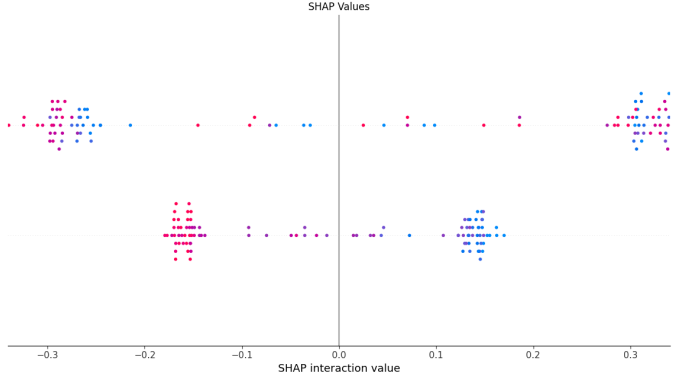


Explainable Artificial Intelligence

Felix Neubürger

2025

Fachhochschule Südwestfalen, Ingenieurs- & Wirtschaftswissenschaften

Abfrage Erwartungen und Vorwissen

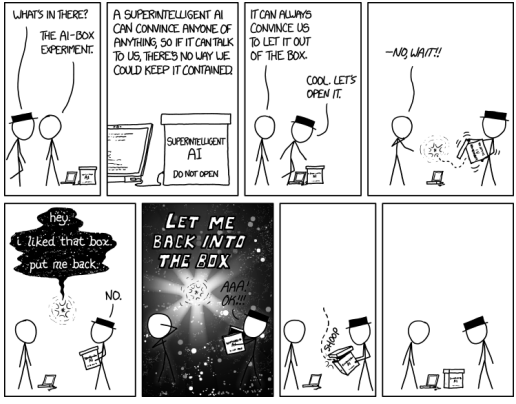
- Was verstehen Sie unter Explainable AI?
- Haben Sie bereits Erfahrungen mit maschinellem Lernen oder KI?
- Welche Erwartungen haben Sie an diese Vorlesung?
- Welche Anwendungsbereiche von KI interessieren Sie besonders?
- Gibt es spezifische Fragen, die Sie in dieser Vorlesung beantwortet haben möchten?

Inhalte der Vorlesung

- Begriffsklärungen
- Erkenntnistheoretischer Exkurs
- Methoden der Explainable AI
- Quantitative Methoden
- Anwendung der gelernten Methoden in einem Beispiel

Ziele der Vorlesung - Welche Fragen sollen beantwortet werden?

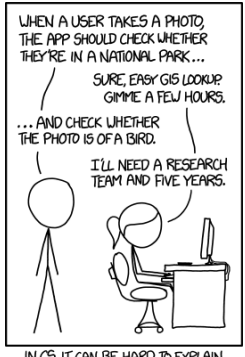
- Wofür Explainable AI?
- Was bedeutet Explainable AI?
- Interpretable AI?
- Trustworthy AI?
- Wie funktioniert das mathematisch?
- Wie schaffe ich Transparenz für Stakeholder?



[https://xkcd.com/1450/]

Format der Vorlesung - Wie sollen diese Fragen beantwortet werden?

- Theroretischer Teil mit Folien
- Selbststudium mt einem Lehrbuch[a]
- Praktischer Teil in Gruppen an einem Projekt
- Gruppengröße 2 oder 3 Personen
- Einzelarbeit möglich wenn eigenes Thema vorhanden
- Abgabe der Ausarbeitung einen Tag vor der Veranstaltung in der Blockwoche
- Vorstellung der Projektergebnisse in der Blockwoche
- Gewichtung der Bewertung Projektausarbeitung (50%) und Vortrag (50%)



[https://xkcd.com/1425/]

[a] https://christophm.github.io/interpretable-ml-book/

Künstliche Intelligenz (KI)

- Teilgebiet der Informatik
- Automatisierung intelligenten Verhaltens
- Maschinelles Lernen als Unterbereich

Maschinelles Lernen (ML)

- Unterbereich der KI
- Algorithmen lernen aus Daten
- Treffen von Vorhersagen oder Entscheidungen
- Deep Learning basiert auf künstlichen neuronalen Netzen

Explainable Artificial Intelligence (XAI)

- Ansätze zur Verständlichkeit von KI-Entscheidungen
- Wichtig für Vertrauen und Transparenz
- Beispiel: Erklärungen in der medizinischen Diagnostik

Definitionen und Unterschiede

Interpretierbarkeit

- Fähigkeit, die internen Mechanismen eines Modells zu verstehen.
- Ermöglicht direkte Einsicht in die Funktionsweise des Modells.
- Beispiel: Lineare Regression, Entscheidungsbäume.

Erklärbarkeit

- Fähigkeit, die Entscheidungen oder Vorhersagen eines Modells verständlich zu machen.
- Oft durch zusätzliche Methoden bei komplexen Modellen erreicht.
- Beispiel: Neuronale Netze mit Post-hoc-Erklärungen.

Erkenntnistheoretische Aspekte

- **Wissenserwerb:** Wie tragen Interpretierbarkeit und Erklärbarkeit zum Verständnis von KI-Entscheidungen bei?
- **Vertrauen:** Inwiefern beeinflusst die Nachvollziehbarkeit von Modellen das Vertrauen der Nutzer?
- **Transparenz vs. Komplexität:** Balance zwischen detaillierter Einsicht und praktischer Anwendbarkeit.
- **Ethische Verantwortung:** Bedeutung von Erklärbarkeit für ethisch vertretbare KI-Systeme.

Bedeutung der Interpretierbarkeit im Maschinellen Lernen

- **Definition:**
  Interpretierbarkeit bezeichnet das Maß, in dem ein Mensch die Ursache einer Entscheidung eines Modells nachvollziehen kann.

- **Warum ist Interpretierbarkeit wichtig?**
  - **Vertrauensbildung:**
    Nutzer vertrauen eher Modellen, deren Entscheidungswege sie verstehen.
  - **Fehleranalyse:**
    Verständnis für Modellentscheidungen erleichtert das Erkennen und Beheben von Fehlern.
  - **Einhaltung gesetzlicher Vorgaben:**
    In sensiblen Bereichen wie Medizin oder Finanzen sind nachvollziehbare Entscheidungen oft gesetzlich vorgeschrieben.

Herausforderungen und Begriffsabgrenzungen

- **Herausforderungen:**
  - Fehlende einheitliche Definition von Interpretierbarkeit erschwert Kommunikation und Forschung.
  - Kompromiss zwischen Modellkomplexität und Interpretierbarkeit oft notwendig.
- **Abgrenzung zu verwandten Begriffen:**
  - **Erklärbarkeit (Explainability):**
    Fähigkeit, interne Mechanismen eines Modells verständlich zu machen.
  - **Transparenz:**
    Ausmaß, in dem die Funktionsweise eines Modells offenliegt.
  - **Vertrauen:**
    Maß, in dem Nutzer darauf vertrauen, dass ein Modell korrekte und faire Entscheidungen trifft.

EU-Regulierung und Erklärbare Künstliche Intelligenz

Die Europäische Union hat den Artificial Intelligence Act verabschiedet,[1] der am 1. August 2024 in Kraft trat.[2]

EU AI Act: Überblick

- **Ziel:** Einführung eines risikobasierten Klassifizierungssystems für KI-Anwendungen.
- **Risikokategorien:**
  - **Unzulässiges Risiko:** Verbotene KI-Anwendungen.
  - **Hohes Risiko:** Strenge Anforderungen an Transparenz, Sicherheit und Compliance.
  - **Geringes oder minimales Risiko:** Weniger strenge oder keine spezifischen Anforderungen.

Erklärbarkeit als zentrale Anforderung

- **Transparenzpflichten:** Anbieter müssen Informationen bereitstellen, die es ermöglichen, die Funktionsweise von KI-Systemen zu verstehen.
- **Vertrauenswürdigkeit:** Erklärbare KI fördert das Vertrauen der Nutzer und erleichtert die Akzeptanz von KI-Technologien.

[1] Regulation (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 über harmonisierte Vorschriften für künstliche Intelligenz. Verfügbar unter: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[2] Pressemitteilung der Europäischen Kommission: "AI Act tritt in Kraft". Verfügbar unter: https://ec.europa.eu/commission/presscorner/detail/de/ip_24_1234

Herausforderungen des EU AI Acts für Unternehmen

- **Komplexität der Regulierung:**
  Der AI Act verfolgt einen risikobasierten Ansatz, bei dem KI-Systeme in verschiedene Risikoklassen eingeteilt werden. Unternehmen müssen ihre KI-Anwendungen entsprechend einstufen und die jeweiligen Anforderungen erfüllen.[3]
- **Standardisierung und technische Umsetzung:**
  Die Entwicklung harmonisierter Standards für Hochrisiko-KI-Systeme ist komplex und zeitaufwendig. Verzögerungen können zu Unsicherheiten bei der Implementierung führen und Innovationen hemmen.[4]
- **Vermeidung von Innovationshemmnissen:**
  Es besteht die Sorge, dass strenge Regulierungen Innovationen im Bereich der Künstlichen Intelligenz behindern könnten. Unternehmen müssen Wege finden, um sowohl den gesetzlichen Anforderungen zu entsprechen als auch ihre Innovationsfähigkeit zu bewahren.[5]
- **Wettbewerbsfähigkeit im internationalen Kontext:**
  Unternehmen in Regionen mit weniger strengen Vorschriften könnten schneller Innovationen umsetzen und dadurch Wettbewerbsvorteile erlangen. Europäische Firmen stehen vor der Herausforderung, trotz strengerer Regulierungen konkurrenzfähig zu bleiben.[6]

[3] https://www.it-schulungen.com/wir-ueber-uns/wissensblog/welche-anforderungen-stellt-der-eu-ai-act.html
[4] https://www.connect-professional.de/security/der-ai-act-chancen-nutzen-risiken-managen.332959.html
[5] https://www.dps-bs.de/blog/der-ai-act-weichenstellung-fuer-kuenstliche-intelligenz-in-europa/
[6] https://de.linkedin.com/pulse/der-eu-ai-act-chancen-und-herausforderungen-f%C3%BCr-andreas-quandt-ljxne

Methoden der Explainable AI (XAI)

Global interpretierbare Modelle:

- Lineare Regression
- Entscheidungsbäume
- Regelbasierte Modelle

Post-hoc Erklärungen:

- Lokale Methoden (z.B. LIME, SHAP)
- Visualisierungen (z.B. Feature Importance, PDPs)
- Gegenbeispiele (Counterfactual Explanations)

Surrogatmodelle:

- Vereinfachte Modelle, die komplexe Modelle approximieren

Globale Methoden

Feature Importance:

- Bewertung der Bedeutung einzelner Merkmale

Permutation Feature Importance:

- Bewertung durch Permutation der Merkmale

Partial Dependence Plots (PDP):

- Einfluss eines Merkmals auf die Vorhersage

Global Surrogates:

- Erklärbares Modell, das ein komplexes Modell nachahmt

Lokale Methoden

LIME:

- Lokale lineare Approximationen des Modells

SHAP:

- Spieltheorie-basierte Quantifizierung von Merkmalbeiträgen

Counterfactual Explanations:

- Minimale Änderungen für eine andere Vorhersage

Visualisierungen

Feature Importance:

- Balkendiagramme zur Darstellung der Merkmalsbedeutung

Partial Dependence Plots (PDP):

- Einfluss eines Merkmals auf die Vorhersage

Individual Conditional Expectation (ICE):

- Individuelle Effekte von Merkmalen für einzelne Datenpunkte

Lineare Modelle

Lineare Regression:

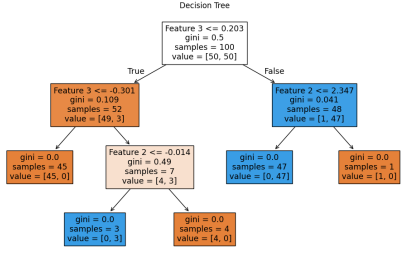
- Modell: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$
- β_0: Achsenabschnitt
- β_i: Regressionskoeffizienten
- ϵ: Fehlerterm

Generalisierte Additive Modelle (GAMs):

- Modell: $Y = \beta_0 + f_1(X_1) + \cdots + f_p(X_p) + \epsilon$
- f_i: Glatte Funktionen für nichtlineare Beziehungen

Entscheidungsbäume

- Rekursive Partitionierung des Merkmalsraums
- Jeder Knoten: Entscheidung basierend auf Merkmal und Schwellenwert
- Ziel: Maximierung der Homogenität in den Blättern



Modellagnostische Methoden

Permutation Feature Importance:

- Permutation eines Merkmals
- Messung des Anstiegs des Vorhersagefehlers
- Signifikanter Anstieg = hohe Bedeutung

Partielle Abhängigkeitsdiagramme (PDPs):

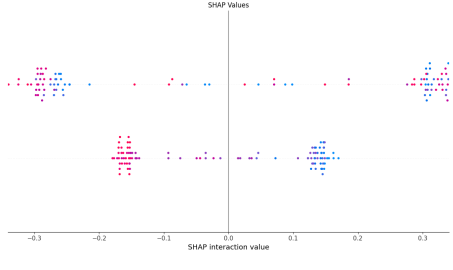
- Zeigen durchschnittliche Wirkung eines Merkmals
- Berechnung: $\hat{f}_{X_S}(x_S) = \mathbb{E}_{X_C}[\hat{f}(x_S, X_C)]$

Akkumulierte lokale Effekte (ALE):

- Messen durchschnittliche Änderung der Vorhersage
- Berechnung: $ALE_j(x) = \int_{x_{min}}^{x} \mathbb{E}\left[\frac{\partial \hat{f}(x)}{\partial x_j} \mid x_j = z\right] dz$

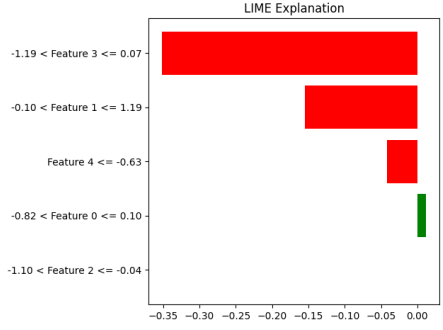
SHAP-Werte

- Basieren auf kooperativer Spieltheorie
- Beitrag jedes Merkmals zur Vorhersage:
- $\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$



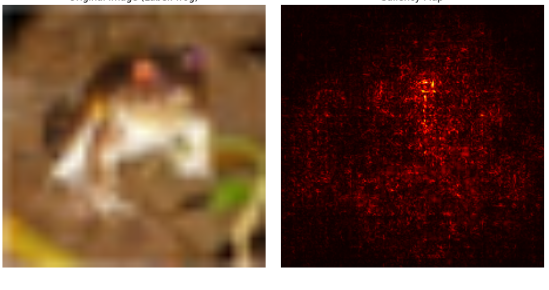
Lokale Surrogatmodelle: LIME

- Einfaches Modell (z.B. lineare Regression) lokal anpassen
- Künstliche Datenpunkte in der Nähe generieren
- Gewichtung basierend auf Ähnlichkeit zur Instanz



Methoden zur Interpretation neuronaler Netzwerke

- **Feature Visualization:**
  - Visualisierung der Merkmale, auf die Neuronen reagieren.
  - Ermöglicht Einblicke in die von Neuronen erkannten Muster.
- **Saliency Maps:**
  - Identifikation von Eingabebereichen, die den größten Einfluss auf die Ausgabe haben.
  - Darstellung der Bedeutung einzelner Pixel oder Merkmale.
- **Layer-wise Relevance Propagation (LRP):**
  - Rückverfolgung der Entscheidung des Netzwerks auf die Eingabedaten.
  - Zuweisung von Relevanzwerten zu einzelnen Eingabeelementen.



Werkzeuge und Bibliotheken für erklärbare KI

- **Captum:**
  - PyTorch-Bibliothek für Interpretierbarkeitsmethoden.
  - Unterstützt Techniken wie integrierte Gradienten und DeepLIFT.
- **ELI5:**
  - Bibliothek zur Erklärung von ML-Modellen und Vorhersagen.
  - Unterstützt verschiedene Modelle wie Sklearn, XGBoost und Keras.
- **SHAP-Bibliothek:**
  - Implementierung der SHAP-Werte für verschiedene Modelltypen.
  - Ermöglicht detaillierte Analysen der Merkmalsbeiträge.

Erklärung für Sentiment-Analyse mit LLM

- **Ziel:** Generierung und Speicherung einer Erklärung für eine Sentiment-Analyse-Aufgabe.
- **Vorgehen:**
  - Verwendung eines vortrainierten Sprachmodells (LLM) aus der Hugging Face Transformers-Bibliothek.
  - Durchführung der Sentiment-Analyse auf einem Beispieltext.
  - Visualisierung der Erklärung (z.B. Token-Wichtigkeit) als Balkendiagramm.
- **Ergebnis:**
  - Die Bedeutung einzelner Tokens wird grafisch dargestellt, um die Entscheidungsfindung des Modells zu verdeutlichen.

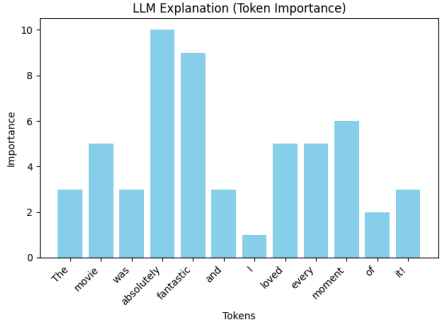


Abbildung: Visualisierung der Token-Wichtigkeit für Sentiment-Analyse

Interpretation von LLMs (Large Language Models)

- **Herausforderungen:**
  - Hohe Komplexität und Anzahl der Parameter erschweren die Nachvollziehbarkeit.
  - Entscheidungen basieren auf nichtlinearen Beziehungen zwischen Tokens.
- **Ansätze zur Interpretation:**
  - **Attention Visualisierung:**
    - Darstellung der Aufmerksamkeit (Attention Scores) zwischen Tokens.
    - Tool: BertViz.
  - **Feature Attribution:**
    - Identifikation der wichtigsten Tokens für eine Vorhersage.
    - Tools: Captum, SHAP.
  - **Neuronale Aktivierungen:**
    - Analyse der Aktivierungsmuster einzelner Neuronen.
    - Tool: Neuroscope.

Bibliotheken zur Interpretation von LLMs

- **BertViz:**
  - Visualisierung der Attention-Matrizen in Transformer-Modellen.
  - Unterstützt Modelle wie BERT, GPT-2.
  - `https://github.com/jessevig/bertviz`
- **Transformers Interpret:**
  - Feature-Attribution-Methoden für Hugging Face Modelle.
  - Unterstützt LIME, Integrated Gradients.
  - `https://github.com/cdpierse/transformers-interpret`
- **Captum:**
  - PyTorch-Bibliothek für Interpretierbarkeit.
  - Unterstützt Integrated Gradients, Layer Conductance.
  - `https://captum.ai/`
- **SHAP:**
  - Shapley-Werte für Feature Attribution.
  - Unterstützt Transformer-Modelle.
  - `https://github.com/slundberg/shap`

XAI Standardwerke

- **Interpretable ML Book** (Molnar)
  - https://christophm.github.io/interpretable-ml-book/
  - Umfassender Leitfaden zu SHAP, LIME, etc.
- **Explainable AI** (Springer 2019)
  - DOI: 10.1007/978-3-030-28954-6
  - Forschungsbeiträge zu Deep Learning Interpretability

XAI Toolkits & Frameworks

- **Quantus**
  - `https://github.com/understandable-machine-intelligence-lab/Quantus`
  - 35+ Metriken für XAI-Evaluation
  - JMLR Paper: Quantus Paper
- **SHAP**
  - `https://github.com/slundberg/shap`
  - Shapley Values für Feature Importance
- **AI Explainability 360** (IBM)
  - `https://aix360.mybluemix.net/`
  - Enthält ProtoDash, CEM

XAI Forschungsarbeiten

- **”Explainable AI: A Review”** (Gilpin 2018)
  - `https://arxiv.org/abs/1801.00631`
  - Systematische Klassifikation
- **LIME Paper** (Ribeiro 2016)
  - `https://arxiv.org/abs/1602.04938`
  - Lokale Erklärbarkeit
- **”Sanity Checks Revisited”** (Hedström 2023)
  - Verbesserte MPRT-Metriken

References I