Woche 03 Daten: Vorverarbeitung von strukturierten Daten

Skript

Erarbeitet von
Ann-Kathrin Selker

Lernziele 1
Inhalt 1
Einstieg 2
Datenbeschaffung 2
Datenbereinigung 2
Fehlerbehandlung 4
Abschluss 4
Quellen 4
Weiterführendes Material 5
Disclaimer 5

Lernziele

- mögliche Fehlerquellen bei Daten identifizieren
- Lösungsstrategien bei Fehlern angeben können

Inhalt

Wo kommen unsere Daten überhaupt her und wie machen wir sie nutzbar?

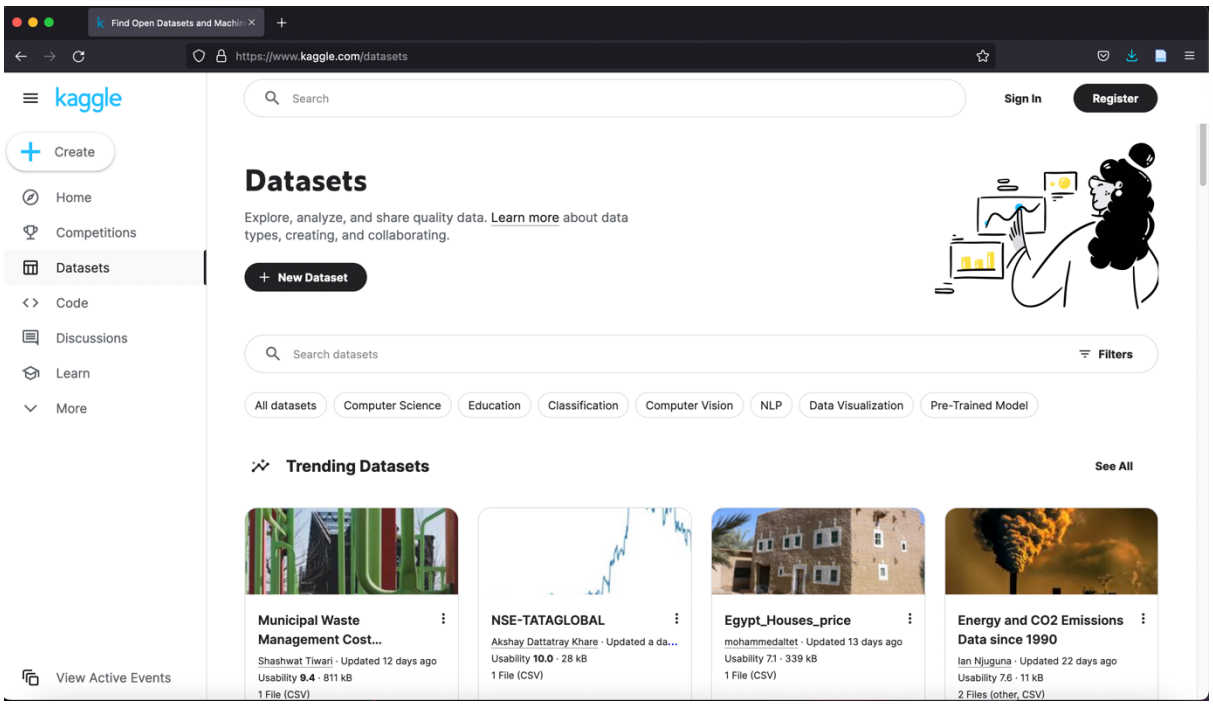
Projekt „KI für alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Einstieg

Um ein Machine-Learning-Modell zu trainieren, braucht man sogenannte Trainingsdaten. Welche Art von Trainingsdaten hängt von der späteren Anwendung ab. Ein Algorithmus, der Bilder in Kategorien einteilen soll, braucht Bilder aus diesen Kategorien, um zu lernen. Ein Algorithmus, der das Wetter für morgen vorhersagt, braucht Wetterdaten aus vergangenen Jahren, um zu lernen.

Datenbeschaffung

Ein Datenobjekt besteht aus Features, z. B. Name, Geburtsjahr, Matrikelnummer, Studiengang etc. für Studierende. Falls du die Daten selber beschaffst oder aber selbst den Auftrag gibst, ist es sehr wichtig, dir vorher gut zu überlegen, welche Features relevant für die gewünschte Anwendung sind. Falls sich später herausstellt, dass du ein wichtiges Feature vergessen hast, ist es oft zeitaufwändig und teuer, dieses nachträglich zu sammeln, wenn es denn überhaupt möglich ist. Es gibt aber auch viele öffentlich verfügbare Datensätze, zum Beispiel über die Webseite Kaggle.com in der Rubrik Datasets. Mit diesen Daten kann man unter anderem seine eigenen Fertigkeiten trainieren.



Einblendung Screenshot Kaggle.com (Quelle [1])

Datenbereinigung

Leider sind Daten nicht immer so perfekt, wie wir sie haben wollen. Daten können zum Beispiel unvollständig sein, sie können fehlerhaft sein oder auch einfach teilweise irrelevant

für die Aufgabe, die zu erfüllen ist. Hier muss dann vor dem Trainieren nachgeholfen werden. Es gilt die Faustregel: „Müll rein, Müll raus“. Wenn mit Daten von schlechter Qualität trainiert wird, wird das Ergebnis am Ende auch schlecht werden. Achtung, das Bereinigen der Daten ist nicht nur sehr wichtig, sondern häufig auch der aufwändigste Schritt beim Erstellen von Machine-Learning-Algorithmen!

In diesem Video beschäftigen wir uns mit Beispielen von strukturierten Daten. Ich zeige einige Fehler, die häufig auftreten können. Es kommt natürlich immer auf die spätere Anwendung an, ob diese Fehler korrigiert werden müssen und falls ja, wie viel Zeit darin investiert werden sollte. Angenommen, dein Datensatz ist diese Tabelle.

Matrikelnummer	Name	Vorname	Wohnort	Studiengang	letzte Rückmeldung
1234567	Müller	Jana	Düsseldorf	Informatik	
1234568	Müller	Janina	Düsseldorf	Informatik	
Park	1337788	Choi	Ratingen	Medizin	01.09.22
21136790	Patel	Rishi	Neuss	Wirtschaftsmathematik	25.09.22
	Wu	Stefan		Anglistik	2022-08-14
2244511	Ivanova	Valeria	Düsseldorf	Jura	
2311198	Wagner	Milan	Köln	Philosophie	

Einblendung Tabelle mit Studierendendaten

Sie ist natürlich viel zu klein, um damit später tatsächlich ein Machine-Learning-Modell sinnvoll trainieren zu können. Als erstes gucken wir uns an, welche Daten unvollständig sind. Hier können wir zeilen- und spaltenweise vorgehen. Jede Spalte entspricht einem Feature.

Sehen wir uns zuerst leere Felder der Tabelle an. Bei Stefan Wu fehlen sowohl die Matrikelnummer als auch der Wohnort. Die Matrikelnummer lässt sich eventuell recherchieren, zum Beispiel durch Vergleiche mit anderen Datensätzen. Hier muss immer der Kostenaufwand beachtet werden. Der Wohnort kann auch approximiert werden: Studierende der Regel wohnen in Düsseldorf oder naher Umgebung, sodass Düsseldorf hier ein guter Kandidat für den richtigen Wohnort ist. Aber Achtung: Alle Änderungen an den Daten können Auswirkungen haben! Eine Möglichkeit, mit fehlenden Daten umzugehen, ist es, die entsprechenden Datenobjekte oder Features mit vielen Lücken einfach zu löschen. Je nachdem, wie viele Datenobjekte bzw. Features dies betrifft, kann dadurch aber ein Großteil der Daten verloren gehen! Manchmal fehlen Daten, weil es sie nicht gibt, zum Beispiel beim Feature "letzte Rückmeldung". Studierende im ersten Semester, die sich noch nie zurückgemeldet haben, haben hier natürlich kein Datum stehen. Hier müsstest du dir dann gut überlegen, wie du mit dem Fall umgehst.

Fehlerbehandlung

Als Nächstes gucken wir uns mögliche Fehler in den Daten an. Auch hier musst du wieder abwägen, wie viel Zeit du aufwendest und welchen Nutzen dir die Fehlerkorrektur bringt. Ein häufiger Fehler sind falsche Datentypen in Spalten. Das Feature Matrikelnummer enthält siebenstellige Zahlen. Wir können aber sehen, dass in einer Zeile eine achtstellige Zahl und in einer Zeile ein String steht. Hier handelt es sich also um Fehler. Manche Features können nur vorher festgelegte Werte annehmen, wie zum Beispiel das Feature Studiengang. Rishi Patel belegt angeblich den Studiengang Wirtschaftsmathematik, den es aber an der HHU nicht gibt. Hier muss auch nachgebessert werden. Beim Feature letzte Rückmeldung sehen wir, dass unterschiedliche Datumsformate verwendet wurden, die vereinheitlicht werden müssen. Solche Fehler treten auch häufig bei Preisen auf, wenn unterschiedliche Währungen verwendet und umgerechnet werden müssen.

Häufig müssen wir Daten aus verschiedenen Quellen zusammenführen. Hier solltest du darauf achten, welche Features gleich, aber vielleicht anders benannt sind, und welche gleich benannt sind, aber unterschiedliche Inhalte haben. Außerdem können so Duplikate auftreten, die entfernt werden sollten. Aber Achtung: Jana Müller mit Matrikelnummer 1234567 und Janina Müller mit Matrikelnummer 1234568 können Duplikate mit Tippfehlern sein, aber auch Zwillinge, die sich zusammen an der Uni eingeschrieben haben.

Im vorliegenden Beispiel habe ich Fehler per Hand gesucht und bereinigt. Bei der Anzahl der Datenobjekte, mit denen später trainiert wird, ist dies nicht ohne einen unzumutbaren Arbeitsaufwand zu machen. Hilfe bietet zum Beispiel Python mit extra Funktionen zur Datenbereinigung. Es gibt aber auch automatische Datenbereinigungstools.

Je nach Anwendung werden nicht alle Features benötigt, die vorliegen. Das Entfernen von irrelevanten Features und Datenobjekten beschleunigt nicht nur das Training, sondern kann auch die Genauigkeit des trainierten Modells erhöhen.

Abschluss

In diesem Video habt ihr gelernt, mögliche Fehlerquellen bei Daten zu identifizieren und Lösungsstrategien für diese Fehler anzugeben.

Quellen

Quelle [1] kaggle.com, abgerufen am 06.10.2022

Weiterführendes Material

https://www.kaggle.com/getting-started/250322
https://www.kaggle.com/learn/data-cleaning

Disclaimer

Transkript zu dem Video „Woche 03 Daten: Vorverarbeitung von strukturierten Daten“, Ann-Kathrin Selker.