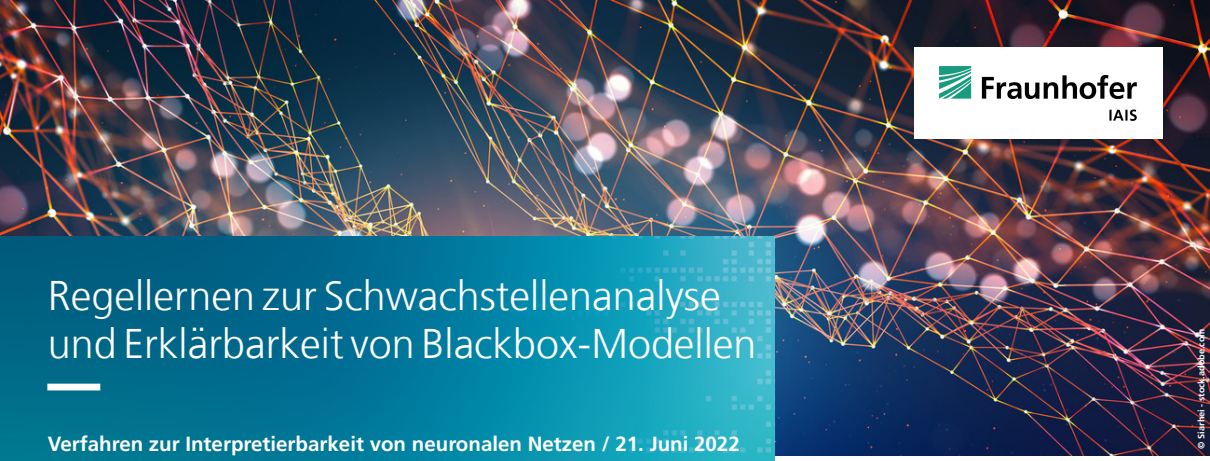


Regellernen zur Schwachstellenanalyse und Erklärbarkeit von Blackbox-Modellen

Verfahren zur Interpretierbarkeit von neuronalen Netzen / 21. Juni 2022

Dr. Daniel Becker

Transparenz, Interpretierbarkeit und Erklärbarkeit

Transparenz als Dimension von Vertrauenswürdigkeit

Fairness

Vermeidung unfairer Voreingenommenheit

Autonomie & Kontrolle

Angemessene Aufgabenverteilung zwischen Mensch und KI-Anwendung

Transparenz

Nachvollziehbarkeit, Reproduzierbarkeit, Erklärbarkeit, Auditfähigkeit

Verlässlichkeit

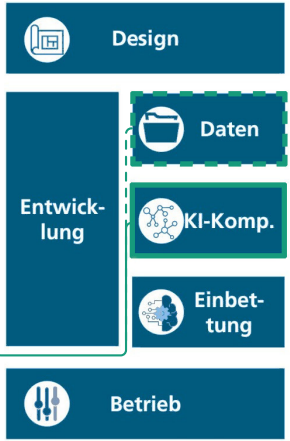
Verlässlichkeit im Regelfall, Robustheit, Unsicherheitsabschätzung

Sicherheit

Funktionale Sicherheit, Integrität und Verfügbarkeit

Datenschutz

Schutz von personenbezogenen Daten und geschäftsrelevanter Information



Icon Quellen: https://iconmonstr.com/construction-35-png; https://iconmonstr.com/folder-20-png/

Fraunhofer IAIS

Transparenz, Interpretierbarkeit und Erklärbarkeit

Wenn Genauigkeit, Performanz und Zuverlässigkeit nicht ausreichen

Übergeordnetes Ziel

Gerechtfertigtes Vertrauen in automatisierte Entscheidungen von **KI-Lösungen** schaffen

Zu welchem Zweck?

- Entscheidung rechtfertigen
- **Fehler verstehen und korrigieren**
- **KI-Modell verbessern**
- **Erkenntnisgewinn**
  - ↳ neue Zusammenhänge in Daten
  - ↳ besseres Problemverständnis

In welchem Kontext?

- Entwurf und **Entwicklung**
- Betrieb bzw. Nutzung
  - ↳ Hochrisiko-Anwendungen
  - ↳ Verkehrs- und Gesundheitswesen, Recht, Militär

Für wen?

- (Laien-) Benutzer
  - ↳ Endnutzer*innen und Besitzer*innen der KI-Lösung
  - ↳ betroffene Personen
  - ↳ Regierungen und Gesetzesvertreter
- **Technische und Domänenexperten**
  - ↳ Data Scientists und Softwareentwickler*innen
  - ↳ Auditor*innen

Transparenz, Interpretierbarkeit und Erklärbarkeit

Abwägungen

Interdimensionale Risiken mangelnder Transparenz

- Verletzung von Anforderungen im Zusammenhang mit menschlicher **Autonomie und Kontrolle**
- Un**fairness** gegenüber betroffenen Personen
- Mangelnde **Zuverlässigkeit**

Transparenz nicht immer erforderlich

Systeme und Anwendungsfälle mit

- Niedrigen Risiken
  - ↳ z.B. automatisches Auspielen von Werbung
- Tauglichen Strategien zur Fehlererkennung und -behandlung
  - ↳ z.B. Zugangskontrolle mit Unsicherheitsabschätzung

Mögliche Zielkonflikte

- Weniger effiziente oder effektive Systeme
- Eingeschränkte Designmöglichkeiten
- Tendenz zu weniger tauglichem und vielfältigem Systemverhalten
- Einschränkungen bei Datenschutz und Sicherheit
  - ↳ z.B. Aufdeckung heuristischer Regeln zur Betrugserkennung
  - ↳ z.B Extraktion von Geschäftsgeheimnissen

Transparenz, Interpretierbarkeit und Erklärbarkeit

Interpretierbarkeit versus Erklärbarkeit von Modell und Daten

Interpretierbarkeit

- Mensch kann Modell und Daten **als Ganzes durchdenken**

oder

- **Jede Modellkomponente** (Eingabe, Parameter und Berechnung) erlaubt intuitives Verständnis

oder

- Modell verhält sich **bewiesen korrekt** für alle (neuen) Eingaben

Ableitung von Erklärung für beliebige Modellvorhersage möglich

Erklärbarkeit

- Gegeben **auf Ebene einzelner Ein-/Ausgabe Instanzen** (Vorhersage, Entscheidung)
- Gleichbedeutend mit **lokaler Interpretierbarkeit**

Erklärung

Menge von Merkmalen aus der **interpretierbaren Datendomäne**, die bei gegebener Eingabe zur Ausgabe **beitragen**

Weil $A(x_i) = a$ & $B(x_i) = b$, ist $y_i = o$

Verständlichkeit nur subjektiv bewertbar

Schwachstellenanalyse für Blackbox-Modelle

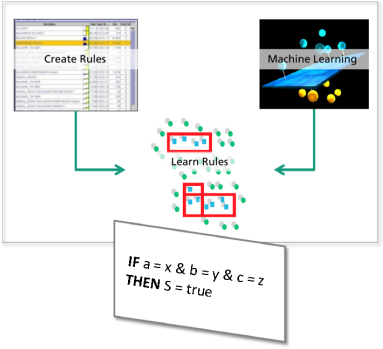
Automatisierte Regelsuche nach Fehlerbedingungen

Herausforderungen

- Blackbox-Modelle komplex und Leistungsfähig, aber **Vorhersagen schwer bis gar nicht verständlich**
- **Riesiger Suchraum** für potentielle Fehlerzustände
- Grund für **Versagen** im Einzelfall schwer **erklär- oder generalisierbar**

Lösungsansatz

- Subgruppensuche findet Regeln für Fehlerfälle mit maximaler „Interessantheit" (Größe und Genauigkeit) bzgl. Suchkriterium
- Voraussetzung: aussagekräftige **Metadaten vorliegend oder generierbar**



Fraunhofer IAIS

Schwachstellenanalyse für Blackbox-Modelle

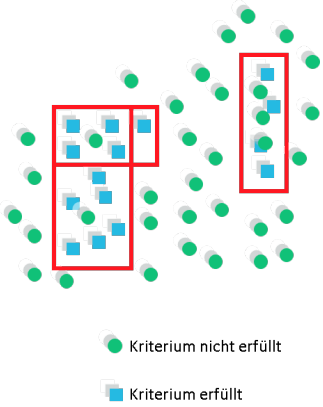
Subgruppensuche mit dem Fraunhofer RuleCreator

Idee

- Finden von interessanten Subgruppen (Teilmengen)
  - Definiert durch **Konjunktionen** von Attributs-Bedingungen (z.B. „Akzent = Friesisch & Geschlecht = männlich“)
  - Möglichst **groß** und mit möglichst stark gegenüber Mittel **erhöhtem Anteil** von Elementen, die Suchkriterium erfüllen
- Attribute: **kategorisch** (oder geeignet diskretisiert)

Features

- **Transparent:** interpretierbare (verifizier- und erweiterbare) Regeln
- **Integrierbar** in menschengemachte Regelsysteme
- **Hohe Suchraumabdeckung:** finden aller statistisch relevanten Regeln
- **Effizient:** parallele, in-memory Suche implementiert



Fraunhofer IAIS

Schwachstellenanalyse für Blackbox-Modelle

Beispiel: Semantische Segmentierung

Anwendungsfall

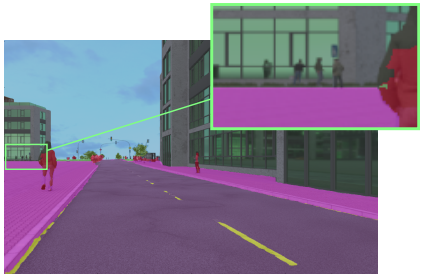
- Neuronales Netz klassifiziert Pixel eines Autokamerabildes (Kategorien: Fußgänger*in, Fahrbahn, Fahrzeug, …)

Metadaten

- Verfügbar z.B. als Labels des Trainingsdatensatzes (hier: Umrahmungen von Fußgänger*innen)

Beispielregeln

- Fußgänger*in nicht erkannt, wenn
  - Umrahmungsgröße < Schwellwert (height < 8 & width < 12)
  - Vertikale Position der Umrahmung > Bildmitte (ycoord > 644)



Fußgänger*innen-Umrahmungen mit Erkennungsgüte (IoU)

id	xcoord	ycoord	width	height	size	IoU	…
0	1071	535	21	59	1239	0.92	
1	94	475	16	34	544	0.00	
2	1520	535	37	94	3478	0.70	
3	228	484	18	34	612	0.00	
4	1068	527	17	48	816	0.97	
5	1039	544	50	126	6300	0.71	
…							

Schwachstellenanalyse für Blackbox-Modelle

Statistische und Visuelle Regelanalyse

Beispielregel

WENN Umrahmung in oberer Bildhälfte DANN IoU $= 0$

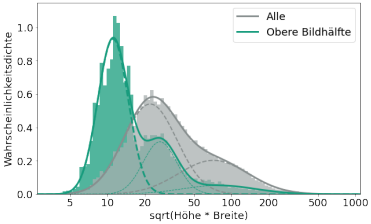
Hypothesen

1. Horizont im Datensatz relativ häufig in Bildmitte (ycoord = 640)
2. Umrahmung in oberer Bildhälfte relativ häufig
   nah am Horizont ➡ relativ klein ➡ schlecht erkannt

Analyse

- **66%** aller Umrahmungen **unterdurchschnittlich klein** (zwischen 25 und 400 Pixel)
- **70%** dieser kleinen Umrahmungen mit **IoU** $= 0$
- Bestätigt durch **visuelle Prüfung/Validierung** (z.B. mit **ScrutinAI**)

Größenverteilung der Fußgänger*innen-Umrahmungen



Fraunhofer IAIS

Schwachstellenanalyse für Blackbox-Modelle

Interpretierbarkeit und Erkenntnisgewinn

Größenregel

WENN Umrahmungsgröße kleiner als Schwelle DANN IoU $= 0$

- Einfach interpretierbar auch für Nicht-Experten
- Erkenntnis:
  - ↳ Erkennung versagt bei sehr kleinen Umrahmungen (**Modell**)

Horizontregel

WENN Umrahmung in oberer Bildhälfte DANN IoU $= 0$

- Nicht ad-hoc interpretierbar
- Hypothesenbildung und „Expertenanalyse" nötig
- Erkenntnisse:
  - ↳ Erkennung versagt bei sehr kleinen Umrahmungen (**Modell**)
  - ↳ Relativ viele kleine Umrahmungen am Horizont (**Daten**)
  - ↳ Potentiell Zusammenhänge über **Umwelt** ableitbar

Globale Blackbox-Erklärungen

- **Regel:** statistisch relevante Gruppe gleich **erklärbarer Fehl-Erkennungen**
- **Erschöpfend** bei gegebener Merkmalsmenge
- **Regelgüte ablesbar** an statistischen Konfidenzmaßen
- **Modell-agnostisch**
- **Datentyp-agnostisch**

Schwachstellenanalyse für Blackbox-Modelle

Beispiel: Klassifizierung von Elementen auf Webseiten

Anwendungsfall

- Modell erkennt bestimmte Webseitenelemente (Such-Button, Eingabemasken, …) anhand von Screenshots

Metadaten

- Webseiten: strukturierte Dokumente in maschinenlesbarer Sprache (HTML, CSS, …)
- Automatisiert extrahier- und generierbar mit etablierten Werkzeugen (z.B. reguläre Ausdrücke, Bildverarbeitung, …)

Beispielregeln

- Such-Button nicht erkannt, wenn
  Kontrast < Schwellwert & Umriss = Rechteck & Symbol = Auge

Fraunhofer IAIS

Schwachstellenanalyse für Blackbox-Modelle

Beispiel: Inhaltsanalyse von Texten in Dokumentdatenspeicher

Anwendungsfall

- Modell klassifiziert Texte in Artikeldatenbank

Metadaten

- Üblicherweise in Datenspeicher integriert
- Erweiterbar (manuell oder maschinell generierte Attribute)

Beispielregeln

- Artikel nicht als Satire erkannt, wenn
  - Ressort = Politik & Region = Südostasien
  - Länge < 1000 Zeichen & Anteil IT-Fachbegriffe > 15%


```
{
  "Documents": [
    {
      "Title": "De Finibus Bonorum et Malorum",
      "Type": "Article",
      "Author": "Cicero",
      "Date": "45 BC",
      "Keywords": [ "fake", "Latin", "filler", "text" ],
      ...
      "Content": "Lorem ipsum dolor sit amet, consectetur ad
    },
    {
      ...
```

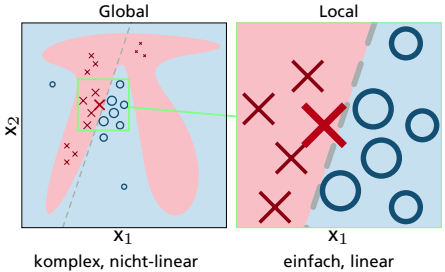

Fraunhofer IAIS

Regelbasierte Erklärungen für KI-Modelvorhersagen

Local Interpretable Model-agnostic Explanations (LIME)

- **Idee:** Approximation von Blackbox-Modell mit interpretierbarem Ersatzmodell (Surrogat-Modell)
  (siehe M. T. Ribeiro, S. Singh, and C. Guestrin. In SIGKDD, 2016)
- Erklärung durch **lokale Approximation:**
  1. Wahl der Eingabeinstanz für Erklärung
  2. Generation lokaler Störungen um Instanz
     (Art abhängig von Modellierungsaufgabe)
  3. Training eines intrinsisch interpretierbaren (lokalen) Modells auf Störungen und den Vorhersagen des (globalen) Blackbox-Modells
  4. Interpretation des lokalen Modells als Erklärung der Blackbox-Entscheidung

lokales Modell: **penalized linear regression**



Globales Modell: binäre Klassifikation (rote Kreuze, blaue Kreise) auf Basis der Merkmale x_1 und x_2

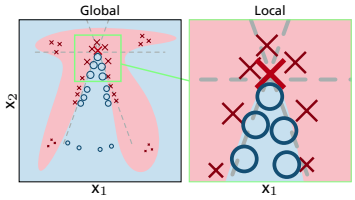
Lokales Modell: x_1 ist entscheidende(re)s Merkmal in Umgebung der untersuchten Instanz

Regelbasierte Erklärungen für KI-Modelvorhersagen

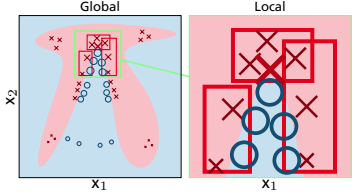
Finden lokaler, interpretierbarer Regeln per Subgruppensuche

- **Idee:** Ersatz der linearen Regression in LIME durch Subgruppensuche (RuleCreator)
- **Features:**
  - Weitgehend **Modell-** und **Datentyp-agnostisch** (wie LIME)
  - Höhere **Stabilität** der Erklärungen
  - **Interpretierbar** von Menschen
  - **Finden aller** statistisch relevanten **Erklärungen**
  - **Konfidenzmaße** „mitgeliefert" (Genauigkeit, Trefferquote, …)

Je nach Instanz, **Regressionslösung instabil**



Subgruppensuche findet alle lokalen Erklärungen



Regelbasierte Erklärungen für KI-Modelvorhersagen

Proof of Concept: binäre Textklassifikation von Newsgroup Emails

- **Trainingsdaten:** Teilmenge aus **20 newsgroups** mit den Labels „christlich" und „atheistisch"
  (siehe http://qwone.com/~jason/20Newsgroups/)
- **Blackbox:** Random-Forest Klassifizierer mit 500 Schätzern (über 90% Genauigkeit nach Training)
- **Erklärung** der Testinstanz: Blackbox-Klassifikation basiert auf **Emailheader**
  (siehe https://github.com/marcotcr/lime)

Testinstanz

Vorhersage: $P(\text{atheistisch}) = 0.56$

Korrektes Label = „atheistisch"

```
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to
…
```

LIME (wichtigste 3 Wörter)

Wort	Regressiongew.
`Host`	0.158
`Posting`	0.148
`NNTP`	0.097

Subgruppensuche (wichtigste 3 Regeln)

WENN [...] **DANN** Label = „atheistisch"

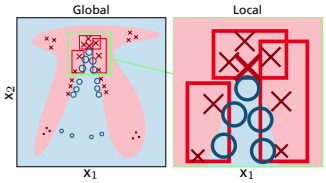
1. „`Host` $= 1$" Genauig.: 100%, Trefferq.: ~33%
2. „`Posting` $= 1$" Genauig.: 100%, Trefferq.: ~33%
3. „`NNTP` $= 1$" Genauig.: ~99%, Trefferq.: ~33%

Zusammenfassung

Regelsuche zum Finden globaler und lokaler Erklärungen für Blackbox-Modelle

Globale Erklärungen

- Geeignet zur globalen (Schwachstellen-)Suche und Analyse
- Basiert auf vorhandenen oder generierbaren Metadaten
- Model- und Datentyp-agnostisch



Lokale Erklärungen

- Alternative zu bzw. Ergänzung von „Standard"-LIME
- Mehrere lokale Erklärungen pro Instanz möglich
- Potenziell höhere Stabilität als LIME
- Einsetzbar bei multi-modalen Daten (z.B. Bilder und Sprache)