KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock 10 Robust/Hybrid/Robust AI: 10_01Frage_Interview

Interview zur Interpretierbarkeit von künstlicher Intelligenz

Erarbeitet von
Marc Feger M.Sc.

Lernziele

- Du kannst, die Grundkonzepte des Themas „Interpretable AI“ erläutern
- Du verstehst, dass Modelle des überwachten Lernens oft als Black-Boxen verstanden werden und daher einer sorgfältigen Interpretation bedürfen
- Du erkennst, dass es nicht ausreicht, Modellen allein aufgrund gut erscheinender Ergebnisse Vertrauen zu schenken

Inhalt

Einstieg

Quelle [1]

Marc: Ich begrüße euch herzlich zu unserem heutigen Interview im Rahmen des Vertiefungskurs „KI für alle“, das sich spezielleren Themen im Bereich der künstlichen Intelligenz widmet. In diesem Interview werden wir das Thema „Die Interpretierbarkeit“ von KI-Modellen erkunden. Mein Name ist Marc Feger und ich darf unseren heutigen Gast Christopher Molnar begrüßen. Christoph ist ein angesehener Experte rund um unser heutiges Thema und Autor des Buches „Interpretable Machine Learning, A Guide for Making Black Box Models Explainable“. Wir werden mit Christoph über die Bedeutung der Interpretierbarkeit von KI-Modellen sprechen, die mittlerweile in vielen Anwendungen von KI eine entscheidende Rolle spielt.

Interview

Quelle [1, 2, 3, 4]

Marc: Lasst uns einen Moment innehalten und an Kurs 1 zurückdenken. Dort haben wir die Grundlagen von Machine Learning besprochen, also Techniken, die Modelle befähigen, aus Daten zu lernen und basierend darauf, Vorhersagen zu treffen. Lasst uns einen genauen Blick auf das überwachte Lernen werfen, in erster Linie auf Klassifikation. Überwachtes Lernen bietet eine Vielzahl von Möglichkeiten und kann sich dabei in vielen Bereichen bereits durchaus mit menschlicher Leistung vergleichen lassen. Einmal trainiert müssen diese Modelle in der Regel nicht mehr modifiziert werden, was sie zu sogenannten „Push-Button Modellen“ macht. Allerdings hat das überwachte Lernen auch Nachteile.

So lässt sich die zentrale Frage, was genau die Modelle aus den Daten gelernt haben, oftmals nicht beantworten. Hierdurch ist es schwer, die genauen Lagen nachzuvollziehen und diese Modelle werden somit als Black Boxen wahrgenommen.

In unserem Gespräch werden wir diese Vor- und Nachteile genauer beleuchten und darüber diskutieren, wie wir überwachte Lernmodelle interpretierbar machen können. Wann sollte ich einem Modell vertrauen? Dazu zwei typische Beispiele. In unserem ersten Beispiel geht es um die Verwendung von Machine Learning, um die biologischen Eigenschaften anhand der Augenstruktur zu identifizieren.

Ähnliche Vorgehen gibt es bereits für Chromosomen und Knochen-Eigenschaften, beispielsweise im Bereich der Kriminalistik. Das klingt zunächst vielversprechend, führt aber zu einem Problem. Der Klassifikator, der für diese Aufgabe entwickelt wurde, hat tatsächlich gelernt, Mascara zu erkennen, anstatt das biologische Geschlecht korrekt zu bestimmen. In unserem zweiten Beispiel geht es um die Früherkennung von Hautkrebs in medizinischen Aufnahmen. Hier tritt eine andere Schwierigkeit auf. Der Klassifikator erkennt nicht den Krebs selbst, sondern die Hilfsmarkierung von Voruntersuchungen.

Christoph, in deinem Buch betonst du die Bedeutung der Interpretierbarkeit von Machine-Learning-Modellen. Kannst du uns näher erläutern, was du mit dem Aspekt der Interpretierbarkeit meinst und wie wir Modelle in solchen Fällen effektiv erklären können? Daher die Frage also, warum vertrauen wir nicht einfach unseren Modellen?

Christopher: Ja, genau. Interpretierbarkeit, da fängt erstmal die Schwierigkeit an. Was ist eigentlich Interpretierbarkeit? Da gibt es nämlich jetzt nicht so eine klare mathematische Definition, sondern so eine Definition von dem Begriff ist, dass der Grad, zu dem man Entscheidungen z. B. verstehen kann oder auch generell den Entscheidungsmechanismus eines Modells. Also man sieht schon, da kommt es auch darauf an, was können Menschen eigentlich verstehen, weil das ja auch z. B. unterschiedlich ist zwischen verschiedenen Personengruppen. Die einen können vielleicht mit einer mathematischen Formel umgehen, die anderen nicht. Ich persönlich sehe den Begriff Interpretierbarkeit als so ein Oberbegriff, um verschiedene Methoden und Ansätze zu sammeln, die versuchen, die Entscheidungen von Machine-Learning-Modellen irgendwie verständlich zu machen, zusammenzufassen, sodass man die groben Entscheidungen des Mechanismus versteht und zum Teil auch einzelne Vorhersagen erklären kann.

Jetzt zur Frage, warum vertraut man nicht einfach überhaupt einem Model. Also die erste Instanz wäre, dass man guckt, wie gut ist das Modell, also wie gut klassifiziert das oder wie gut macht es Vorhersagen. Und dann könnte man jetzt sehen, das Modell funktioniert sehr gut, aber das Problem ist auch hinter so einer guten Performance kann sich eben verstecken, dass das Modell Abkürzungen nimmt. Oder man irgendeine andere Art von Fehler vielleicht auch als Entwickler eingebaut hat. Zum Beispiel, dass mit dem, das bei Krebs-Klassifikationen die Marker genommen werden zum Klassifizieren. Und statt zum Beispiel, um Hautkrebs zu klassifizieren, die eigentlichen Hautstellen zu untersuchen. In diesem Fall würden wir es nicht erkennen, ohne zusätzliche Tools, dass das Modell diese Abkürzung nimmt. Das heißt, wir brauchen irgendeine Form, um ins Modell hineinzuschauen, um zu verstehen, was passiert und auf welchen Entscheidungsregeln und Mechanismen es tut.

Um das zu tun hat man so grundlegend drei verschiedene Ansätze. Der erste Ansatz ist, dass man sagt, man nimmt nur Modelle her, die grundsätzlich so eine einfache Struktur haben, dass man am Ende auch noch versteht, wie die Entscheidung, die Vorhersage, zustande kommt. Also wenn man zum Beispiel an so etwas wie Entscheidungsbäume denkt oder Lineare-Regressionsmodelle. Aber da kommt es natürlich auch wieder darauf an, dass die Anwender das dann verstehen, das Modell, also nicht jeder sieht eine Liste von Koeffizienten und weiß dann direkt, wie man das interpretiert. Und da ist es dann auch fließend, weil Entscheidungsbaum, wenn der sehr tief wird und sehr verästelt, dann kann man den natürlich auch nicht mehr so leicht interpretieren. Der zweite Ansatz ist, dass man versucht, die Modelle, die man schon hat, dass man da versucht, einzelne Aspekte daraus zu verstehen. Also zum Beispiel bei einem Neural Network, das Bilder klassifiziert, könnte man jetzt versuchen, einzelne Neuronen in dem Netzwerk irgendeine Bedeutung zuzuweisen, indem man zum Beispiel ein Bild sucht, dass dieses Neuron maximal aktiviert. Das sind so diese modellspezifischen Ansätze und dann gibt es noch den dritten Teil, das sind sogenannte modellagnostische Ansätze, bei denen sagt man, okay das Modell ist komplex

und eine Blackbox und ich will da gar nicht reinschauen. Anstatt dessen schaue ich mir an, wenn ich meine Eingabedaten verändere, wie verändert sich dann die Vorhersage. Und daraus kann ich dann auch relativ viel ableiten, zum Beispiel, was die wichtigsten Faktoren waren, also die wichtigsten Features, welche, bei Bildklassifikation, zum Beispiel, welche Teile vom Bild wichtig waren für die Entscheidungen usw.

Marc: Können alle Modelle ohne Weiteres interpretiert werden? Ich frage mich, ob es möglich ist, sowohl Entscheidungsbäume als auch neuronale Netze gleichermaßen zu interpretieren, ohne spezielle Unterschiede zwischen ihnen zu berücksichtigen.

Christopher: Ja, das ist eine gute Frage. Bei der Größe, also da kommt es wieder drauf an, welche Strategie man nimmt. Wenn man von vornherein schränkt man seine Modellwahl ein auf diese interpretierbaren Modelle oder nimmt man eher diese modellagnostischen Ansätze. Bei dem modellagnostischen Ansatz, da ist es tatsächlich egal, welches Modell zugrunde liegt. Der einzige Unterschied ist, welchen, wenn man verschiedene Datentypen hat, also eine Erklärung für ein Bild-Klassifikationsmodell sieht natürlich anders aus wie für tabellarische Daten, weil einfach die Daten auch schon anders aussehen und die Erklärungen an sich sind immer auf der Ebene oder meistens auf der Ebene von den Eingabefeatures. Das heißt, wenn man ein Bild erklären möchte, bekommt man dann, oder sollte man Methoden benutzen, die dann Teile auf dem Bild highlighten und bei tubulären Daten hat man dann eher Methoden, die einzelnen Spalten aus dem Datensatz dann zum Beispiel eine Wichtigkeit zu weisen oder dann, wenn man einzelne Vorhersagen erklären möchte, dann diese Vorhersagen aufteilen sozusagen auf welches Feature wie relevant war für die Vorhersage. Aber natürlich je komplexer das Modell wird, desto weniger können wir diesen ersten Aspekt nutzen, dass man ins Modell reinschaut, um versucht was zu verstehen. Also wenn das Ursprungsmodell einfach ein Entscheidungsbaum war, können wir uns mal plotten lassen, gucken, welche Entscheidungsregel hat das Modell gelernt und daraus auch versuchen, das Modell besser zu verstehen. Wenn wir aber ein tiefes neuronales Netzwerk haben, da können wir sozusagen nicht einfach einzelne Teile so leicht anschauen und direkt verstehen, wofür die sind und dann das Modell dadurch besser verstehen. Und je komplexer ein Modell ist, desto mehr Interaktionen und so weiter, kann es auch modellieren zwischen den Eingabedaten. Das heißt, auch mit diesen modellagnostischen Ansätzen, die erstmal nicht unterscheiden zwischen verschiedenen Modelltypen, also die auf jeden Modelltypen anwendbar sind, die sind ja auch immer nur Vereinfachungen vom Modell. Und je komplexer ein Modell ursprünglich war, desto mehr verliert man sozusagen und die Erklärung ist immer nur eine Art Zusammenfassung oder eine sehr grobe Erklärung, was passiert oder wie das Modell Entscheidungen trifft.

Marc: Christoph, wenn du die Bedeutung der Tierbarkeit in einem übergeordneten Kontext betrachten würdest, welchen Stellenwert würdest du ihr zuschreiben?

Christopher: Ja, meine Antwort ist wahrscheinlich ein bisschen biased, weil ich so tief drin bin in dem Thema. Aber der Grund, warum ich auch tief drin bin in den Themen ist, weil ich denke, dass es sehr wichtig ist. Also Interpretierbarkeit hat auch ganz verschiedene Bedeutungen. Also wir haben Interpretierbarkeit für verschiedene Gründe. Das fängt an beim Entwickeln von den Modellen, da kann man Interpretierbarkeit benutzen, um so eine Art Debugging zu betreiben, Auditieren von Modellen, also man kann dann eben solche

Fehler schon direkt finden zu dem Zeitpunkt. Man kann Interpretierbarkeit aber auch benutzen, um mehr über die Daten zu lernen. Also wenn man zum Beispiel auch so Modelle in der Wissenschaft benutzt, Machine Learning und das Modell eigentlich mehr Mittel zum Zweck ist und man herausfinden möchte, was in den Daten so passiert, dann kann man mit Interpretierbarkeit auch noch mal sein Modell ein bisschen durchleuchten, also zum Beispiel gucken, was waren die wichtigsten Input-Features, um daraus zu lernen. Aber auch, wenn zum Beispiel Personen von Modell Entscheidungen betroffen sind, also so dieses typische Beispiel ist okay, der Kredit wurde abgelehnt, weil das Modell gesagt hat, eine Person ist nicht kreditwürdig. Da kann man dann auch Erklärungen erzeugen, um zu sagen, okay, warum wurde jetzt diese Kreditanfrage abgelehnt. Also Interpretierbarkeit spielt in verschiedenen Ebenen schon, wie ich denke, eine wichtige Rolle.

Marc: Zum Ende unseres Videos, was möchtest du unseren Zuschauer mit auf den Weg geben?

Christoph: Stand ist momentan, dass es jetzt relativ viele Tools gibt. Also vor ein paar Jahren, als ich angefangen habe dazu zu forschen, da gab es jetzt eigentlich auch noch nicht so viele. Mittlerweile gibt es da richtig viele Tools und die sind auch gar nicht so schwer zu benutzen. Und auch wenn man praktische Projekte macht, also ich mache auch gerade wieder so ein Machine Learning Challenge mit. Es bringt einfach auch unglaublich viel zur Modellentwicklung, wenn man Interpretierbarkeitsmethoden benutzt, weil grundsätzlich Machine Learning, da kann man natürlich immer einfach einen Algorithmus draufwerfen und erzeugt dann irgendeine Vorhersage und es sieht dann erst mal gut aus, aber wenn man dann nicht weiterweiß, dann ist es oft irgendwie cool, wenn man so einen Einblick auch sozusagen bekommt, was das Modell eigentlich macht, was die wichtigsten Features waren. Und meine Empfehlung ist auch, die Methoden einfach mal auszuprobieren. Da, gerade wenn man praktische Anwendung hat, das einfach mal zu benutzen, vor allem die modellagnostischen Methoden, die kann man dann auch immer so im Nachhinein noch nutzen, wenn das Modell schon trainiert ist. Und ich denke, es hat auch einen großen Mehrwert, wenn man das macht, um das Modell zum Beispiel zu verbessern.

Marc: Christoph, vielen Dank, dass du an unserem kurzen Interview teilgenommen hast. Wenn euch das Thema Interpretierbarkeit von KI-Modellen weiter interessiert, freuen wir uns, dass ihr unsere folgenden Videos rund um das Thema und in die weitere Literatur von Christoph reinschaut. Bis dahin.

Quellen

Quelle [1] Molnar, C. (2024, May 26). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Quelle [2] Kuehlkamp, A., Becker, B., & Bowyer, K. W. (2017). Gender-From-Iris or Gender-From-Mascara? http://arxiv.org/abs/1702.01304

Quelle [3] Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein,T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., & Haenssle, H. A. (2019). Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. https://doi.org/10.1001/jamadermatol.2019.1735

Quelle [4] Schnellbacher, M. (2016, März 15). AlphaGo schlägt weltbesten menschlichen Go-Spieler. https://entwickler.de/programmierung/alphago-schlagt-weltbesten-menschlichen-go-spieler

Quelle [5] https://huggingface.co

Weiterführendes Material

Molnar, C. (2024, May 26). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Disclaimer

Transkript zu dem Video „Themenblock 10 Robust/Hybrid/Robust AI: 10_01Frage_Interview“, Marc Feger.