KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Datenbeschaffung und -aufbereitung: 04_03Aufbereitung_Konsistenz

Konsistente Daten

Erarbeitet von
Dr. Ann-Kathrin Selker

Lernziele

- Du kannst den Zusammenführprozess wiedergeben
- Du kannst mögliche Fehler beim Zusammenführen von Daten wiedergeben

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

Inhalt

Es ist sehr selten, dass alle Daten, die du für deine Anwendung brauchst, bereits in einem einzigen Datensatz zusammengefasst sind. Nachdem du alle relevanten Datensätze gesammelt hast, musst du diese also erst einmal kombinieren, bevor die Bereinigung starten kann. In einem „Datentagebuch" kannst du genau notieren, welche Aktionen du auf deinen Daten ausgeführt hast. So kannst du jederzeit deine Ergebnisse reproduzieren und auch Fehler erkennen.

Ein guter erster Schritt beim Zusammenführen ist das Prüfen der Features. Stimmen die Feature-Bezeichnungen überein? Gibt es gleiche Feature-Bezeichnungen, die aber eigentlich etwas Unterschiedliches abbilden?

Auch sonstige Inkonsistenzen können in deinen Daten auftauchen. So können sich zum Beispiel Datumsformate unterscheiden oder manche Datensätze enthalten das Alter, andere das Geburtsdatum. Bei manchen Datensätzen unterscheiden sich bei Preisen vielleicht die Währung, und mal wird das metrische und mal das imperiale Einheitensystem verwendet. Bei manchen Ratings werden fünf Sterne benutzt, woanders vielleicht drei oder zehn. Länder können ausgeschrieben in verschiedenen Sprachen, als Ländercodes oder ISO-Kodierung referenziert werden. Namen bzw. Adressen können als Ganzes als ein Feature auftauchen oder getrennt nach Vor- und Nachnamen bzw. Straße, Hausnummer, Postleitzahl und Stadt. Du solltest dir auch bewusst sein, dass verschiedene Datensätze verschiedene Biases enthalten können, zum Beispiel wegen unterschiedlicher Messgeräte.

Auch bei unstrukturierten Daten kann es natürlich zu Problemen beim Kombinieren mehrerer Datensätze kommen. Bilder zum Beispiel brauchen in der Regel die gleiche Größe, damit sie von Machine-Learning-Modellen wie CNNs verarbeitet werden können. Falls du Bilder dann mit einem schwarzen Rahmen oder Ähnlichem vergrößerst, kann es sein, dass diese Bilder fälschlicherweise am schwarzen Rahmen und nicht an ihrem Motiv erkannt werden.

Außerdem führt das Zusammenführen von Daten oft zu Lücken: Wenn z. B. ein Datensatz Features enthält, die im anderen nicht vorhanden sind, resultiert das in fehlenden Werten. Diese Datenpunkte können dann nicht ohne Behandlung der fehlenden Werte zum Lernen verwendet werden, da sie vom zu trainierenden Machine-Learning-Modell nicht verarbeitet werden können.

Für das Zusammenführen von Datensätzen gibt es viele Programme. Du kannst deine Daten aber natürlich auch mit Python oder SQL selber zusammenführen. Dieses Video hat dir dafür häufige Fehlerquellen aufgezeigt.

Weiterführendes Material

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. & Müller, K. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-08987-4

https://realpython.com/pandas-merge-join-and-concat/

https://www.kaggle.com/code/crawford/python-merge-tutorial/notebook

Disclaimer

Transkript zu dem Video „04 Datenbeschaffung und -aufbereitung: Konsistenz“, Ann-Kathrin Selker