Woche 09 Theorie: K-means Clustering

Skript

Erarbeitet von
Katja Theune

Lernziele 1
Inhalt 2
Einstieg 2
Was ist Clustering? 2
K-means Clustering – Beispiel 3
K-means Clustering – Finden von Clustern in Daten 4
Interpretation der Cluster 7
Abschluss 8
Weiterführendes Material 9
Disclaimer 10

Lernziele

- Definieren, was Clustering ist
- Erläutern der Idee und Vorgehensweise des k-means Clustering
- Anwenden der Vorgehensweise des Verfahrens auf ein neues Beispiel
- Beispiele nennen, wozu man k-means Clustering verwendet

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

Inhalt

Einstieg

Viele von uns haben sich doch bestimmt schonmal gefragt, warum wir bei Streamingdiensten andere Serien vorgeschlagen bekommen als unsere Freund*innen. Oder auch z. B. beim Surfen im Internet andere Werbung angezeigt bekommen. Hier ist zumeist nichts anderes passiert, als dass wir aufgrund unserer Eigenschaften und unseres Kauf- oder Sehverhaltens in verschiedene Kund*innengruppen eingruppiert worden sind und dann gezielt auf unsere Gruppe abgestimmte Werbung erhalten haben. Eine Methode, die dafür häufig verwendet wird, ist das Clustering. Es gehört zum sogenannten unsupervised learning.

Was ist Clustering?

Aber was ist Clustering eigentlich genau? Auch hier gibt es wieder viele verschiedene Methoden, aber sie verfolgen eine gemeinsame Idee. Sie versuchen, unsere vorliegenden Trainingsdaten in Gruppen, oder eben auch Cluster genannt, einzuteilen. Das Ziel ist dabei, dass sich Beobachtungen innerhalb dieser Cluster sehr ähnlich sind, sich Beobachtungen aus verschiedenen Clustern aber möglichst gut voneinander unterscheiden.

Einblendung Piktogramme mit drei Benutzer*innengruppen

Beim Clustering sind aber, im Gegensatz zu überwachten Klassifikationsmethoden, die Gruppen, denen die Beobachtungen zugeordnet werden, nicht im Vorhinein bekannt oder irgendwie definiert. Sogar ganz im Gegenteil. Die Interpretation und Benennung der gefundenen Cluster muss im Nachhinein von den Anwendern und Anwenderinnen selbst übernommen werden. Clustering ist, wieder im Gegensatz zu vielen überwachten Lernmethoden, weniger für Prognosen gemacht, sondern eher für die Generierung von neuem Wissen. Oft weiß man vorher gar nicht, was man eigentlich sucht und findet. Das Ziel ist es also, in den Daten ganz neue Gruppen zu finden und diese dann ggf. für weitere Analysen, z. B. auch zur Klassifikation, weiterzuverwenden.

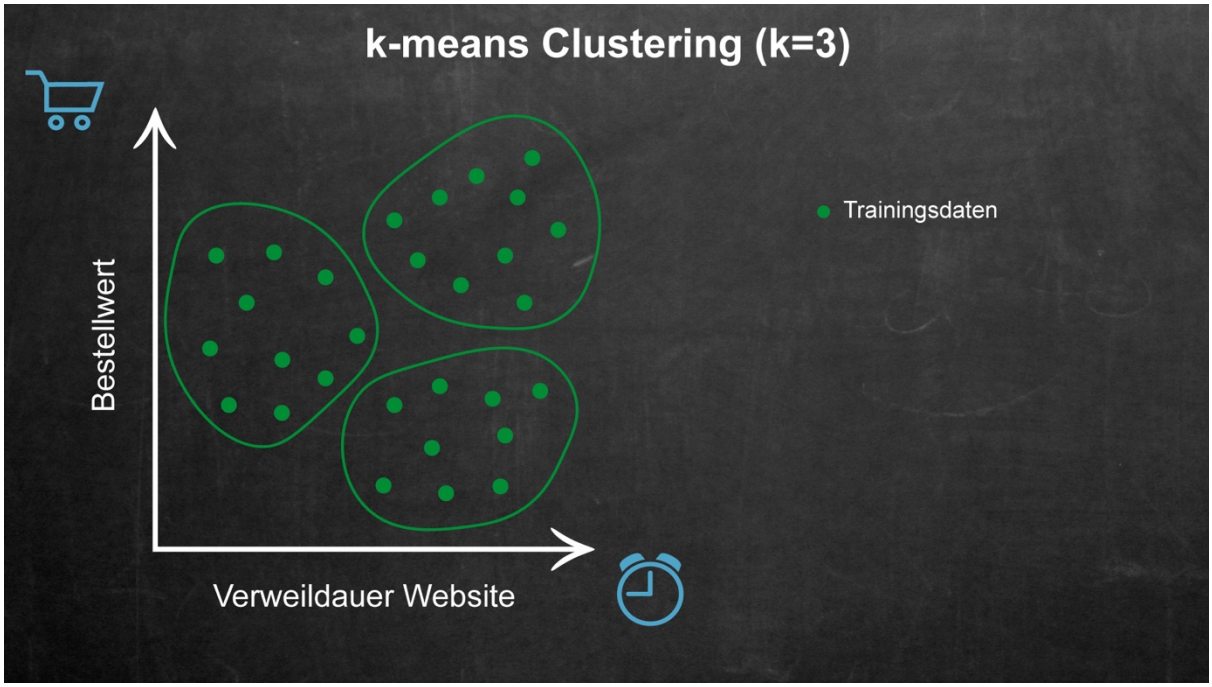
Beim Clustering können wir verschiedene Arten unterscheiden. Da wäre z. B. die Unterscheidung in hierarchische und partitionierende Algorithmen, um mal zwei Schlagwörter zu nennen. Hierarchische Verfahren fassen entweder nacheinander einzelne Beobachtungen zu Clustern zusammen oder versuchen andersherum, nach und nach die Gesamtheit der Daten sinnvoll in Cluster aufzuspalten. Partitionierende Verfahren gehen dagegen von einer gegebenen Gruppierung der Daten aus und versuchen, einfach ausgedrückt, in mehreren Schritten durch Anpassungen diese Gruppierung zu verbessern. Zu diesen partitionierenden Verfahren gehört auch der sehr beliebte und häufig angewendete k-means Algorithmus. Um diesen kümmern wir uns jetzt im Folgenden etwas genauer.

K-means Clustering – Beispiel

Dafür ist es zunächst sinnvoll, sich mal ein konkretes Beispiel anzuschauen, um die Idee des k-means Clustering besser zu verstehen. Ein klassisches Anwendungsfeld ist die sogenannte Markt- bzw. Kund*innensegmentierung. Im Marketing wird damit versucht, Werbung besser auf verschiedene Kund*innentypen anzupassen.

Einblendung Piktogramme mit drei Benutzer*innengruppen

Wir können uns jetzt vorstellen, dass wir genau das für unseren Onlinehandel machen wollen. Uns liegen dafür verschiedene features von unseren Kunden und Kundinnen vor, anhand derer wir glauben, sie unterscheiden zu können. Das können z. B. demographische features sein wie das Alter oder das Geschlecht, aber auch Informationen zu ihrem Kaufverhalten. Z. B. so etwas wie der durchschnittliche Bestellwert und die Verweildauer auf unserer Website. Um es anschaulich zu halten, verwenden wir jetzt nur die zwei letztgenannten features, und zwar den durchschnittlichen Bestellwert und die durchschnittliche Verweildauer auf unserer Website.

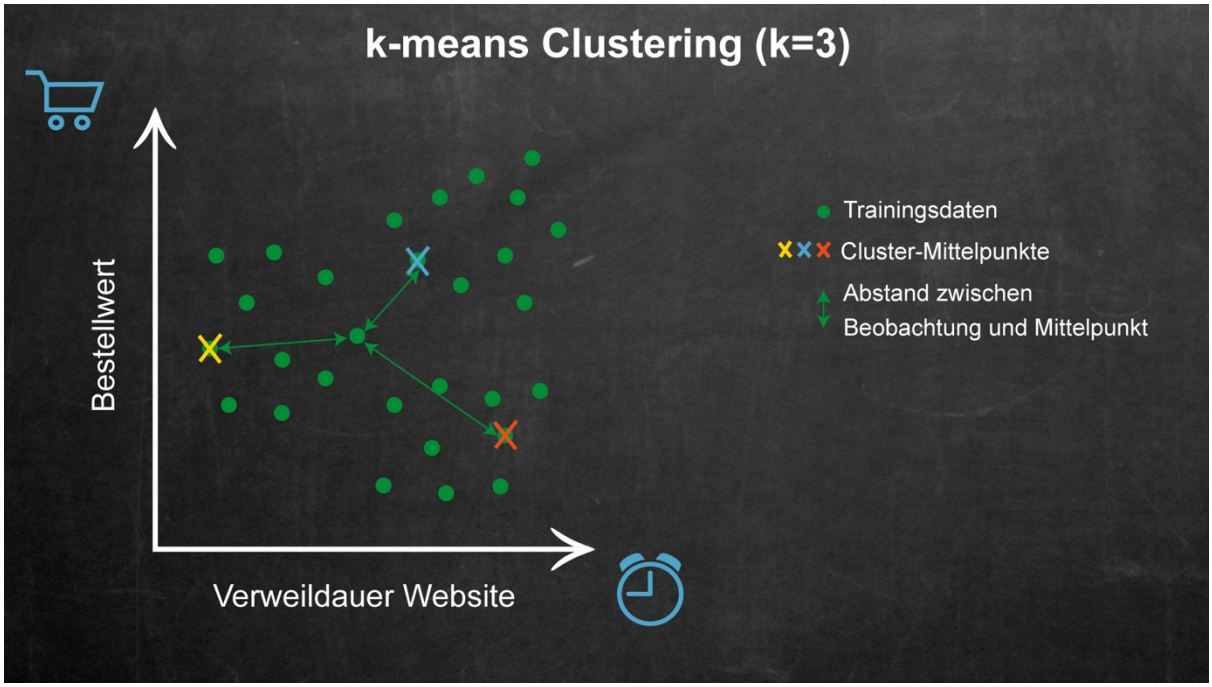


Wir tragen jetzt unsere Trainingsdaten in einem Koordinatensystem als grüne Datenpunkte ein. Da wir zwei features verwenden, haben wir ein zweidimensionales Koordinatensystem. Wir tragen die durchschnittliche Verweildauer auf der Website auf der horizontalen Achse und den durchschnittlichen Bestellwert auf der vertikalen Achse ab. Je weiter rechts die Beobachtungen liegen, desto länger verweilen sie auf der Website. Je weiter oben die Beobachtungen liegen, desto höher ist ihr durchschnittlicher Bestellwert.

Wir sehen bereits rein optisch, dass sich an drei Stellen die Beobachtungen häufen. Das deutet auf drei mögliche Cluster hin. Hier sind sie grün umkreist. Aber wie findet man diese Cluster nun mit Methoden des Maschinellen Lernens? Im Prinzip folgen die Algorithmen unserer optischen Herangehensweise, so auch das k-means Clustering.

K-means Clustering – Finden von Clustern in Daten

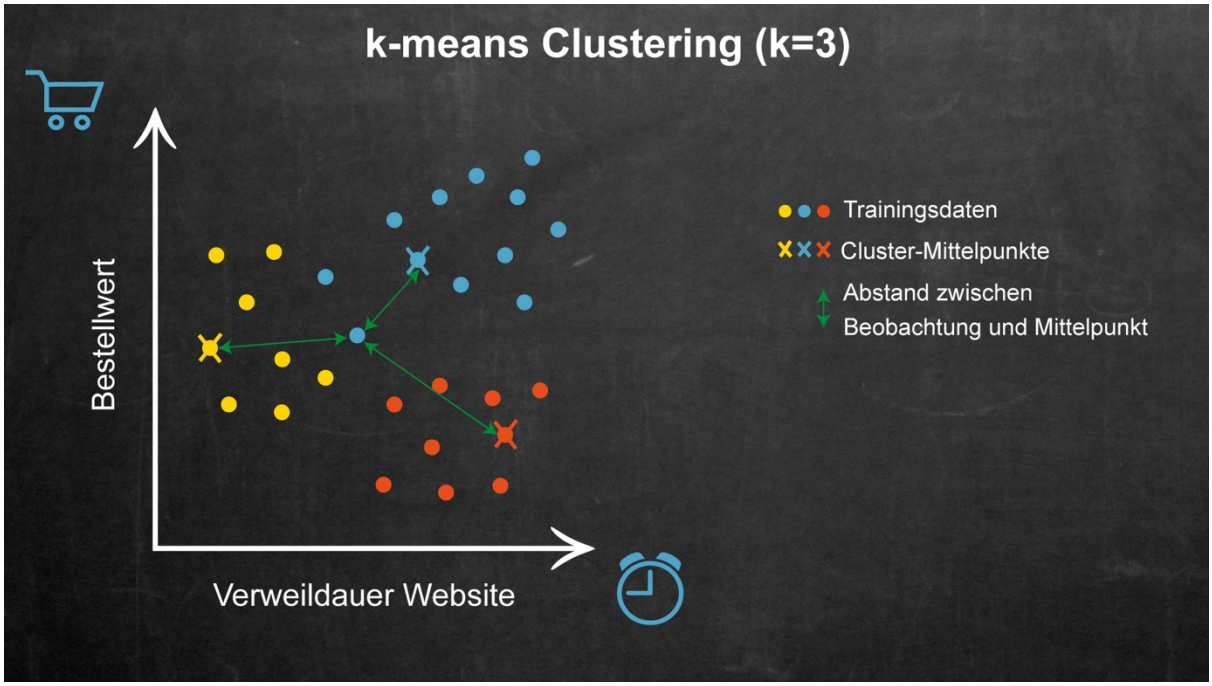
Bei dieser Methode müssen wir die Anzahl der Cluster, die wir in den Daten vermuten, zu Beginn vorgeben. Diese Anzahl bezeichnen wir mit k. In unserem Beispiel vermuten wir ja k=3 Cluster. Wie der Name k-means schon vermuten lässt, werden die Cluster durch ihre Mittelwerte bzw. Mittelpunkte repräsentiert. Der Algorithmus beginnt jetzt mit einer zufälligen Aufteilung der vorliegenden Daten in k Cluster. Dafür gibt es wieder verschiedene Herangehensweisen. Häufig werden z. B. zufällig k Trainingsdaten als erste Cluster-Mittelpunkte verwendet. Hier sind diese durch die drei Kreuze in Gelb, Blau und Orange gekennzeichnet.



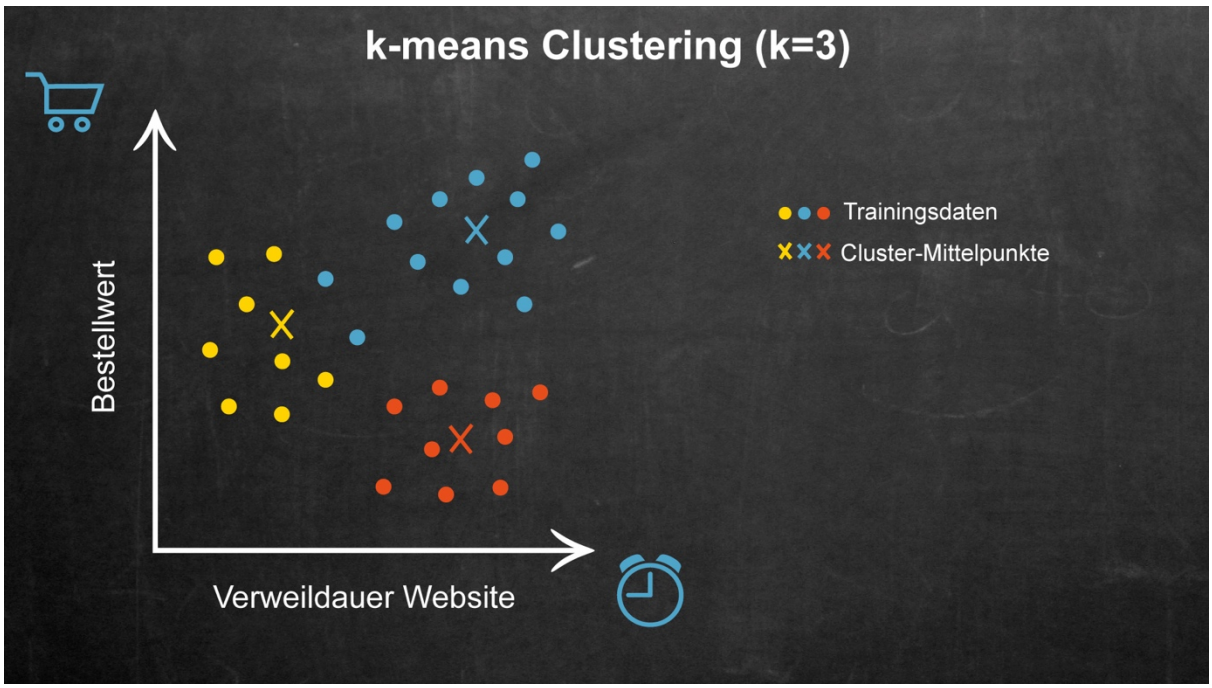
Haben wir diese k Mittelpunkte, können wir für jede Beobachtung ihre Abstände zu jedem der k Mittelpunkte bestimmen. Diese Abstände kann man wiederum mit verschiedenen Formeln berechnen, häufig wird die sogenannte Euklidische Distanz verwendet. Sie entspricht der Länge der Geraden, die wir zwischen zwei Beobachtungen ziehen und theoretisch mit einem Lineal abmessen könnten.

Diese Abstände einer beispielhaften Beobachtung zu den Mittelpunkten sind hier durch die drei grünen Pfeile gekennzeichnet. Wir ordnen jetzt jede Beobachtung dem Cluster zu, dessen Mittelpunkt der Beobachtung am nächsten liegt.

CC BY

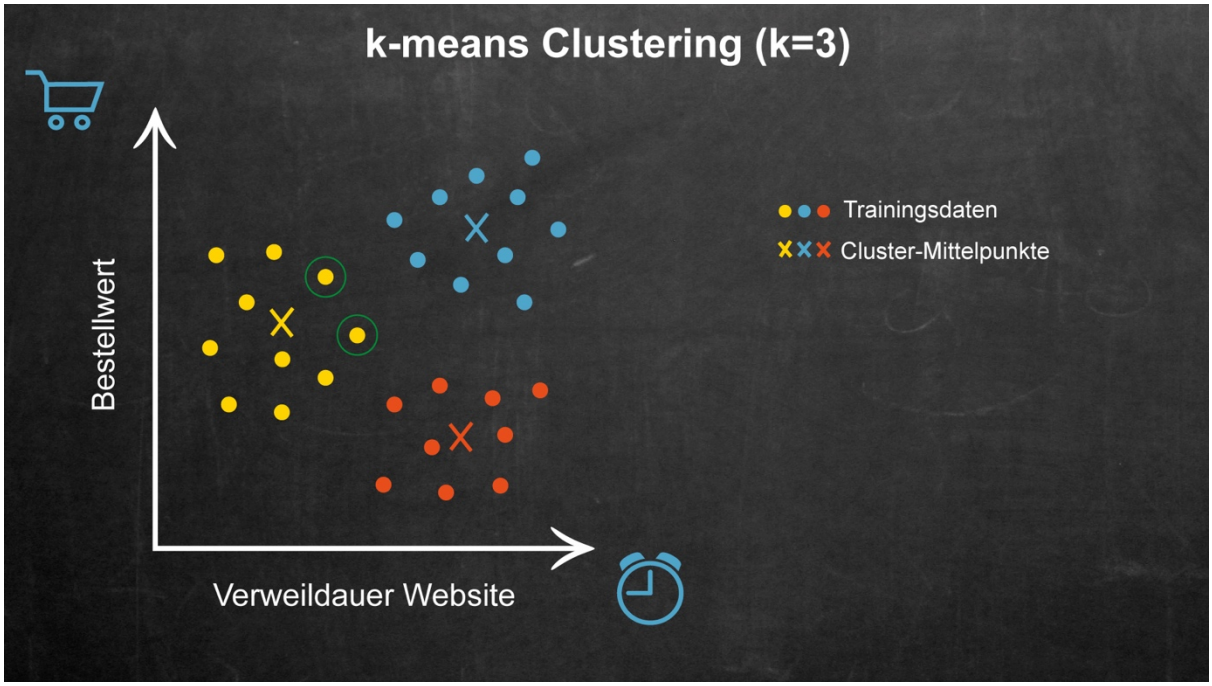


Zur besseren Veranschaulichung stellen wir unsere drei zufällig entstandenen Cluster farblich unterschiedlich dar. Die beispielhafte Beobachtung würde jetzt zum blauen Cluster gehören, da sie am nächsten an dem blauen Mittelpunkt liegt. Die zufällig vorgegebenen Mittelpunkte entsprechen jetzt natürlich nicht mehr den wahren Mittelpunkten der drei entstandenen Cluster. Um diese wahren Mittelpunkte zu berechnen, können wir einfach alle Beobachtungen in dem jeweiligen Cluster nehmen und mit ihren Werten für die verwendeten features den wahren neuen Cluster-Mittelpunkt berechnen. Die drei neu berechneten Mittelpunkte sind wieder durch die Kreuze, jetzt in veränderter Position, gekennzeichnet.



CC BY

Sie müssen nicht einer der Beobachtungen in den Trainingsdaten entsprechen. Mit diesen drei neuen Mittelpunkten können wir nun wieder die Abstände zu jeder Beobachtung berechnen und die Beobachtung dem Cluster zuordnen, dessen Mittelpunkt die geringste Distanz zu ihr aufweist. Das führt jetzt zu einer Umverteilung der Beobachtungen zu den Clustern. In unserem Beispiel gehören nun zwei blaue Beobachtungen zum gelben Cluster. Sie sind grün umkreist.



Diese Schritte der Neuberechnung der Mittelpunkte und die Umverteilung der Beobachtungen zu den Clustern wiederholen wir z. B. so oft, bis der Algorithmus keine Umverteilung mehr vornimmt, jede Beobachtung also in ihrem Cluster bleibt. Wir fassen die einzelnen Schritte noch einmal kurz zusammen:

Einblendung der einzelnen Schritte

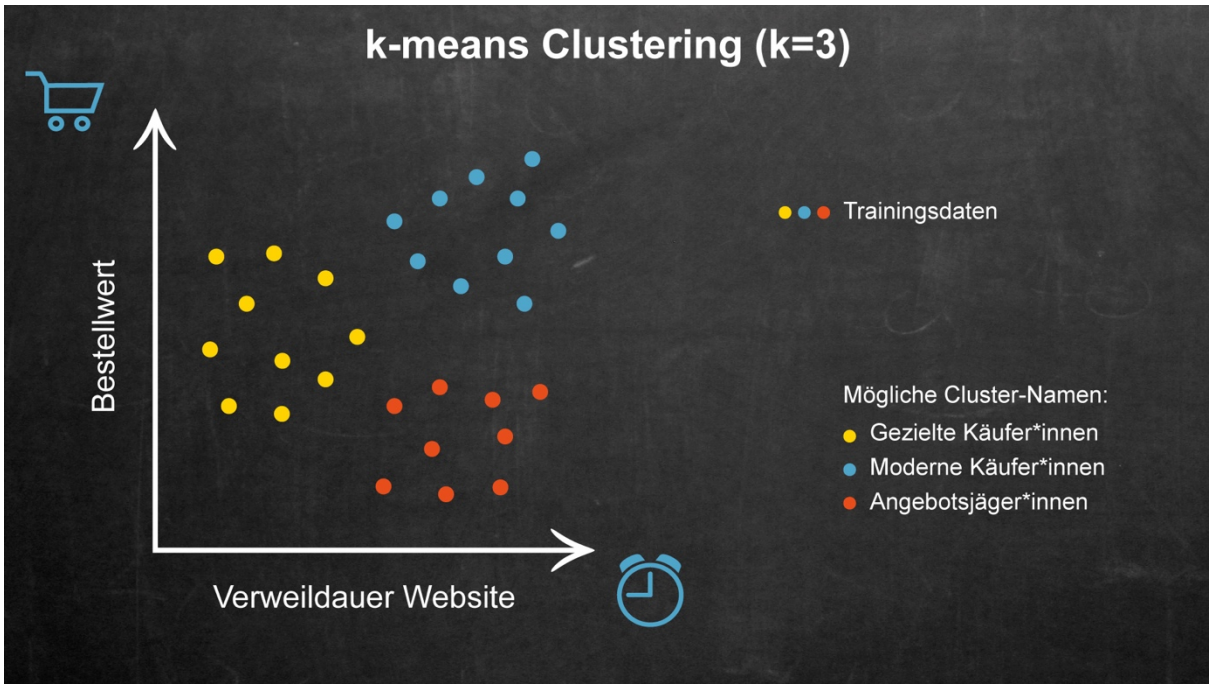
1. Vorgabe der ersten (zufälligen) k Cluster-Mittelpunkte
2. Zuordnung aller Beobachtungen zu ihrem nächstgelegenen Cluster (-Mittelpunkt)
3. Berechnung der Mittelpunkte der neu entstandenen Cluster
4. Wiederholung von Schritt 2 und 3
5. Stopp, wenn z. B. keine Umgruppierung mehr erfolgt

Insgesamt verfolgt der Algorithmus also das Ziel, dass Beobachtungen in dem gleichen Cluster sich möglichste nahe, also sich sehr ähnlich sind und Beobachtungen aus verschiedenen Clustern dagegen sehr unähnlich sind. Bei uns sollen demnach die Kunden und Kundinnen unserer Website in einem Cluster am besten eine ähnliche Verweildauer und einen ähnlichen Bestellwert aufweisen.

CC BY

Interpretation der Cluster

Die so entstandenen finalen k Cluster haben jetzt aber nicht direkt auch eine inhaltliche Bedeutung. Wir als Anwender und Anwenderinnen müssen uns diese Cluster genau ansehen, interpretieren und ggf. benennen. Man kann sich dafür z. B. die features der Beobachtungen in den einzelnen Clustern anschauen. Zudem müssen wir beurteilen, ob die gefundenen Cluster uns überhaupt sinnvoll erscheinen und uns z. B. für weitere Analysen weiterhelfen können.



Kommen wir nochmal zu unserem Beispiel der Kund*innensegmentierung zurück und sehen uns die features in den finalen Clustern an. Im blauen Cluster haben Kunden und Kundinnen für beide features eher hohe Werte, also eine eher lange Verweildauer und einen hohen Bestellwert. Dagegen haben Beobachtungen im orangenen Cluster zwar auch eine hohe Verweildauer, aber einen niedrigeren Bestellwert. Man könnte z. B. vermuten, dass Beobachtungen im orangenen Cluster eher Angebotsjäger*innen sind; so nennen wir nun auch dieses Cluster. Beobachtungen im blauen Cluster könnten eher Kund*innen sein, die nach modernen Neuheiten und damit teureren Produkten suchen. Wir nennen dieses Cluster „Moderne Käufer*innen“. Personen im gelben Cluster scheinen nur kurz zu verweilen und haben einen eher durchschnittlichen Bestellwert. Dieses Cluster könnte man vielleicht als „Gezielte Käufer*innen“ betiteln. Mit diesen Informationen könnten wir nun individuellere Werbemaßnahmen für die einzelnen Gruppen ableiten.

Ein Vorteil des k-means Clustering ist seine intuitive und nachvollziehbare Herangehensweise und die Möglichkeit, neues Wissen zu generieren. Ein Nachteil ist, dass wir eine Anzahl an Clustern vorgeben müssen, welche in der Praxis nicht immer direkt ersichtlich ist. Zudem ergeben sich nicht immer sinnvolle und interpretierbare Cluster.

CC BY

Abschluss

Wir kennen jetzt ein sehr beliebtes Verfahren aus dem unsupervised learning, das k-means Clustering. Mit ihm können wir unsere Daten in vorher unbekannte Gruppen einteilen und so neues Wissen generieren. Im Prinzip folgt der Algorithmus unserer optischen Herangehensweise an ein solches Problem.

Einblendungen drei Benutzer*innengruppen und kleine Grafik mit Clustern

Anwendungsbeispiele gibt es neben der Kund*innensegementierung z. B. im Bereich der Medizin oder Psychologie. Hier können mittels Clustering Patient*innengruppen gefunden und ähnlich zu Werbemaßnahmen gezieltere Therapieansätze abgeleitet werden.

CC BY

Weiterführendes Material

Fachbücher:

Guter Einstieg ins Thema, anschaulich erläutert, keine Formeln oder tiefergehende methodische Erläuterungen:

Müller, D. (2019). k-Means-Algorithmus – Finde deine Mitte. In K. Kersting, C. Lampert, & C. Rothkopf (Hrsg.), *Wie Maschinen lernen* (S. 81-88). Springer, Wiesbaden.

Auch wenn dieses Buch mit R anstatt Python arbeitet, anschauliche Erklärung der Methoden, tiefergehende methodische Erläuterungen:

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

- Chap. 9: Finding Groups of Data – Clustering with k-means

Klassisches Werk für Statistisches/Maschinelles Lernen, tiefergehende methodische Erläuterungen:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2. Auflage). Springer.

- Chap. 12.4: Clustering Methods

Videos/Kurse:

So lernen Maschinen: #5 Unüberwachtes Lernen – Clustering.
https://ki-campus.org/videos/solernenmaschinen

Etwas weitergehender Einstieg ins Thema, anschaulich erläutert:

AMALEA - Angewandte Machine Learning Algorithmen, Woche 3, Kap. 4: K-Means-Clustering.
https://learn.ki-campus.org/courses/amalea-kit2021/items/70bKEpLQwYK8eZG5ovlDoX

Disclaimer

Transkript zu dem Video „Woche 09 Theorie: K-means Clustering“, Katja Theune.
Dieses Transkript wurde im Rahmen des Projekts ai4all des Heine Center for Artificial Intelligence and Data Science (HeiCAD) an der Heinrich-Heine-Universität Düsseldorf unter der Creative Commons Lizenz CC-BY 4.0 veröffentlicht. Ausgenommen von der Lizenz sind die verwendeten Logos, alle in den Quellen ausgewiesenen Fremdmaterialien sowie alle als Quellen gekennzeichneten Elemente.