

Erklärung von Klassifikatoren des maschinellen Lernens
durch vielfältige kontrafaktische Erklärungen

Explaining Machine Learning Classifiers through Diverse Counter-
factual Explanations

Fabian Wagner

Seminar

Betreuer: Prof. Dr. Jörn Schneider, Marvin Schneider

Trier, 11.02.2026

Kurzfassung

Inhaltsverzeichnis

1	Einleitung und Problemstellung	1
2	Theoretische Grundlagen	2
2.1	Grundlegende Begriffe	2
2.2	Definition von DiCE	2
3	Methodisches Vorgehen	4
3.1	Zentrale Konzepte	4
3.2	Optimierung	4
4	Evaluation der Methode	5
4.1	Evaluationsmetriken	5
4.2	Bewertung	6
5	Anwendbarkeit und Bedeutung in der Praxis	7
5.1	Relevanz in der Praxis	7
5.2	Demonstration eines Beispiels	7
6	Zusammenfassung und Ausblick	8
	Literaturverzeichnis	9
	Eigenständigkeitserklärung	10

Einleitung und Problemstellung

Künstliche Intelligenz etabliert sich zu einem festen Bestandteil der Arbeitswelt und dem Alltag. Neben dem Einsatz als Chatbot oder Assistenten 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.

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

iiiiii HEAD 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 als ein sogenannter Feature Vector, wobei die Dimension des Vektors gleich der Anzahl an kategorischen und numerischen Merkmalen ist. Kategorische Merkmale sind ...

===== 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. *llllllll* 40b0b912b72bae12dae93ee0ce3cc7c58e88d19

Die Eingabe erfolgt als ein sogenannter *Feature Vector*, 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 *ungewollten Klasse* resultiert. Ein *Counterfactual* (CF) ist hingegen eine leicht veränderte Eingabe. Das CF wird so gewählt, dass nicht die ursprünglich erhaltene, ungewollte Klasse das Ergebnis ist, sondern eine sogenannte *CF Klasse*.

2.2 Definition von DiCE

Diverse **C**ounterfactual **E**xplanations 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 der Bewertung, Nachvollziehbarkeit sowie dem Aufzeigen von Schwächen oder einem Bias des Modells.

Methodisches Vorgehen

3.1 Zentrale Konzepte

- Diversity via dpp

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

- Proximity

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

- Sparsity -

3.2 Optimierung

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

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?

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

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

- Validity

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

- Proximity

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

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

- Sparsity

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

- Diversity

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

- Count-Diversity

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

4.2 Bewertung

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

Anwendbarkeit und Bedeutung in der Praxis

5.1 Relevanz in der Praxis

5.2 Demonstration eines Beispiels

Zusammenfassung und Ausblick

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

Literaturverzeichnis

- 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.

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	-