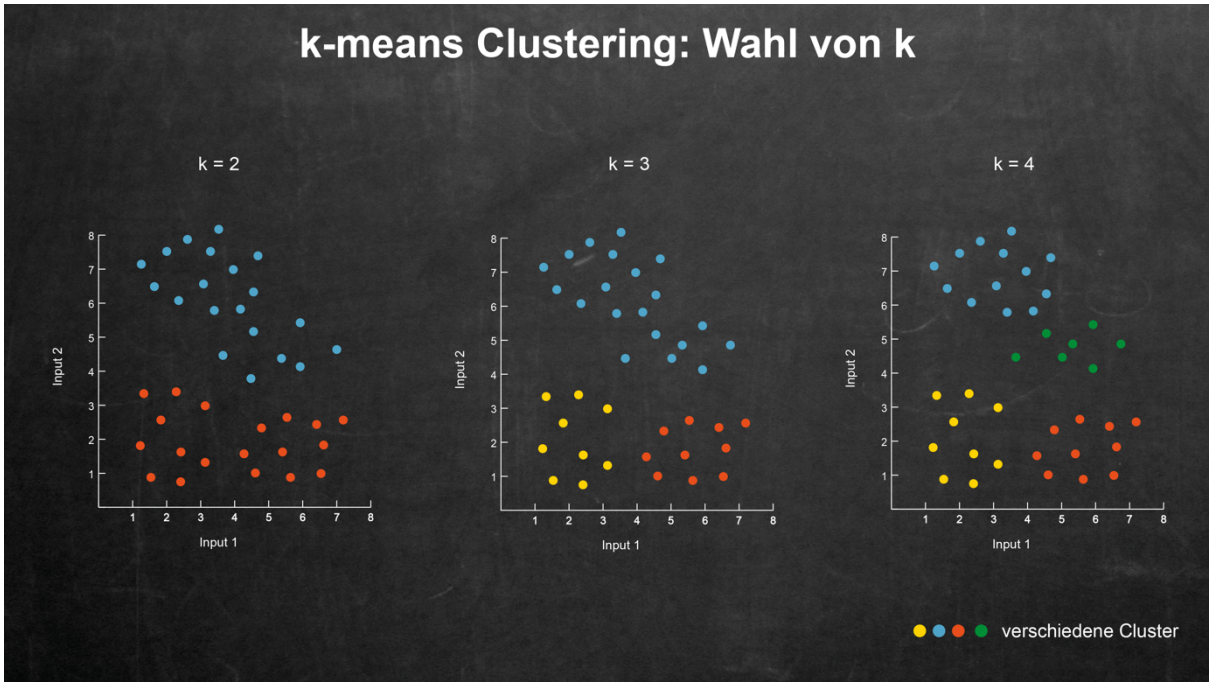
Stellvertretend für partionierende Verfahren, schauen wir uns jetzt das beliebte k-means Clustering an. Stellen wir uns vor, dass wir mit zwei Laborwerten Patient*innengruppen finden wollen, um gezielte Therapieansätze abzuleiten.

Wir wählen ein k von 2, geben also vor, dass zwei Cluster gebildet werden sollen. Wir beginnen mit einer zufälligen Bestimmung von k Clustermittelpunkten. Die Beobachtungen werden nun auf Basis ihrer kürzesten Distanz zu diesen Mittelpunkten jeweils genau einem Cluster zugeteilt. Sehr häufig wird dafür die Euklidische Distanz verwendet. Hier dargestellt durch die orangenen Pfeile. Es ist aber auch die Verwendung anderer Distanzmaße möglich.



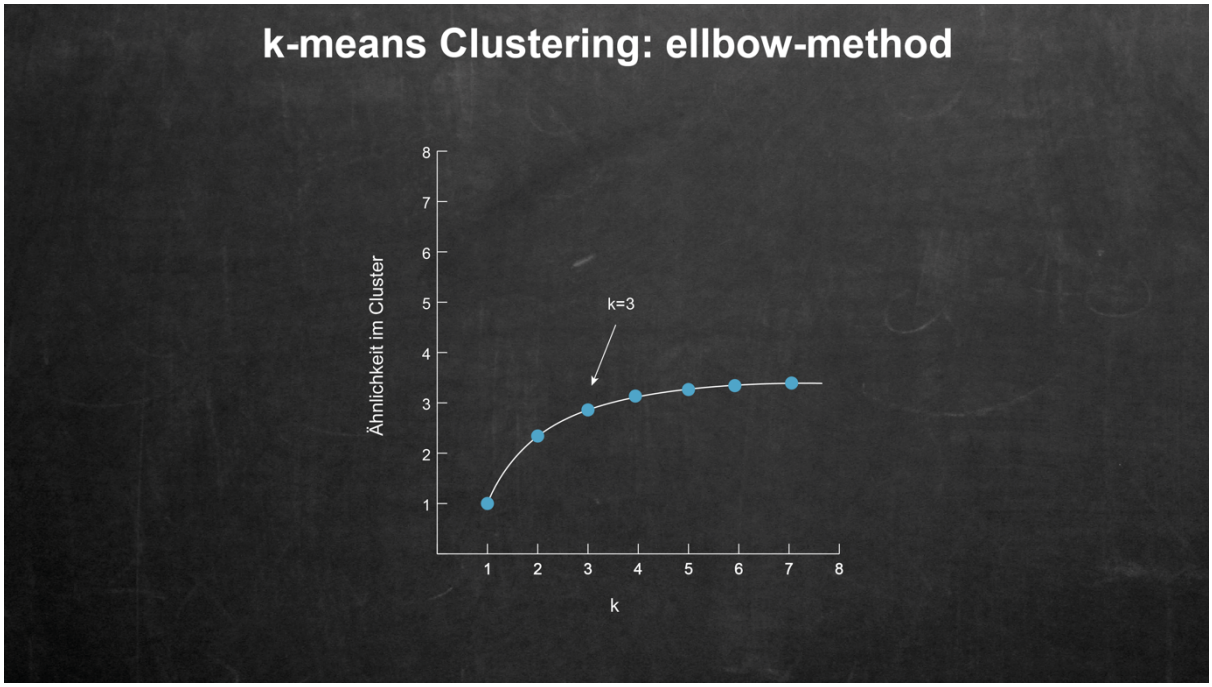
Nach dieser Zuteilung der Beobachtungen zu den Clustern müssen wir die neuen Cluster-Mittelpunkte berechnen und die Beobachtungen wieder den nächsten Mittelpunkten bzw. Clustern zuordnen. Während des gesamten Prozesses können unsere Beobachtungen also die Cluster wechseln. Dieser Prozess wird fortgeführt, bis keine Umverteilung der Beobachtungen zu den Clustern mehr erfolgt. Hier im Beispiel sehen wir aber schon die finalen Cluster. Die so entstandenen Cluster müssen wir dann interpretieren und ggf. benennen. Hier ergeben sich z. B. zwei Patient*innengruppen, eine mit hohen und eine mit niedrigen Laborwerten.

Partionierende Verfahren: Wahl von k



Bei dem k-means Verfahren hat die Wahl der Anzahl an Clustern k einen sehr großen Einfluss auf die finalen Cluster. Wir sehen hier beispielhaft unterschiedliche Resultate, wenn wir k = 2, 3 oder 4 wählen. Eine „falsche" Wahl kann zu wenig sinnvollen Clustern führen.

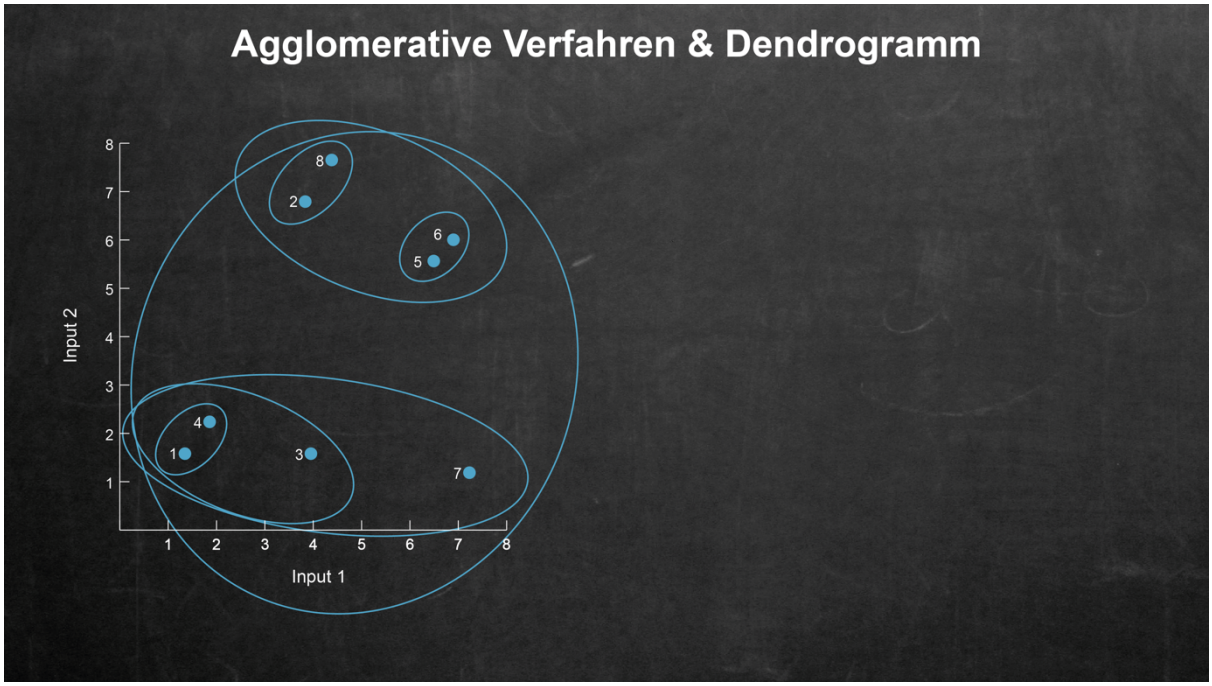
Um das „richtige" k zu finden, gibt es leider kein Geheimrezept. Bei der sogenannten Ellbogen-Methode, oder englisch elbow-method wird z. B. die Ähnlichkeit der Beobachtungen in den einzelnen Clustern gegen die Anzahl an Clustern k visualisiert. Man wählt dann das k, ab welchem eine Erhöhung von k nicht mehr zu einer starken Verbesserung der Ähnlichkeit führt. Also das k, ab dem wir den Knick, also den Ellbogen, sehen.



Manchmal haben wir im Vorhinein auch schon eine Idee, wie groß k in unserem Anwendungsfall sein könnte oder wir haben hierfür spezielle Restriktionen. Zum Beispiel könnten für eine medizinische Studie nur zwei verschiedene Patient*innengruppen vorgegeben sein.

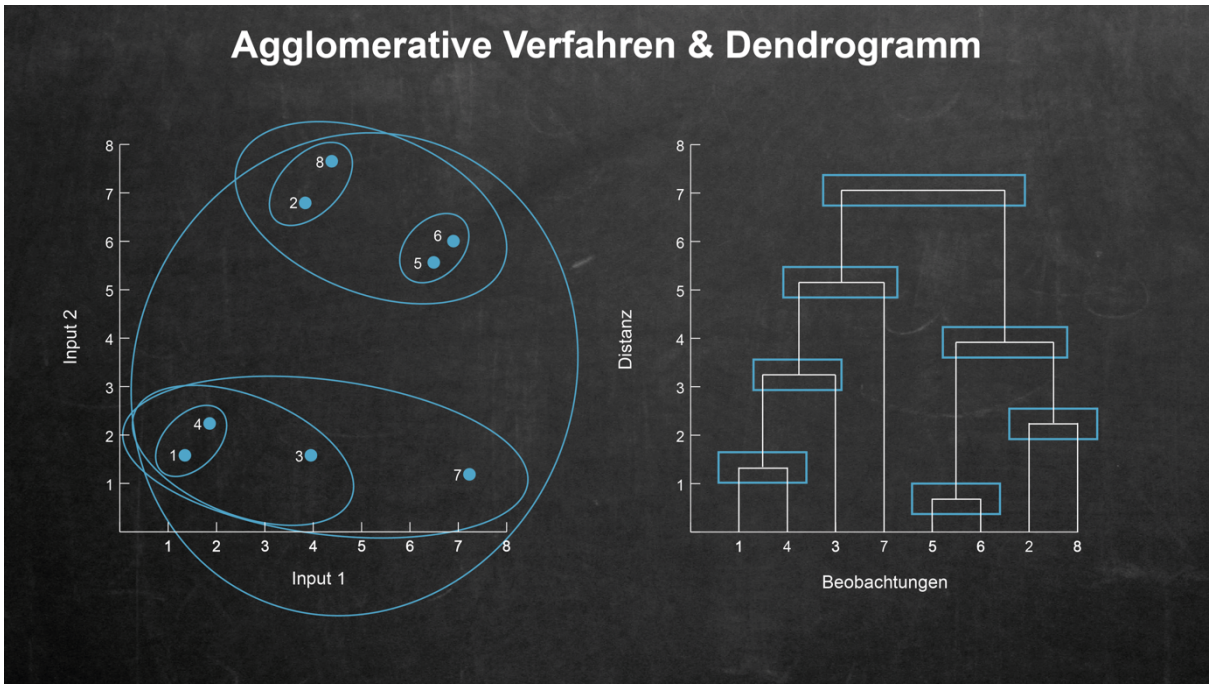
Hierarchische Verfahren: agglomerativ & divisiv

Eine Alternative zu den partionierenden Verfahren sind hierarchische Verfahren.
Bei agglomerativen Verfahren beginnen wir damit, dass alle Beobachtungen ein eigenes Cluster bilden. Nach und nach werden sie in immer größere Cluster zusammengeführt. Dabei werden ganz zu Anfang die zwei ähnlichsten Beobachtungen bzw. Cluster vereint. Diese Ähnlichkeit wird wie beim k-means Verfahren mit Distanzmaßen gemessen. Danach werden wieder die zwei ähnlichsten Cluster vereint. Das Zusammenführen wird so lange fortgeführt, bis alle Beobachtungen in einem einzigen Cluster sind. Agglomerative Verfahren heißen daher auch bottom-up Verfahren.



Divisive Verfahren nennt man dagegen auch Top-down Verfahren. Hier beginnen wir mit einem einzigen Cluster, in dem sich alle Beobachtungen befinden. Dieses wird dann Schritt für Schritt in immer kleiner werdende Cluster aufgeteilt. Hier wird der Prozess so lange fortgeführt, bis alle Beobachtungen in einem eigenen Cluster sind.

Hierarchische Verfahren: Dendrogramm



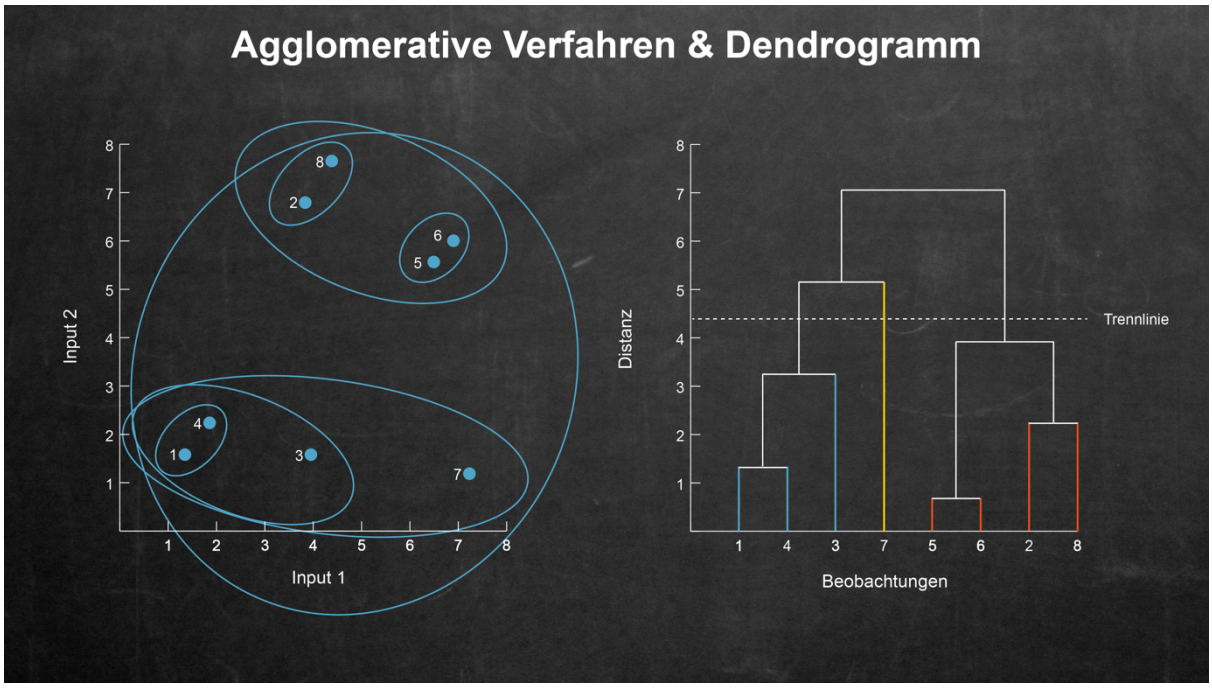
Praktischerweise kann man den Prozess des Zusammenführens oder Teilens der Cluster bei hierarchischen Verfahren sehr übersichtlich in einem sogenannten Dendrogramm darstellen, hier rechts im Bild. Horizontal sind hier die Beobachtungen 1 bis 8 abgetragen

und vertikal die, hier rein fiktiven, Distanzwerte. Die Höhe der Blöcke repräsentiert also die Distanz zwischen den Beobachtungen bzw. Clustern im Koordinatensystem links. Je niedriger der Block, also je weiter unten Beobachtungen zusammengeführt werden, desto ähnlicher sind sich diese. Die Reihenfolge bzw. Distanz der Beobachtungen auf der horizontalen Achse im Dendrogramm sagt hier dagegen nichts über ihre Ähnlichkeit aus. Z. B. ist im Koordinatensystem Beobachtung 5 viel näher an 6 als an 7.

Schauen wir uns mal das agglomerative Verfahren genauer an. Beobachtung 5 und 6 sind sich am nächsten und werden in einem ersten Schritt zusammengeführt … Danach Beobachtung 1 und 4 … dann 2 und 8 … Als nächstes sehen wir, dass das Cluster mit den Beobachtungen 1 und 4 zu Beobachtung 3 am nächsten liegt, sie werden zusammengeführt usw. …

Einmal zusammengeführte Cluster bleiben dann auch für den Rest des Prozesses zusammen. Einmal vorgenommene Entscheidungen werden also strikt eingehalten. Die Cluster werden nur verfeinert oder verallgemeinert. Daher auch der Name „hierarchisch".

Mit einem einzigen Dendrogramm können wir jetzt jede gewünschte Anzahl an Clustern auswählen. Dazu können wir in das Dendrogramm auf einer bestimmten Höhe, also ab einem bestimmten Distanzwert, eine Trennlinie einfügen (siehe Grafik unten). Hier würden wir z. B. drei Cluster erhalten. Eines, dass die Beobachtungen 1, 4 und 3 enthält, hier in Blau. Eines, das nur Beobachtung 7 enthält, hier in Gelb. Und eines, dass die restlichen Beobachtungen 5, 6, 2 und 8 enthält, hier in Orange.
So gibt uns ein Dendrogramm Informationen über den gesamten Clustering-Prozess und damit auch über das Ergebnis für verschiedene Werte von k.

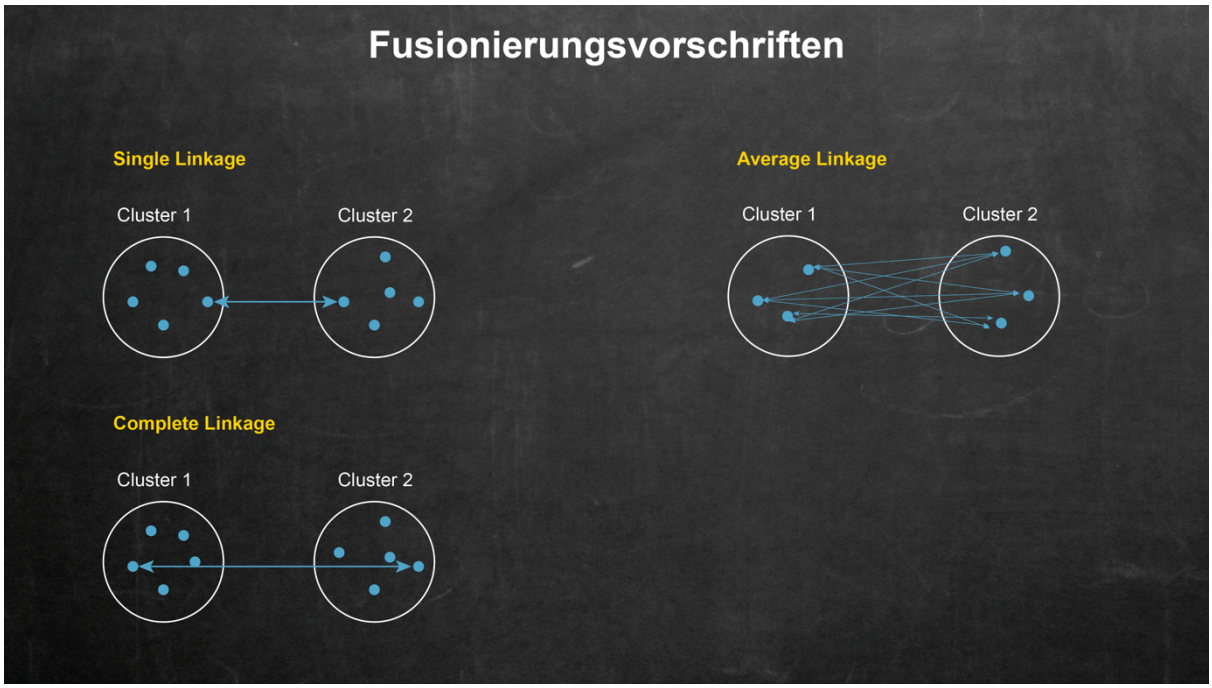


Hierarchische Verfahren sind insbesondere dann sinnvoll, wenn wir z. B. auch in den zu findenden Clustern eine Hierarchie vermuten. Wir können uns hier z. B. ein Clustering von Mitarbeitenden einer Firma mit verschiedenen Aufgabenleveln vorstellen.

Der Prozess bei divisiven Verfahren findet im Prinzip genau andersherum zum agglomerativen Verfahren statt.

Hierarchische Verfahren: Fusionierungsvorschriften

Für das Zusammenführen von Clustern mit mehreren Beobachtungen müssen wir noch überlegen, wie genau das passieren soll. Hier benötigen wir nämlich noch neben dem Distanzmaß, das ja bei der Fusion einzelner Beobachtungen reicht, noch Informationen, wozwischen genau diese Distanz bei mehreren Beobachtungen gemessen werden soll. Wir benötigen also sogenannte Fusionierungsvorschriften.



Die single linkage Vorschrift sucht die zwei sich am nächsten liegenden Beobachtungen in den jeweiligen Clustern. Sie verwendet also die kürzeste Distanz zwischen zwei Beobachtungen. Die beiden Cluster mit der kleinsten kürzesten Distanz werden fusioniert.

Die complete linkage Vorschrift nutzt dagegen die größte Distanz zwischen zwei Beobachtungen in den jeweiligen Clustern. Hier werden die beiden Cluster mit der kleinsten größten Distanz miteinander fusioniert.

Die average linkage Vorschrift misst alle paarweisen Distanzen zwischen den Beobachtungen in den jeweiligen Clustern und berechnet die durchschnittliche Distanz. Die beiden Cluster mit der kleinsten durchschnittlichen Distanz werden fusioniert.
Das ist sozusagen ein Kompromiss zwischen den beiden anderen genannten Extremen und nicht so anfällig gegenüber Ausreißern.

Außer diesen dreien gibt es noch einige weitere Fusionierungsvorschriften. Je nachdem, welche Vorschrift wir wählen, können sich natürlich unterschiedliche Fusionierungen und damit sehr unterschiedliche Dendrogramme ergeben.

Diskussion, Vor- und Nachteile

Alle vorgestellten Clustering-Verfahren haben den Vorteil, sehr intuitiv zu sein. Sie entsprechen im Prinzip unserer optischen Herangehensweise an diese Art von Problem. Insbesondere liefern hierarchische Verfahren aber auch eine übersichtliche Darstellung der Cluster-Bildung. Mit beiden Verfahren können wir neues Wissen generieren. Allerdings liegt auch die Interpretation der gefundenen Cluster in unserer Hand, was sowohl ein Nachteil als auch ein Vorteil sein kann.

Ein Nachteil beim k-means Verfahren ist im Gegensatz zu hierarchischen Verfahren, dass die Anzahl an Clustern im Vorhinein festgelegt werden muss.

Abschluss

Wir haben hier die Idee und Vorgehensweise partionierender und insbesondere hierarchischer Verfahren kennengelernt. Bei Letzterem müssen wir nicht die Anzahl an Clustern im Vorhinein bereits vorgeben. Hierarchische Verfahren geben uns auch hilfreiche Informationen über den gesamten Clustering-Prozess. Welches Verfahren sich am besten eignet, hängt wie immer vom jeweiligen Anwendungsfall ab.

Weiterführendes Material

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3. Auflage). Morgan Kaufmann.

James, G., Witten, D., Hastie, T., & Tibshirani, R., & Tylor, J. (2023). *An Introduction to Statistical Learning - with Applications in Python*. Springer.

Lantz, B. (2015). *Machine learning with R* (2. Auflage). Packt Publishing Ltd, Birmingham.

Disclaimer

Transkript zu dem Video „Clustering: vom Sortieren bis zum Explorieren: Verfahren des Clustering“, Dr. Katja Theune.