Themenblock Clustering: vom Sortieren bis zum Explorieren: 06_06Diskussion_ClusVsKlass

Clustering vs. Klassifikation

Erarbeitet von
Dr. Katja Theune

Lernziele

- Du kannst den Unterschied zwischen Clustering und Klassifikation erläutern
- Du kannst erläutern, wie Clustering als Vorverarbeitungsschritt für eine Klassifikation angewendet werden kann

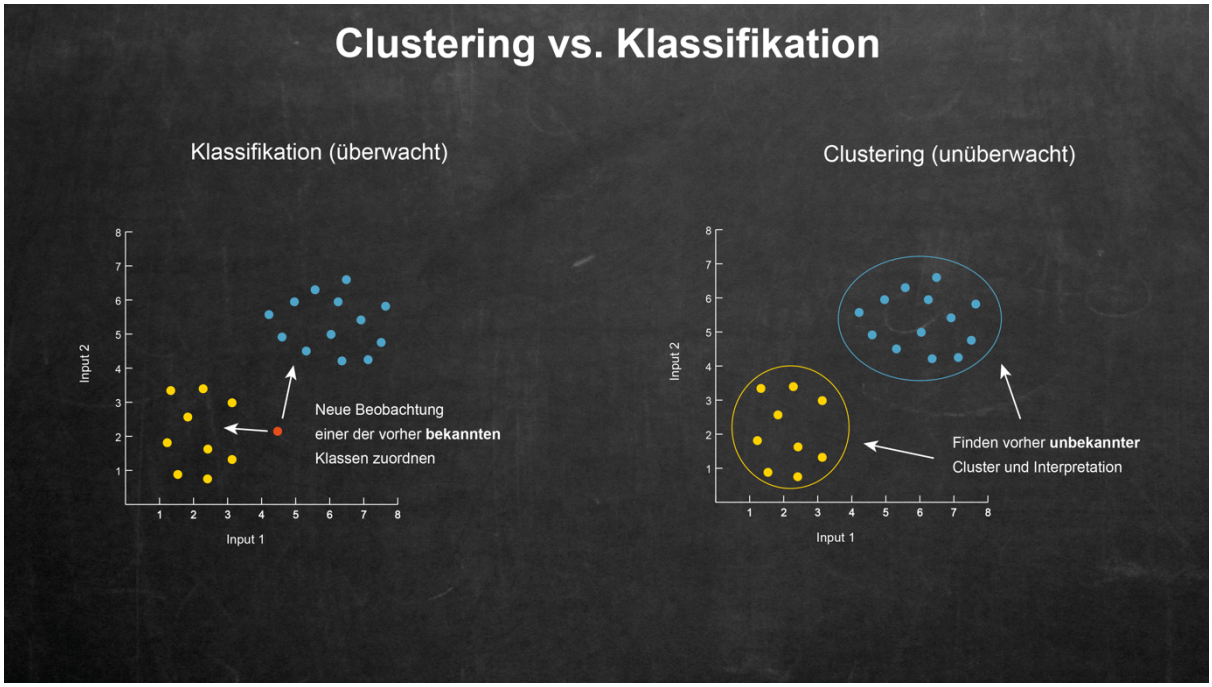
Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Clustering vs. Klassifikation

Clustering und Klassifikation haben auf den ersten Blick viele Gemeinsamkeiten. Bei beiden geht es darum, Beobachtungen in Gruppen, Klassen oder eben Cluster einzuteilen. Beobachtungen in derselben Klasse oder demselben Cluster sollen sich sehr ähnlich sein, also ähnliche Eigenschaften haben. Beobachtungen zwischen diesen Klassen bzw. Clustern sollen sich dagegen am besten deutlich unterscheiden.

Wir können hier z. B. an das k-nearest neighbours Verfahren zur Klassifikation und das k-means Verfahren beim Clustering denken. Bei beiden geht es also um Ähnlichkeiten zwischen Beobachtungen, die bei beiden Verfahren auch mit Distanzmaßen gemessen werden.



Ein bedeutender Unterschied ist aber, dass die Klassifikation zum überwachten und das Clustering zum unüberwachten Lernen gehört. Zur Klassifikation verwenden wir schon vorher definierte Outputs, also bekannte Klassen, und teilen neue Beobachtungen in diese ein. Wir kennen also auch die Anzahl an vorhandenen Klassen.

Das Clustering gehört dagegen zum unüberwachten Lernen. Hier kennen wir üblicherweise nicht vorher die vorhandenen Gruppen oder Cluster und auch nicht ihre Anzahl. Es geht mehr darum, in den Daten uns noch ganz unbekannte Strukturen, Muster und eben auch Gruppen zu finden, in die sich Beobachtungen aufgrund ihrer Ähnlichkeiten aufteilen lassen. Da diese Gruppen vorher nicht definiert sind, müssen wir sie selbst interpretieren. Üblicherweise müssen wir uns hier auch selbst für die „richtige" oder „sinnvollste" Anzahl an Clustern entscheiden.

Als Beispiel könnten wir mit Clustering-Verfahren versuchen, auf Basis bestimmter medizinischer und demographischer Merkmale verschiedene uns noch unbekannte Patient*innengruppen zu finden, um individuellere Therapiemaßnahmen abzuleiten. Bei der Klassifikation würden wir dagegen die Patient*innengruppen bzw. Klassen bereits kennen und neue Patient*innen in diese Klassen einsortieren, um eine geeignete Therapie einzuleiten.

Clustering als Vorverarbeitungsschritt

Nicht selten wird Clustering aber auch als Vorverarbeitungsschritt für eine Klassifikation verwendet. So kann man in einem ersten Schritt mit Hilfe des Clustering uns vorher noch nicht bekannte Gruppen in den Daten finden, diese interpretieren und benennen.
Diese gefundenen Gruppen können wir dann in einem zweiten Schritt als Klassen in einer überwachten Klassifikation verwenden, um z. B. ein Prognosemodell zu erstellen und die relevanten Inputs für die Klassenprognosen herauszufinden.

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: Clustering vs. Klassifikation“, Dr. Katja Theune.