Interpretierbare KI

Neue Methoden zeigen Entscheidungswege künstlicher Intelligenz auf

Machine Learning erzeugt heute KI-Systeme, die Entscheidungen schneller treffen als ein Mensch. Darf dieser sich aber entmündigen lassen? Neue Methoden machen Entscheidungswege transparent und nachvollziehbar und schaffen damit Vertrauen und Akzeptanz – oder sie decken Missverständnisse auf.

Von Andreas Holzinger

Autonomes Fahren, Gesichtserkennung, Sprachverstehen und Empfehlungssysteme werden heute als KI-Systeme verwirklicht, zumeist in Form neuronaler Netze. Diese entstehen nicht durch manuelle Programmierung, sondern durch maschinelles Lernen, indem sie automatisiert mit vorgegebenen großen Mengen an Beispieldaten trainiert werden – das sogenannte Deep Learning. Am Ende ist allerdings nur sehr schwer nachzuvollziehen, wie neuronale Netze tatsächlich Entscheidungen treffen, zahlreiche versteckte Ebenen zwischen Eingabe- und Ausgabeschicht und Millionen von Parametern machen sie zu äußerst schwer zu durchschauenden Black-Box-Modellen.

Der Grund, warum Deep-Learning-Verfahren trotzdem zunehmend eingesetzt werden: Abstrakte Algorithmen finden in komplexen und hochdimensionalen Datenmengen Muster, die kein Mensch jemals in der Lage wäre zu entdecken. Also wird den Algorithmen die Lösungsfindung zwangsläufig überlassen.

Hier hat das Forschungsgebiet der „explainable artificial intelligence“, also einer interpretierbaren KI, seinen Ursprung. Transparente Entscheidungswege stellen eine Riesenchance für KI-Lösungen dar. Die ihnen vorgeworfene Undurchsichtigkeit könnte vermindert und die Akzeptanz bei den Nutzern gefördert werden. Denn die Angst vor einem Kon-

trollverlust des Menschen durch KI ist groß. Themen wie das autonome Fahren und die unklare Entscheidungsfindung des Fahrzeugs, beispielsweise im Extremfall kurz vor einer Unfallkollision, stehen längst in der öffentlichen Diskussion. Ebenso die Frage, inwieweit KI medizinische Entscheidungen unterstützen oder sogar selbst treffen darf.

Die EU-Datenschutzgrundverordnung macht seit Mai mit dem „Recht auf Erklärung“ Transparenz notwendig, das heißt, auf Anfrage sind die zugrunde liegenden Kriterien offenzulegen, nach denen eine bestimmte Entscheidung getroffen wurde. Zum Beispiel ist auf die Frage „Warum wurde mir der Kredit verweigert?“ die Antwort „Der Computer hat so entschieden“ keinesfalls ausreichend.

Ein KI-System kann seine internen Entscheidungsgrundsätze aber nicht selbst erklären. Dazu müsste es die zugrunde liegende Problemstellung – den Kontext – *verstehen*. Genau das ist der Knackpunkt: Keine Methode der KI beherrscht heute das kontextuale Verstehen. Das würde nämlich nicht nur bedeuten, formal Beziehungen (f: X→Y) zu erkennen, sondern auch kausal die Zusammenhänge (wie die Ursache für X→Y) und daraus Schlüsse zu ziehen.

Das Fernziel der interpretierbaren KI sind Verfahren, die eine Verknüpfung zwischen statistischen Lernmethoden und großen Wissensrepräsentationen herstellen und deren KI-Algorithmen schließlich nachvollziehbar, verständlich und erklärbar ablaufen.

Von Natur aus transparent

Auch wenn der Ansatz der Nachvollziehbarkeit angesichts neuronaler Netze neuartig erscheint, so gibt es bereits eine lange Tradition transparenter Glass-Box-Systeme, die im Gegensatz zu Black Boxes einen Einblick in ihre innere Funktionsweise erlauben.

Seit Beginn der KI werden Expertensysteme erforscht, die linearen Modellen entsprechen. Diese folgen einfachen linearen Funktionen und können somit im Gegensatz zu den nicht-linearen neuronalen Netzen leicht durchschaut werden. Weitere transparente Expertensysteme sind beispielsweise in Entscheidungsbäumen organisiert und häufig sogar als sogenannte Random Forests, also als „Wälder“ aus mehreren unabhängigen Entscheidungsbäumen. Diese sind während des Lernprozesses parallel gewachsen. Jeder einzelne Baum aus diesem Wald trägt zur Entscheidungsfindung mit einer bestimmten Gewichtung bei. Das hat den großen Vorteil, dass nachvollzogen werden kann, welcher Baum zu welchem Ergebnis tendiert.

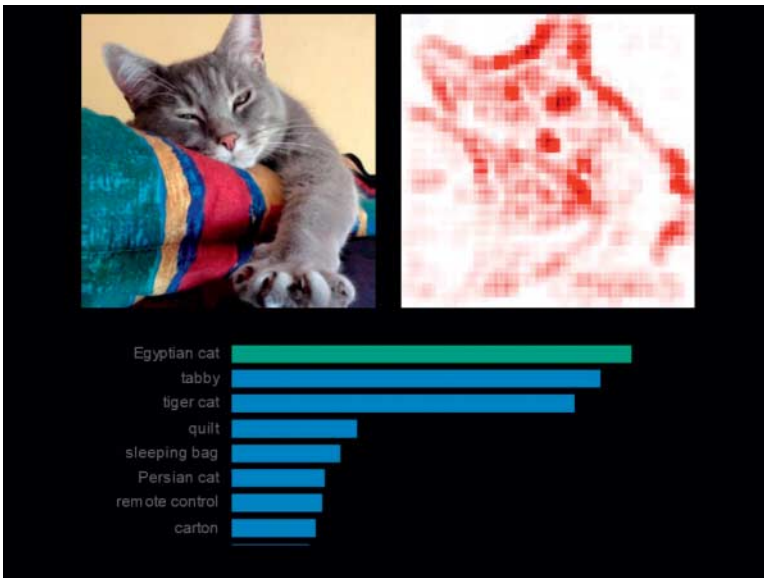
Ein weiteres typisches Beispiel für Glass-Box-Ansätze ist das interactive Machine Learning (iML) [1]. Hier wird dem Menschen eine Chance gegeben, in einen Algorithmus direkt einzugreifen, zum Beispiel in Form eines einfachen Spiels, in dem jeder Teilnehmer ein messbar gutes oder schlechtes Lösungsverhalten vorgibt (siehe Kasten „Gamification optimiert Traveling Salesman“). Damit gebildete Präferenzen für Teillösungen erlauben es, den Lösungsalgorithmus schrittweise nachzuvollziehen.

Erklärung für einen konkreten Fall

Darüber hinaus kommen bei komplexen neuronalen Netzen die sogenannten Post-Hoc-Erklärungsmethoden zum Tragen. Post-Hoc, lateinisch für „nach diesem (Ereignis)“, meint Methoden, die nicht das ganze Modell, sondern jeweils eine erhaltene Lösung erklären. Es wird also ein Datensatz klassifiziert und anschließend dafür eine nachvollziehbare Begründung geliefert.

Die drei derzeit bekanntesten Post-Hoc-Methoden heißen BETA, LRP und LIME. Bei allen handelt es sich um Prototypen, die frei verfügbar auf der Entwicklerplattform GitHub zur Verfügung stehen. Konkrete Anwender-Tools gibt es aber noch nicht.

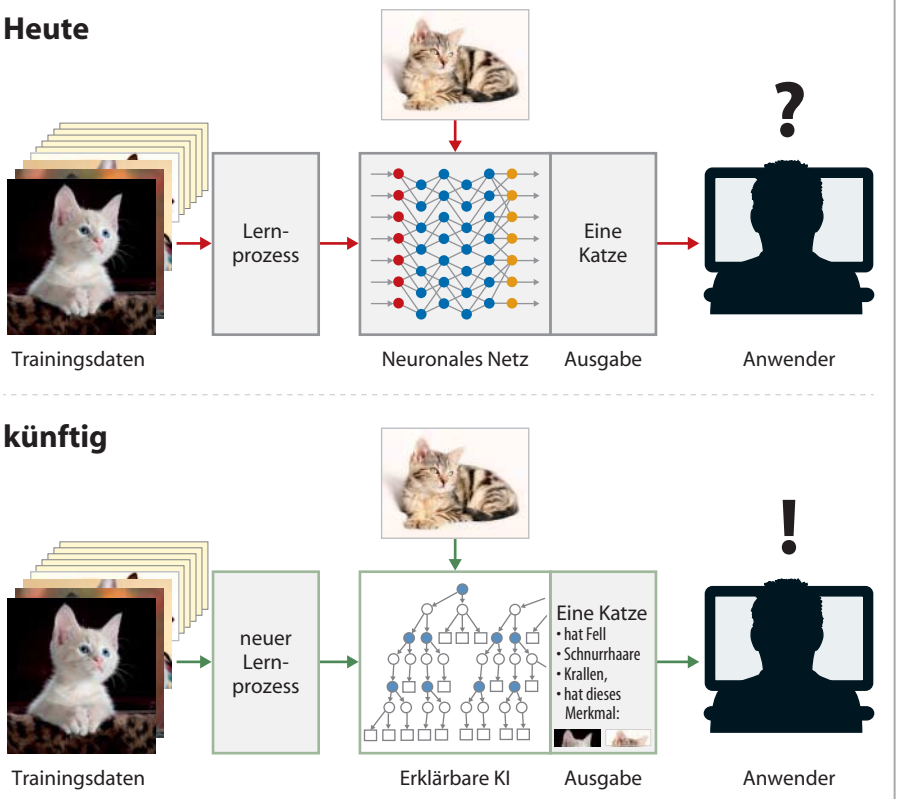
Bei BETA (Black Box Explanations through Transparent Approximations, [2]) werden für den Menschen nachvollziehbare Erklärungen für das Verhalten eines Klassifikators gesucht, beispielsweise einer KI, die Tumore in Aufnahmen als gutartig oder bösartig einschätzt. BETA-Werkzeuge unterstützen in diesem Fall den Arzt dabei, sinnvolle Teilbereiche (sogenannte „Subspaces“) zu identifizieren. Subspaces sind durch gemeinsame Eingangsmerkmale gekennzeichnet und für diese Bereiche folgt der Klassifikator einem überschaubaren Regelsatz. So klären zwei Subspaces beispielsweise, wie sich das Modell für Patienten über 50 Jahre im Vergleich zu Patienten unter 30 Jahren verhält. Anhand der Subspaces wird die Komplexität der Klassifikationsentscheidungen besser durchschaubar. Die Bewertung des Verhaltens in den Subspaces erfordert allerdings Sachkenntnis (Kontextverstehen), weshalb bei dieser Methode ein Fachexperte wie im Beispielfall der Arzt eingebunden werden muss – Zusammenhänge verstehen können heute auch die besten KI-Systeme nicht. Das heißt, mit BETA-Unterstützung können Experten interaktiv untersuchen, wie sich



Im LRP-Verfahren (Layer-Wise Relevance Propagation) wird ermittelt, welcher Input welchen Anteil am Ergebnis hatte; eine Heatmap (rechts) markiert die Bereiche mit dem größten Gewicht. Dabei offenbart sich hier: Es sind fälschlicherweise auch Teile des Bettbezugs in die Entscheidung eingeflossen.

Entscheidung wird transparent

Heutige neuronale Netze liefern ein Klassifikationsergebnis, bilden aber kein Vertrauen beim Anwender. Interpretierbare KI könnte in Zukunft Entscheidungen fällen und gleichzeitig offenbaren, worauf sich diese Entscheidungen gründen. Der Anwender kann damit das Klassifikationsergebnis besser einschätzen und gegebenenfalls begründetes Vertrauen in das KI-System setzen.



das Black-Box-Modell in interessanten Teilbereichen verhält.

Menschliche Experten sind oft in der Lage, aufgrund ihres fachlichen Verständnisses interessante Bereiche nicht nur zu erkunden, sondern gegebenenfalls auch zu korrigieren. Daher kann bei der Suche nach Erklärungen zugleich die Genauigkeit des ursprünglichen Modells verbessert werden.

Rückwärts durchs neuronale Netz

Das LRP-Verfahren (Layer-Wise Relevance Propagation, [3]) stellt eine weitere allgemeine Lösung zum Verstehen von Klassifikationsentscheidungen dar. Stark vereinfacht erlaubt LRP, die Entscheidungsprozesse in neuronalen Netzen rückwärts ablaufen zu lassen. Dabei wird nachvollziehbar, welcher Input welchen Einfluss auf das jeweilige Ergebnis hatte. Ist beispielsweise ein neuronales Netz mit genetischen Daten und damit zusammenhängenden Erkrankungen trainiert, so kann im speziellen Fall mit LRP nicht nur analysiert werden, mit welcher Wahrscheinlichkeit ein Patient eine bestimmte genetische Erkrankung hat, sondern auch anhand welcher Merkmale in den Eingabedaten diese Entscheidung getroffen wurde. In Zukunft könnte mit diesem Wissen eine genau auf den individuellen Patienten abgestimmte Krebstherapie ermittelt werden.

Interessant an diesem Ansatz ist, dass der Einfluss jedes Merkmals auf das Ergebnis zum Beispiel durch Heatmaps – ähnlich wie im Bild einer Wärmebildkamera – visualisiert werden kann. Stärker eingefärbte rote Bereiche tragen beispielsweise mehr zur Gesamtentscheidung bei als schwächer eingefärbte Bereiche. So werden die Ergebnisse für menschliche Experten nachvollziehbar, was zum Beispiel bei der Bestimmung von Erkrankungswahrscheinlichkeiten in klinischen Studien hilfreich ist oder zur Identifikation von Risikofaktoren bei Kreditinstituten oder – um bei einem bekannten Beispiel zu bleiben – welcher Bereich am relevantesten beiträgt, damit eine Katze als Katze erkannt wird. Am Beispiel der Klassifikation von Fotos, Texten oder handschriftlichen Ziffern kann diese Methode auf dem Server des Fraunhofer HHI ausprobiert werden, samt Anzeige erhellender Heatmaps (siehe ct.de/y45s).

Was gab den Ausschlag?

Die LIME-Methode (Local Interpretable Model-agnostic Explanations, [4]) ist darauf ausgerichtet, den Einfluss von Teilen der Eingangsparameter, den sogenannten Instanzen, zu klären. Eine Instanz kann etwa eingegrenzte Bildbereiche oder Bildmuster umfassen oder einen bestimmten Abschnitt von Patientendaten. Es wird nun versucht, Instanzen zu finden, deren Einfluss auf das Klassifikationsergebnis durch nachvollziehbar einfache Funktionen beschrieben werden kann.

Mit jeder Instanz arbeitet das Verfahren dann separat weiter. Ihre Eingangsparameter werden permutiert und ein Ähnlichkeitsmaß zur ursprünglichen Instanz berechnet. Nun lässt man das zu erklärende Modell Vorhersagen für jede dieser Permutationen treffen, der Einfluss der Änderungen auf die Vorhersagen kann so nachvollzogen werden. Schritt für Schritt wird auf diesem Weg aus dem Eingaberaum eine bedeutende Instanz herausgeschält, deren Einfluss sich mit einer überschaubaren Funktion wiedergeben lässt.

Schließlich kann LIME den Wert bestimmen, den jede Instanz zur Gesamtentscheidung beigetragen hat, und stellt diesen grafisch als Balken mit Prozentwerten dar. So können beispielsweise Ärzte überprüfen, welche Einzelkomponenten in welchem Ausmaß zu einer Entscheidung beigetragen haben, beispielsweise dass ein betrachteter Tumor gutartig ist. Experten haben damit eine Chance zu überprüfen, ob die Ergebnisse plausibel sind. Es wird allerdings lediglich dargestellt, welche Instanz mit welcher Ausprägung zum Ergebnis beigetragen hat. Es kann nicht erklärt werden, warum das so ist – dies bleibt vorerst dem Menschen vorbehalten.

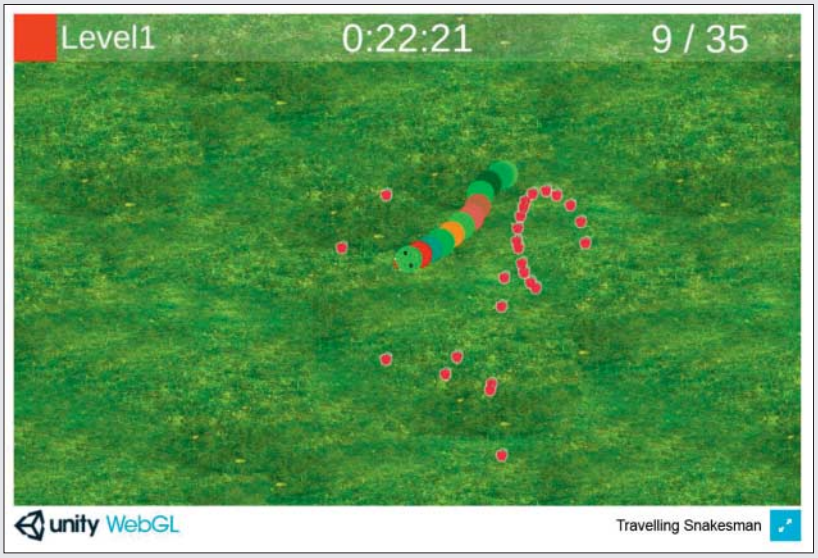
In einem anderen Beispiel bei der Unterscheidung von Katzen- und Hundebildern würde LIME ergründen, welche Bildausschnitte die Entscheidung für die Klassifikation besonders beeinflusst haben, etwa die Ohren der abgebildeten Tiere. Auf

Gamification optimiert Traveling Salesman

Wir demonstrieren an einem Beipiel, wie menschliche Intelligenz in einen interaktiven maschinellen Lernansatz (iML) für KI eingebracht werden kann [1]. Dazu wurde als Fallstudie das Traveling-Salesman-Problem gewählt, auch Rundreiseproblem genannt. Dabei handelt es sich um ein Optimierungsproblem, das in vielen Anwendungen des täglichen Lebens auftritt. Das Problem ist aber auch NP-vollständig, das heißt, es lässt sich nicht effizient allgemeingültig lösen. Für die Fallstudie wurde als Spiel „Traveling Snakesman“ entwickelt.

Traveling Snakesman basiert algorithmisch auf einem sogenannten Ant Colony Optimization Framework (ACO). Das zugrunde liegende Modell ist Ameisen auf der Futtersuche nachgebildet, die mit Duftstoffen (Pheromone) kommunizieren, Wege markieren und damit ihr Suchverhalten optimieren. Diese „Ameisenalgorithmen“ werden vielseitig eingesetzt, von der Routenoptimierung im Navi bis zur Proteinfaltung in der Medizin. Ein ACO-Framework besteht aus autonomen Einheiten (Software-Agenten), die kollektiv zur Problemlösung eingesetzt werden.

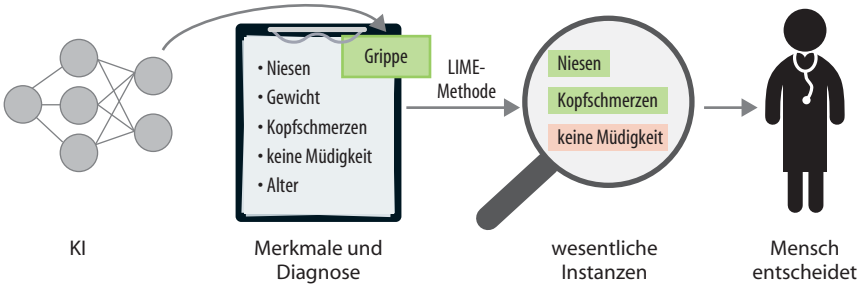
Dieses Modell haben wir so modifiziert, dass Menschen die Rolle eines Agenten einnehmen können, um Pheromonwerte entlang einer Wegstrecke direkt zu beeinflussen. Damit das Ganze für Spieler attraktiv wird, steuert der Mensch zwar keine Ameise, aber eine Schlange über ein Feld, auf dem Äpfel verteilt liegen. Das Spielziel ist es, so schnell wie möglich alle Äpfel abzugrasen. Die Ergebnisse zeigen, dass der Mensch, der eine Tour über den Spielplan führt, einen signifikanten Einfluss auf den Algorithmus und den von ihm gefundenen kürzesten Weg hat. Der Versuch zeigt also ein Beispiel, wie menschliche Intelligenz die maschinelle Intelligenz positiv beeinflusst. Das Experiment ist noch nicht abgeschlossen, Mitspielen ist jederzeit möglich: https://hci-kdd.org/project/iml/.



Gamification bedeutet, dass der Mensch spielerisch die Rolle eines lösungssuchenden Software-Agenten einnimmt und mit seiner Intuition die Gesamtlösung beeinflusst, beispielsweise bei einer Abwandlung des Rundreiseproblems, in dem er alle Äpfel auf der Wiese einsammelt.

KI und Experte gemeinsam

In Hybridsystemen legt eine KI offen, welche Eingangsdaten vor allem zu einer Klassifikationsentscheidung geführt haben. Damit wird die Entscheidungsgrundlage der KI transparent, der menschliche Fachexperte, in diesem Fall ein Arzt, kann so deren Output besser berücksichtigen und für den Einzelfall eine individuelle Entscheidung fällen.



diese Weise ließe sich nachvollziehen, nach welchen Kriterien das KI-System Entscheidungen trifft. Falsch trainierte Systeme lassen sich mit der LIME-Methode gut entlarven. Falls beispielsweise alle Katzenfotos des Trainingsmaterials in geschlossenen Räumen aufgenommen wurden und die Hundebilder im Freien, könnte ein KI-System das Bild einer Katze im Garten als Hundefoto klassifizieren. Solchen Fehlern kann man mit LIME auf die Schliche kommen.

Mensch und KI ergänzen sich

Allerdings, und das gilt leider für alle State-of-the-Art-Methoden, kann derzeit nicht erklärt werden, warum eine Entscheidung getroffen wurde. Dieses Verstehen und Schlussfolgern von Zusammenhängen und auch das Generalisieren aus wenigen Beispielen, genau das können allerdings Menschen sehr gut. Daher werden in Zukunft beide zusammen wichtig sein, KI und der menschliche Fachexperte. Ein Mensch kann helfen, wo die KI an ihre Grenzen kommt, und die KI kann unterstützen, wo Menschen an ihre Grenzen kommen. Ärzte können von monotonen Routineaufgaben entlastet werden, während gleichzeitig, wie Studien belegen, KI-Systeme und menschliche Experten gemeinsam bessere Entscheidungen treffen als jeweils für sich allein [5].

Derart hybride Systeme werden oft in der Medizin verwendet, insbesondere dann, wenn man nicht nur mit Bilddaten, Analysewerten und Ähnlichem zu tun hat, sondern auch mit komplexen Textmengen. Die Medizin gilt als ein Prototyp für nicht-monotones Schließen, wo man unter großer Unsicherheit schlussfolgern und Entscheidungen treffen muss. Zudem ist dieses Aufgabenfeld durch unvollständige Informationen gekennzeichnet. Menschen zeigen gerade in niedrigdimensionalen Problemstellungen unserer Alltagsumgebung sehr gute Intuition, können durch ihre Alltagsintelligenz erstaunlich gut aus wenigen Daten generalisieren und Zusammenhänge erkennen – die Algorithmen können das bis dato nicht.

Die große Chance interpretierbarer KI ist nicht nur, Black Boxes transparent zu machen und damit Vertrauen in KI zu fördern, sondern vor allem ein tieferes Verständnis für vorher unbekannte Zusammenhänge zu fördern. So könnten Ärzte beispielsweise Algorithmen auf interessant erscheinende Daten ansetzen und interaktiv Zusammenhänge ergründen.

Servolenkung fürs Gehirn

Vielleicht der wichtigste Beitrag von erklärbarer KI ist es, aufzuklären, was Ursache ist und was Wirkung. Die größte Gefahr besteht nämlich darin, dass man fälschlich Artefakte in Entscheidungen einbezieht, das sind scheinbare, irreführende Zusammenhänge oder schlichtweg falsche Ergebnisse. Die Unterscheidung zwischen Ursache und Wirkung ist in vielen Anwendungsdomänen wünschenswert, in sicherheitskritischen Bereichen jedoch zwingend erforderlich.

Die große Chance für die Zukunft besteht aus einer Verknüpfung verschiedener bereits bewährter Ansätze, zum Beispiel logikbasierte Ontologien (Wissensrepräsentationen mit formallogischen Bedingungen) mit maschinellem Lernen und mit einem menschlichen Experten zu

einem hybriden Interaktionsmodell zu fusionieren. KI würde dann als eine Art „Servolenkung fürs Gehirn“ unterstützend verwendet werden. Dies würde nicht nur eine Erweiterung menschlicher Intelligenz mit maschineller Intelligenz bedeuten, sondern auch umgekehrt eine Erweiterung der künstlichen Intelligenz durch menschliche Intuition. *(agr@ct.de)* c't

Literatur

[1] Andreas Holzinger, Interactive Machine Learning for Health Informatics, When do we need the human-in-the-loop?: https://braininformatics.springeropen.com/track/pdf/10.1007/s40708-016-0042-6

[2] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, Jure Leskovec, Interpretable and Explorable Approximations of Black Box Models: https://arxiv.org/abs/1707.01154

[3] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, Wojciech Samek, The LRP toolbox for artificial neural networks: www.jmlr.org/papers/volume17/15-618/15-618.pdf

[4] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, Why should I trust you? Explaining the predictions of any classifier: www.jmlr.org/papers/volume17/15-618/15-618.pdf

[5] Arne Grävemeyer, KI erkennt Krebs, Neuronale Netze in der Radiologie, c't 14/2018, S. 52

[6] Sebastian Stabinger, Putin – KGB + NSA = Obama, Word2Vec berechnet Bedeutung, c't 15/2018, S. 182

Motivklassifikation mit Analyse:
ct.de/y45s

Verständlich – unverständlich

Verstehen und Erklären sind Voraussetzungen für Nachvollziehbarkeit. Aber was ist für den Menschen überhaupt verstehbar? Direkt verständlich und damit auch interpretierbar und nachvollziehbar sind Daten und Objekte in der Ebene, maximal im Raum, zum Beispiel Bilder (Matrix aus Pixeln, Graphen, 2D/3D-Projektionen) oder Text. Menschen können Bilder und Texte mit Bezug auf ihr Vorwissen verstehen.

Vorhersagemodelle, die für Menschen interpretierbar sind, bestehen beispielsweise aus Entscheidungslisten mit einer Reihe von Wenn-dann-Aussagen (etwa: wenn hoher Blutdruck, dann droht Schlaganfall). Auf diese Weise kann ein hochdimensionaler, viele Variablen betreffender Merkmalsraum oftmals in einen niedrigdimensionalen und somit menschlich interpretierbaren Entscheidungsraum übertragen werden.

Nicht direkt verständlich und damit auch nicht nachvollziehbar sind höherdimensionale Vektorräume. Ein Beispiel sind die sogenannten Word Embeddings, das heißt jedem Wort wird ein Vektor in einem n-dimensionalen Vektorraum zugeordnet. Nun kann Vektoralgebra betrieben werden, zum Beispiel „Sohn“ – „Mann“ + „Frau“ = „Tochter“ [6]. Angesichts einer schnell steigenden Dimensionenzahl streikt hier aber bald die menschliche Vorstellungskraft.

Letztlich nicht verstehbar sind auch undokumentierte, das heißt noch unbekannte Eingangsmerkmale, wie zum Beispiel Textsequenzen mit unbekannten Wörtern oder unbekannten Symbolen (Chinesisch etwa, wenn man kein Chinesisch versteht).

Algebra im Vektorraum

Das Word Embedding ordnet jedem Wort einen Vektor in einem n-dimensionalen Vektorraum zu. Das sieht im gezeigten Beispiel recht durchschaubar aus, aber mit steigender Dimensionenzahl sind Vektorräume für Menschen kaum nachzuvollziehen, auch grafische Abbildungen mit mehr als drei Dimensionen sind sehr schwierig herstellbar.