Informatik
Hauptcampus

HOCHSCHULE TRIER

Erklärung von Klassifikatoren des maschinellen Lernens durch vielfältige kontrafaktische Erklärungen

Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations

Fabian Wagner

Seminar

Betreuer: Prof. Dr. Jörn Schneider, Marvin Schneider

Trier, 11.02.2026

Kurzfassung

Inhaltsverzeichnis

1

Einleitung und Problemstellung

Künstliche Intelligenz etabliert sich zu einem festen Bestandteil der Arbeitswelt und dem Alltag. Neben dem Einsatz als Chatbot oder Assisstenten treffen KI-Systeme vermehrt Entscheidungen, welche folgenschwere Konsequenzen auf das Leben haben. Die Nachvollziehbarkeit einer solchen Entscheidung ist für den sicheren und fairen Einsatz von KIs von zentraler Bedeutung. Erklärbare KI (xAI) befasst sich mit der Fragestellung die getroffene Klassifikation einer KI für den Menschen verständlich und nachvollziehbar zu machen.

Ziel der Seminararbeit ist die Analyse und Evaluation der Methodik von DiCE [MST20] sowie die Bewertung für die Anwendbarkeit der Methode in der Praxis. Der Ansatz von DiCE beruht auf der Erklärung durch Beispiele, die ähnlich zur ursprünglichen Eingabe sind, jedoch nicht zur gleichen Klassifikation durch die KI führen. Als erstes werden grundlegende Begriffe erklärt, welche für das Verständnis der Arbeit notwendig sind. Anschließend wird das methodische Vorgehen von DiCE untersucht. Hierzu werden die Konzepte Diversity, Proximity, Sparsity, User Constraints sowie Feasibility erläutert und deren Abhängigkeiten betrachtet. Die Bewertung von DiCE erfolgt mithilfe von Evaluationsmetriken für die einzelnen Konzepte, welche zuvor definiert werden. Weiterhin wird ein Beispiel demonstriert, an dem die Verständlichkeit einer solchen Erklärung durch Counterfactuals verdeutlicht werden soll. Abschließend folgt eine Zusammenfassung und kritische Diskussion der vorgestellten Methodik.

2

Theoretische Grundlagen

In diesem Kapitel werden zunächst wichtige Begriffe aus dem Bereich der erklärbaren KI vorgestellt und in dem Kontext von DiCE erläutert.

2.1 Grundlegende Begriffe

Ein **Modell des maschinellen Lernens** (ML-Modell) ist ein algorithmisches System, das mit Trainingsdaten zur Erkennung von Mustern durch überwachtes, unüberwachtes, halbüberwachtes oder bestärkendes Lernen trainiert wurde. Anschließend kann das Modell unbekannte Eingabedaten auf Basis des gelernten Wissens verarbeiten und klassifizieren.

Die Eingabe erfolgt in Form eines sogenannten **Feature-Vektors**, wobei die Dimension gleich der Anzahl an kategorischen und numerischen Merkmalen ist. Das Ergebnis erfolgt über die Angabe einer **Klasse**, zu der eine Eingabe durch das ML-Modell zugeordnet wurde. Die **ursprüngliche Eingabe** bezeichnet den unveränderten Feature Vector, welcher in einer **unerwünschten Klasse** resultiert. Eine **kontrafaktische Erklärung**, auch **Counterfactual** (CF) genannt, ist ein verändertes Gegenbeispiel der Eingabe, welches so gewählt wird, dass nicht die ursprünglich erhaltene, unerwünschte Klasse das Ergebnis ist, sondern eine andere Klasse. Diese erwünschte Klasse wird als **CF Klasse** bezeichnet. [MST20]

Bei ML-Modellen wird zwischen **Black-Box-** und **White-Box-Modellen** unterschieden. Black-Box-Modelle sind schwer erklärbar und für Domänenexperten unverständlich, sodass ein Ergebnis nicht nachvollziehbar für einen Menschen ist. Als White-Box-Modelle werden hingegen ML-Modelle bezeichnet, die erklärbare und nachvollziehbare Resultate liefern. [LG19]

Ein **globaler** Erklärungsansatz verfolgt das Ziel, das gesamte Entscheidungsverhalten eines ML-Modells zu verstehen. Im Gegensatz dazu verfolgen **lokale** Ansätze das Ziel ein Modell in einem eingeschränkten, weniger komplexen Lösungsraum zu erklären. [BATGL+19]

2.2 Definition von DiCE

Diverse **C**ounterfactual **E**xplanations (DiCE) ist ein Erklärungsansatz, um Entscheidungen von ML-Modellen für den Menschen verständlich zu machen. Es handelt sich um ein lokales, post-hoc Verfahren für Black-Box-Modelle. Die Erklärung des Modells findet somit nach der Trainingsphase statt und dient lediglich der Bewertung, Nachvollziehbarkeit sowie dem Aufzeigen von Schwächen oder einem Bias in den Ausgaben des Modells. DiCE generiert zu einer Eingabe verschiedene Counterfactuals, um dem Anwender aufzuzeigen, welche Parameter sich für eine andere Klassifikation durch das ML-Modell ändern müssen. Die Counterfactuals werden so generiert, dass sie möglichst unterschiedlich sind, dabei jedoch möglichst nahe an der ursprünglichen Eingabe bleiben. [MST20]

3

Methodisches Vorgehen

In diesem Kapitel werden die zentralen Konzepte von DiCE betrachtet, welche zur Generierung der Counterfactuals benötigt werden. Darunter fallen Proximität, Diversität und Sparsität. Abschließend wird die Verlustfunktion als Möglichkeit der Optimierung von Counterfactuals untersucht.

3.1 Zentrale Konzepte

Diversität beschreibt, wie sich generierten Counterfactuals voneinander unterscheiden. Eine hohe Vielfältigkeit zeigt dem Anwender nicht nur mehrere Möglichkeiten zum Erreichen einer anderen Klassifikation auf, wodurch sich die Machbarkeit erhöht, sondern lässt auch größere Rückschlüsse auf das Entscheidungsverhalten des ML-Modells schließen. Um Diversität zu berücksichtigen, wird in DiCE das Konzept der **D**eterminantal **P**oint **P**rocesses (DPP) verwendet, um das *Subset Selection Problem* zu lösen. Das Problem beschreibt dabei die Auswahl von wenigen CFs aus einer unendlich großen Menge an möglichen Beispielen, welche zeitgleich gültig als auch divers sind. In Gleichung 3.1 beschreibt $dist(c_i, c_j)$ die Distanz zwischen zwei Counterfactuals. Somit führt eine kleine Ähnlichkeit $K_{i,j}$ der CFs zu einer großen Determinante $det(K)$ und Maximierung der Diversität. [MST20, KT$^+$12]

$$dpp\_diversity = det(K), \text{ mit } K_{i,j} = \frac{1}{1 + dist(c_i, c_j)} \tag{3.1}$$

Diversität alleine ist nicht ausreichend, um einem Anwender eine Erklärung zu geben. Die generierten CFs sollten nicht nur unterschiedlich sein, sondern müssen möglichst nah an der ursprünglichen Eingabe sein. Diese **Proximität** ist für die Machbarkeit von zentraler Bedeutung, da Anwender den meisten Nutzen aus ähnlichen Counterfactuals erhalten. Die Proximität eines CFs ergibt sich aus der negative Distanz zwischen dem Counterfactual c_i und dem Feature-Vektor x. Eine geringe Distanz resultiert in einer hohen Proximität. Die Berechnung der mittleren Proximität einer Menge von CFs ist in Gleichung 3.2 dargestellt.

$$Proximity = -\frac{1}{k}\sum_{i=1}^{k} dist(c_i, x) \tag{3.2}$$

Eine weitere Eigenschaft für die Machbarkeit oder auch Umsetzbarkeit der kontrafaktischen Beispiele ist die **Sparsität**. Nach der Proximität ist auch ein Counterfactual nahe an einer Eingabe, wenn jeder Vektoreintrag minimal geändert wird. Dies ist zwar mathematisch korrekt, vernachlässigt aber den Umstand der Machbarkeit für einen Anwender. Ein Counterfactual ist einfacher umzusetzen, wenn sich möglichst wenige Eigenschaften ändern.

3.2 Optimierung

$$C(x) = argmin\frac{1}{k}\sum_{i=1}^{k} yloss(f(c_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(c_i, x) - \lambda_2 dpp_diversity(c_1, ..., c_k) \tag{3.3}$$

4

Evaluation der Methode

4.1 Evaluationsmetriken

- macht dice auch das was es soll (löst es das problem) - wie gut ist es? - wie lauten die Metriken welche zur auswertung verwendet wurden? - sind die metriken sinnvoll und gut?

- distanzfunktionen in grundlagen kapitel oder hier lassen?

$$dist\_cont(c, x) = \frac{1}{d_{cont}} \sum_{p=1}^{d_{cont}} \frac{|c^p - x^p|}{MAD_p} \tag{4.1}$$

$$dist\_cat(c, x) = \frac{1}{d_{cat}} \sum_{p=1}^{d_{cat}} I(c^p \neq x^p) \tag{4.2}$$

- Validity

$$\%ValidCFs = \frac{|\{uniqueinstancesinCs.t.f(c) > 0.5\}|}{k} \tag{4.3}$$

- Proximity

$$ContinousProximity : -\frac{1}{k} \sum_{i=1}^{k} dist\_cont(c_i, x) \tag{4.4}$$

$$CategoricalProximity : 1 - \frac{1}{k} \sum_{i=1}^{k} dist\_cat(c_i, x) \tag{4.5}$$

- Sparsity

$$Sparsity : 1 - \frac{1}{kd} \sum_{i=1}^{k} \sum_{l=1}^{d} 1_{[c_i^l \neq x_i^l]} \tag{4.6}$$

- Diversity

$$Diversity : \Delta = \frac{1}{C_k^2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} dist(c_i, c_j) \tag{4.7}$$

- Count-Diversity

$$Count - Diversity : \Delta = \frac{1}{C_k^2 d} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{l=1}^{d} 1_{[c_i^l \neq c_j^l]} \tag{4.8}$$

4.2 Bewertung

- ist DICE besser als zb LIME? - Gründe für die Unterschiede

5

Anwendbarkeit und Bedeutung in der Praxis

5.1 Relevanz in der Praxis

5.2 Demonstration eines Beispiels

6

Zusammenfassung und Ausblick

- zusammenfassung - Herausforderungen - wo liegen die grenzen von dice? - was ist problematisch und was benötigt weitere forschung

Literaturverzeichnis

BATGL+19. Barredo Arrieta, Alejandro, Siham Tabik, Salvador García López, Daniel Molina Cabrera, Francisco Herrera Triguero, Natalia Ana Díaz Rodríguez et al.: *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.* 2019.

KT+12. Kulesza, Alex, Ben Taskar et al.: *Determinantal point processes for machine learning.* Foundations and Trends® in Machine Learning, 5(2–3):123–286, 2012.

LG19. Loyola-González, Octavio: *Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View.* IEEE Access, 7:154096–154113, 2019.

MST20. Mothilal, Ramaravind K, Amit Sharma und Chenhao Tan: *Explaining machine learning classifiers through diverse counterfactual explanations.* In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Seiten 607–617, 2020.

A

Eigenständigkeitserklärung

☐ Die vorliegende Arbeit wurde als Einzelarbeit angefertigt.

☐ Die vorliegende Arbeit wurde als Gruppenarbeit angefertigt. Mein Anteil an der Gruppenarbeit ist im untenstehenden Abschnitt *Verantwortliche* dokumentiert:

☐ Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne unzulässige Hilfe Dritter angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate als solche kenntlich gemacht. Darüber hinaus erkläre ich, dass ich die vorliegende Arbeit in dieser oder ähnlicher Form noch nicht als Prüfungsleistung eingereicht habe.

☐ Es ist keine Nutzung von KI-basierten text- oder inhaltgenerierenden Hilfsmitteln erfolgt.

☐ Die Nutzung von KI-basierten text- oder inhaltgenerierenden Hilfsmitteln wurde von der/dem Prüfenden ausdrücklich gestattet. Die von der/dem Prüfenden mit Ausgabe der Arbeit vorgegebenen Anforderungen zur Dokumentation und Kennzeichnung habe ich erhalten und eingehalten. Sofern gefordert, habe ich in der untenstehenden Tabelle *Nutzung von KI-Tools* die verwendeten KI-basierten text- oder inhaltgenerierenden Hilfsmittel aufgeführt und die Stellen in der Arbeit genannt. Die Richtigkeit übernommener KI-Aussagen und Inhalte habe ich nach bestem Wissen und Gewissen überprüft.

Datum

Unterschrift der Kandidatin/des Kandidaten

Nutzung von KI-Tools

KI-Tool	**Genutzt für**	**Warum?**	**Wann?**	**Mit welcher Eingabefrage bzw. -aufforderung?**	**An welcher Stelle der Arbeit übernommen?**
ChatGPT, DeepSeek, Gemini	Korrektur bzgl. Rechtschreibung, Grammatik und Formulierungen	Verbesserung der Textqualität	Während der gesamten Arbeit	Textabschnitte mit der Aufforderung zur Kontrolle	-