Hintergrundwissen

STAND: 15.05.2024

Projekt KILE – Modul Vorurteile (Bias) im maschinellen Lernen

Bundesarbeitskreis Arbeit und Leben e. V.
Hochschule für Technik und Wirtschaft Berlin

Was ist maschinelles Lernen?

Maschinelles Lernen (ML) ist ein Teilbereich der Informatik und befasst sich damit, wie Modelle (Abbilder der Realität, bspw. Prozesse) aus Daten erlernt werden können. Typische Bezeichnungen sind Machine Learning (ML, eng.) und Künstliche Intelligenz (KI), bzw. Artificial Intelligence (AI, eng.). Hierbei ist allerdings zu beachten, dass ML nur ein Teilbereich der Künstlichen Intelligenz ist.

Es gibt viele Prozesse, die bspw. im Sinne der Automatisierung von Arbeitsaufgaben, mit Hilfe von ML modelliert werden. Ein gängiges Beispiel ist die Klassifikation: die Einordnung von Objekten in festgelegte Kategorien (Klassen), welche im industriellen Bereich bspw. genutzt wird um Bauteile anhand von Kamerabildern in „defekt“ und „nicht defekt“ einzusortieren. Im Vergleich zu klassischen Programmen wird beim maschinellen Lernen nicht die komplette Funktionalität der Anwendungen programmiert. Stattdessen wird nur definiert wie das Programm sich zu verhalten hat und welche Daten für das Erlernen dieses Verhaltens verwendet werden. Während des Lernprozesses – welcher auch „Training“ genannt wird – bildet das Programm eigene Regeln und strukturiert sich selbst, um die Aufgabe zu lösen. So werden schrittweise Muster in den Daten erlernt. Bei Bilddaten handelt es sich hierbei z. B. um Farbe, Formen und Strukturen. Ein klassisches Programm hingegen folgt immer einer gewissen Logik, die man als Mensch nachvollziehen kann, aber auch genauestens definiert werden muss, um gute Ergebnisse zu ermöglichen. Die genaue Funktionsweise der ML-Programme sind für Menschen – je nach verwendeter Methode – nicht so einfach nachzuvollziehen, da sie wie bspw. bei neuronalen Netzten auf erlernten Zahlenwerten langer mathematischer Formeln basieren und keiner einfach verständlichen Logik folgen. Das führt auch zu verschiedensten Problemen. Durch die komplexere Funktionsweise des maschinellen Lernens und große Datenmengen kann man heutzutage zwar eindrucksvolle Ergebnisse erzielen, hat aber oft auch mit der Korrektheit der Systeme zu kämpfen. Oftmals ist die Ursache dieser Probleme in den für das Lernen genutzten Daten zu finden. Wenn bspw. Informationen abgerufen werden sollen, die das Modell nicht kennt, wird die nächstwahrscheinliche Antwort gegeben. Man kann es sich so vorstellen, wie ein Schulkind, das für einen Test nicht gelernt hat und versucht mit dem vorhanden Wissen Fragen zu beantworten.

Generell lässt sich maschinelles Lernen bezüglich der Einsatzzwecke der resultierenden Modelle in mehrere Unterkategorien einteilen. Dazu gehören bspw. die bereits erwähnte Klassifikation, die Regression (die Vorhersage von Werten anhand anderer Informationen, bspw. Schuhgröße zu Körpergröße) und Generierung (bspw. Textgeneratoren wie ChatGPT oder Bildgeneratoren wie DALL-E). Der Workshop befasst Großteils mit der Klassifikation, welche sich besonders gut eignet um einfach und anschaulich die Funktionsweise von ML zu erklären.

Was ist Machine Learning for Kids (ML4Kids)?

Machine Learning for Kids ist eine Web-Plattform, die kostenlos für Bildungsinstitutionen zur Verfügung steht, um Schülerinnen und Schülern Konzepte des Maschinellen Lernens näher zu bringen. Dort hat man die Möglichkeit seine eigenen Modelle zu trainieren, anzupassen und zu testen. Die hierbei resultierenden Modelle können zwar nicht mit realen Anwendungen von maschinellem Lernen mithalten, liefern aber ausreichend gute Ergebnisse, um einen Blick hinter die Kulissen zu ermöglichen. Unterstützt werden Texte, Zahlen und Bilder. Dabei hat man die Möglichkeit Daten aus dem Web, sowie von dem eigenen Gerät hinzuzufügen.

In diesem Workshop werden wir uns auf Bilder beschränken, da sich die Lerninhalte damit am einfachsten vermitteln lassen. Allerdings wäre es auch möglich die vorgefertigten, in ML4Kids zur Verfügung stehenden, Beispiele zu nutzen oder sich sein eigenes Konzept zu erstellen.

Was ist ein Bias?

Bias ist ein aus dem Englischen kommender Begriff, der zu Deutsch „Voreingenommenheit“ oder „Vorurteil“ bedeutet. Allgemein ist damit eine positiv oder negativ verzerrte Wahrnehmung der Realität gemeint. Mit dem Begriff „Voreingenommenheit“ sind zwar häufig Vorurteile gegenüber anderen Menschen konnotiert, der Begriff lässt sich aber auch weiterspannen. Generell geht es hierbei um Assoziationen oder Weltbilder, die man (falsch) erlernt hat und welche einen Einfluss auf die eigenen Entscheidungen ausüben. Ein Beispiel sind Dachformen von Häusern. Ist jemand in einer Gemeinde aufgewachsen, wo jedes Haus ein spitzes Dach hat, so wird sie oder er der Meinung sein, dass Häuser vornehmlich derartige spitze Dächer haben. Dieses Modell der Wirklichkeit kollidiert mit der Realität, sobald die Person merkt, dass in anderen Regionen Häuser mit flachen Dächern durchaus üblich sind.

Voreingenommenheit, egal ob beim Menschen oder der Maschine, ist in vielen Bereichen möglich. Bekannt sind insbesondere Vorurteile auf Basis von Alter, Geschlecht, Herkunft, Hautfarbe und Religion. Allerdings können sich Vorurteile überall verstecken und sind zum Teil nicht so einfach zu erkennen.

Gründe für Bias

Beim maschinellen Lernen finden sich die Ursachen für Voreingenommenheit häufig bereits in den Daten. Sind diese nicht repräsentativ oder ungünstig gewählt, kann sich ein ML-Modell Verzerrungen der Realität aneignen. Derartige Verzerrungen können eine Vielfalt von Gründen haben: die Nutzung historischer, veralteter Daten; eine Über- oder Unterrepräsentation bestimmter Klassen; ungenaue Messung; unbetrachtete Variablen; problematische Algorithmen oder Modellevaluation; fehlerhafte Benutzung und Feedbackschleifen. Historische Verzerrungen basieren bspw. auf gesellschaftlichen und technischen Fortschritten bzw. Veränderungen. So sind Daten oft nicht mehr für die Gegenwart repräsentativ, wenn sie in der Vergangenheit erhoben wurden. Bei Repräsentationsverzerrungen hingegen sind gewisse Gruppen innerhalb der Daten häufig nicht ausreichend repräsentiert. Mit derartigen Datensätzen trainierte Modelle können mit Beispielen aus den unterrepräsentierten Gruppen entsprechend nicht gut umgehen und treffen falschen Entscheidungen. Bei Feedbackschleifen beeinflussen aktuelle Modelle sich selbst bzw. verzerren die Trainingsdaten zukünftiger Modelle. Dies lässt sich gut anhand von Produktempfehlungen im Onlinehandel verstehen. Empfiehlt ein Algorithmus bestimmte Produkte, führt das oft dazu, dass diese Produkte häufiger gekauft werden. Ist die Kaufhäufigkeit gleichzeitig ein Kriterium für den Empfehlungsalgorithmus, so werden häufiger verkaufte Produkte wiederum häufiger empfohlen. So entsteht eine Feedbackschleife aus Empfehlung und Verkauf, welche dazu führen kann, dass eine kleine Auswahl von Produkten bevorzugt empfohlen wird, während andere Artikel wenig Präsenz bekommen.

Vorurteile vermeiden

Um Verzerrungen und Vorurteile (Bias) in maschinell lernenden Systemen zu umgehen, gibt es verschiedene Ansätze.

Bei der Nutzung existenter Datensätze sollte generell überprüft werden, woher die Daten kommen und wie die Verteilung der einzelnen Gruppen innerhalb der Daten aussieht. Wurden alle möglichen Gruppen mit aufgenommen? Gibt es Gruppen, die unter- oder überrepräsentiert sind? Waren die Erzeuger des Datensatzes selbst voreingenommen?
Sammelt man selbst Daten, sollte man sich der verschiedenen Gründe von Verzerrungen sowie der eigenen potenziellen Voreingenommenheiten bewusst sein. Entsprechend empfiehlt es sich mit einem diversen Team zu arbeiten, um möglichst viele Blickwinkel abzudecken.
Ist ersichtlich, dass in existenten Daten Verzerrungen vorliegen, gibt es verschiedene Möglichkeiten den Datensatz anzupassen, um die Daten fairer zu verteilen. So können weitere Daten gesammelt werden, um unterrepräsentierte Gruppen stärker auszuprägen. Gleiches kann auch durch eine Reduktion der überrepräsentierten Gruppen erfolgen. Je nach Lernprozess, können

den Daten auch während des Lernens unterschiedliche Gewichtungen gegeben werden um das die Genauigkeit des Modells zu verbessern.

Weitere Informationen

- Link: Bias & Fairness in KI-Systemen

Fachbegriffe		
Machine Learning	Maschinelles Lernen	ML
Artificial Intelligence	Künstliche Intelligenz	AI / KI
Bias	Voreingenommenheit, Vorurteil	
Modell	Ein Abbild der Realität. Im Kontext von ML: eine erlernte Abbildung eines realen Prozesses	
Training	Lernprozess des Modells	
Neuronales Netzwerk	Verknüpfung mehrerer mathematischer Funktionen	
Large Language Model	Großes Textmodell, zum Erstellen und Kategorisieren von Text	LLM