

Einführung Clustering

Erarbeitet von
Dr. Katarina Boland

Lernziele	1
Inhalt	2
Einstieg	2
Weiterführendes Material.....	3
Disclaimer	3

Lernziele

- Du kannst erklären, was man unter Clustering versteht
- Du kannst mindestens einen Anwendungsfall erläutern, für den Clustering verwendet werden kann

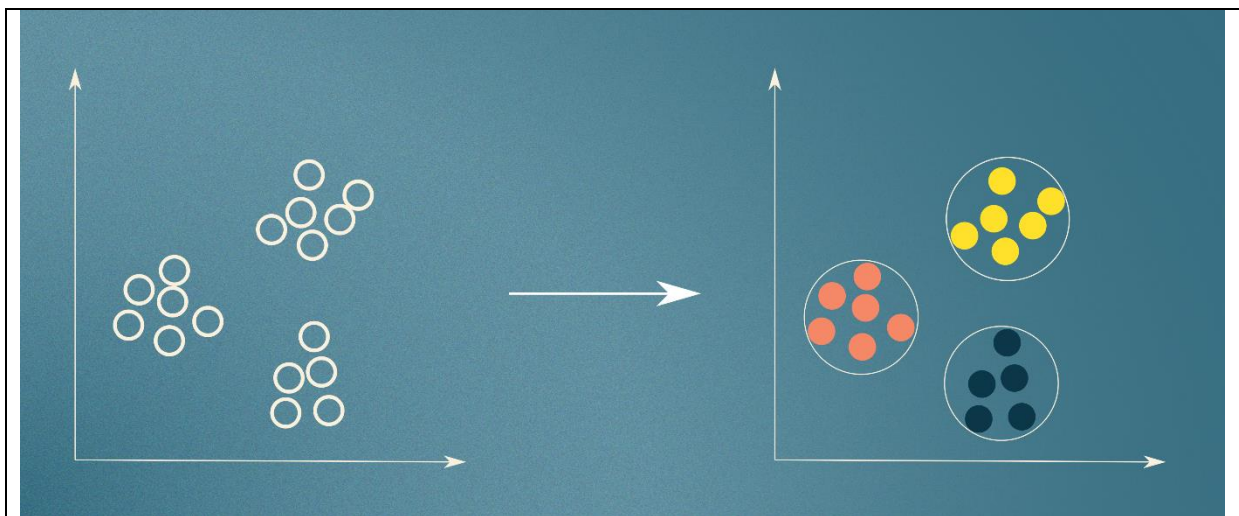
Inhalt

Einstieg

Stell dir vor, du hast eine riesige Menge von Daten, beispielsweise Texte oder Umfragedaten und du möchtest diese in verschiedene Gruppen einteilen: beispielsweise hast du Umfragedaten zum Thema „Umgang mit dem Corona-Virus“ und interessierst dich dafür, was für verschiedene Personengruppen und Strategien es hier gibt. Oder du möchtest eine Übersicht von Themen haben, die in Texten zu finden sind. Wenn du Metainformationen zur Verfügung hast, wie beispielsweise Schlagworte, ist eine Sortierung leicht vorzunehmen. Was aber, wenn du nur die Texte an sich hast und auch keine Informationen darüber, welche Themen überhaupt in den Texten vorkommen können, beispielsweise wenn du untersuchen möchtest, welche Themen zu verschiedenen Zeitpunkten in sozialen Medien diskutiert werden?

Hier kommt „Clustering“ ins Spiel.

Der Begriff „Cluster“ lässt sich übersetzen mit „Gruppe“, „Anhäufung“ oder „Einheit“. Clustering bezeichnet den Prozess, Daten in verschiedene Gruppen einzuteilen bzw. Gruppen in den Daten zu finden.



Gruppen können hier Personengruppen sein, aber auch alle sonstigen Gruppen von Dingen oder Eigenschaften, wie Gruppen von thematisch ähnlichen Texten.

Clustering kann zum Explorieren von Daten eingesetzt werden, um einen zusammenfassenden Überblick zu bekommen und die Daten zu organisieren und zu sortieren. Clustering basiert auf unüberwachtem Lernen und kommt ohne vorher bekannte Labels und Goldstandards aus.

Allerdings wirft das die Frage auf, wie die Ergebnisse von automatischem Clustering überhaupt evaluiert werden können. Dieser und weiteren Fragen widmen wir uns in den kommenden Videos.

Weiterführendes Material

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Comput. Surv., 31(3), 264–323. <https://doi.org/10.1145/331499.331504>

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C.-T. (2017). A review of clustering techniques and developments. Neurocomputing, 267, 664–681. <https://doi.org/10.1016/j.neucom.2017.06.053>

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. Trans. Neur. Netw., 16(3), 645–678. <https://doi.org/10.1109/TNN.2005.845141>

Disclaimer

Transkript zu dem Video „06 Clustering: vom Sortieren bis zum Explorieren: Einführung Clustering“, Katarina Boland.

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](https://creativecommons.org/licenses/by/4.0/) veröffentlicht. Ausgenommen von der Lizenz sind die verwendeten Logos, alle in den Quellen ausgewiesenen Fremdmaterialien sowie alle als Quellen gekennzeichneten Elemente.