LRP

goKI

Layer-Wise Relevance Propagation

Lokal

Modell-spezifisch

Merkmals-Gewichtung

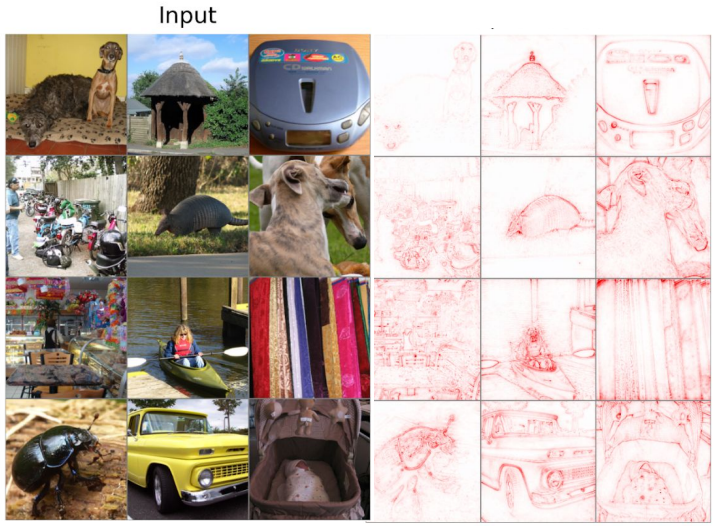
Post-Hoc Erklärung

Evaluieren

Lernen

Methode

Layer-Wise-Relevance Propagation (LRP) ist eine Methode um die Relevanz konkreter Eingabewerte eines neuronalen Netzes zu einer spezifischen Ausgabe zu berechnen. Die Relevanz kann anschließend z.B. für bildbasierte Netze durch eine sog. Heatmap dargestellt werden (siehe rechts). Für die Berechnung der Relevanz wird der zu untersuchende Ergebniswert des Netzes durch Anwendung eigens dafür entwickelter Übertragungsregeln “zurück” gespielt, wobei aber die Summe der Relevanz jeder Netzwerkschicht gleich bleibt.



Anwendungsbereich

Entwickelt für Convolutional Neural Networks (CNNs) und Support Vector Machines (SVMs). Die Methode wurde inzwischen auf andere Arten von Modellen wie Recurrent Neural Networks (RNNs), One-Class Support Vector Machines und K-Means-Clustering erweitert. Sehr schnelle Berechnung, Echtzeitfähig.

Einschränkungen

- Benötigt Zugriff auf das Neuronale Netzwerk und sollte an die Architektur angepasst werden.
- Kann im Vergleich zu anderen Methoden mehr Expertenwissen erfordern um die verschiedenen LRP-Varianten korrekt anzuwenden.
- Parametereinstellungen haben großen Einfluss auf die Qualität der Erklärung.

Paper
https://bit.ly/47l2Kin

Implementation
https://bit.ly/3Qvl3eQ

SHAP

goKI

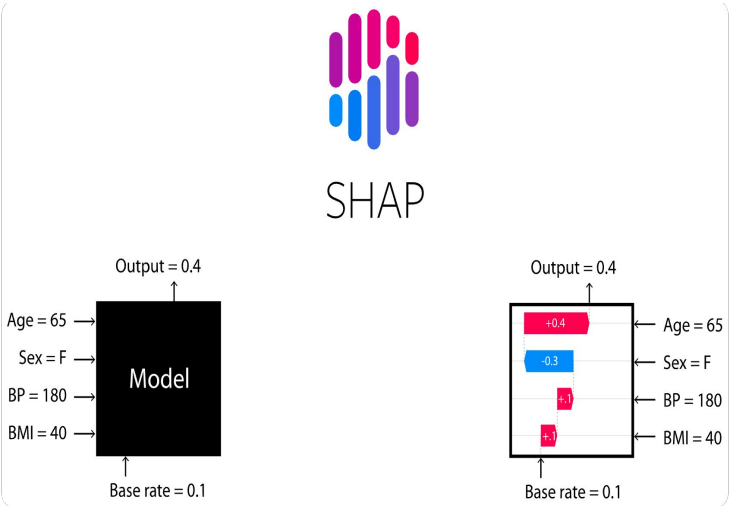
SHapley Additive exPlanations

Lokal/Global | Modell-agnostisch | Evaluieren

Merkmals Gewichtung | Post-Hoc Erklärung | Lernen

Methode

Auf Basis der kooperativen Spieltheorie berechnet SHAP für jedes Eingabemerkmal einen Wert, der ausdrückt, wie stark der Einfluss des individuellen Merkmals auf das Vorhersageergebnis ist. SHAP soll dadurch eine relativ einfache Interpretation der Eingabemerkmale für eine breite Zielgruppe ermöglichen. Diese Methode findet aufgrund ihrer theoretischen Fundierung, der vielfältigen Verwendungsmöglichkeiten und hoher Genauigkeit breite Anwendung in Wirtschaft und Forschung.



Anwendungsbereich

SHAP wird für eine Vielzahl von Modellen empfohlen, einschließlich lineare Modelle, baumbasierte Modelle (wie Entscheidungsbäume und Random Forests), Support Vector Machines (SVMs), neuronale Netze und Ensemblemodelle.

Einschränkungen

- Rechenintensiv.
- Interpretation der SHAP Werte sehr komplex bei korrelierten Merkmalen oder anderer Wechselwirkungen.
- Mögliche Einschränkungen bei der Erfassung aller Aspekte des Modellverhaltens, z. B. bei ungleicher Verteilung im Datensatz.

Paper
https://bit.ly/3s5anJF

Implementation
https://bit.ly/3Yp8nle

LIME

goKI

Lokal Interpretable Model-Agnostic Explanations

Lokal

Modell-agnostisch

Modellvereinfachung

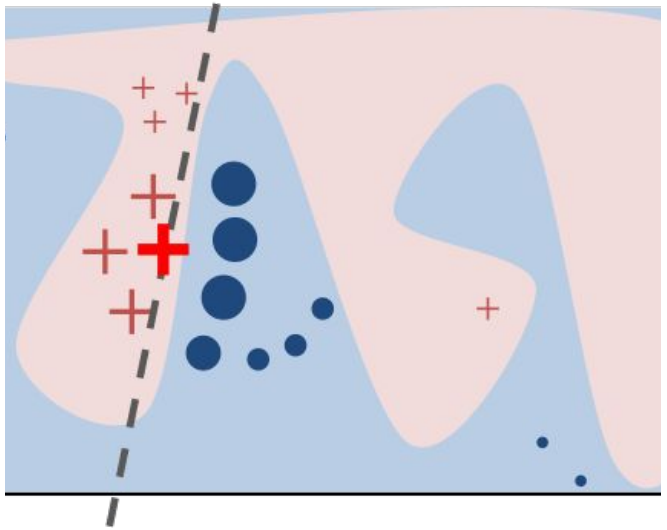
Post-hoc Erklärung

Vertrauen

Verbessern

Methode

LIME generiert lokale Erklärungen. Es vereinfacht komplexes Modellverhalten durch simple und dadurch interpretierbare Regeln. Durch einen Prozess, der als Perturbation Sampling bezeichnet wird, stört LIME Merkmale von Eingabeinstanzen und bewertet die daraus resultierende Auswirkung auf die Ausgabe des Modells. LIME bewertet die Merkmale von Eingabeinstanzen am wichtigsten, deren Störung den größten Einfluss auf das Endergebnis hat.



Anwendungsbereich

LIME ist in der Lage, Vorhersagen beliebiger Modellarchitekturen zu erklären, wie z.B. Random Forests, neuronale Netze und mehr. Im Vergleich zu SHAP wird LIME oft für Datensätze mit vielen Merkmalen bevorzugt, da der Sampling-Ansatz von LIME hilft, die Komplexität zu reduzieren.

Einschränkungen

- Rechenintensiv (wegen benötigten Stichproben).
- Zuverlässigkeit und Vollständigkeit der Erklärungen hängt vom Sampling ab. Es können systemische Muster oder Bias unerkannt bleiben, sowie irreführende Erklärungen auftreten.
- Näherungen erfassen nicht die volle Komplexität bestimmter Modelle oder ihrer Wechselwirkungen.

Paper
https://bit.ly/450sD5l

Implementation
https://bit.ly/3YxpJms

ProtoSeg

goKI

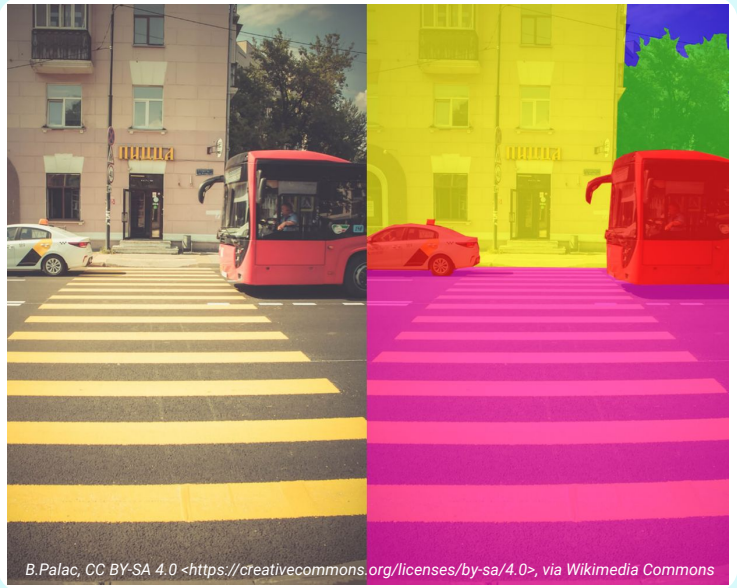
Interpretable Semantic Segmentation with Prototypical Parts

Lokal | Modell-spezifisch | Vertrauen

Prototypen | Interpretable Model | Lernen

Methode

ProtoSeg ist ein Modell für interpretierbare Segmentierung, das Prototypen als Erklärmethode verwendet. ProtoSeg lernt Prototypen, d.h. Referenzbeispiele, für jede Klasse und verwendet diese, um Segmentierungen mit Patches (Fällen) aus der Trainingsmenge zu erzeugen und so zu erklären. Es besteht aus einem Backbone-Netz, einem Prototypen Layer und einem fully-connected Layer. Während das Backbone-Netz das Bild segmentiert, verbinden die Prototyp und die fully-connected Layer die Segmente mit ähnlichen Bildausschnitten.



Anwendungsbereich

ProtoSeg ist ein interpretierbares Modell, d.h. es erzeugt erklärbare Ergebnisse per Design. Es ist daher sowohl ein Modell als auch eine Erklärbarkeitsmethode. Vereinfacht gesagt handelt es sich bei ProtoSeg um eine XAI-Methode für CNN, da ProtoSeg selbst ein Convolutional Neural Network (CNN) ist.

Einschränkungen

- Spezifische Erklärbarkeitsmethode nur für Segmentierungsaufgaben und daher weder Modell- noch Aufgaben-agnostisch.
- Beeinträchtigte Präzision im Vergleich zu Segmentierungsmodellen ohne integrierte Erklärbarkeit (Accuracy-Interpretability Trade-off).

Paper
https://bit.ly/3Oug4sh

Implementation
https://bit.ly/3DP7DCE

CEM

goKI

Contrastive Explanation Method

Lokal | Modell-spezifisch

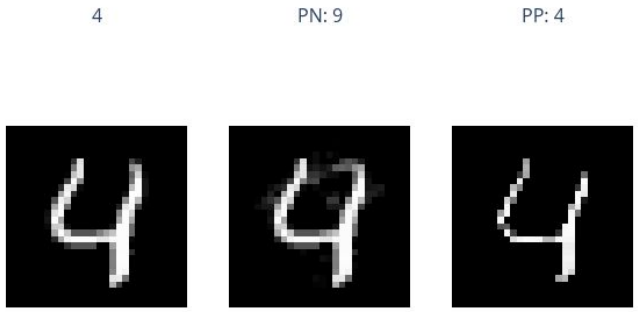
Kontradiktion | Post-hoc Erklärung

Vertrauen

Lernen

Methode

CEM ist eine Methode die mit kontrastiven Erklärungen arbeitet. Diese erläutern, wieso ein bestimmtes Ergebnis und nicht ein anderes anstatt erreicht wurde. Sie unterteilen erklärende Elemente in pertinent Positive (PP) und pertinent Negative (PN) Faktoren. Die PP sind eine minimale Menge an Faktoren, die für ein bestimmtes Ergebnis sprechen. Die PN sind die minimale Menge an zusätzlichen Faktoren die zu einem anderen Ergebnis führen würden. Ein Beispiel für PN wäre das Bild in der Mitte, wo weiße Bereiche kennzeichnen was erforderlich wäre, statt der ursprünglichen Zahl 4 eine 9 zu erkennen.



Erklärungsansatz

CEM ist eine Erklärbarkeitsmethode für Klassifizierungsaufgaben, kann aber auch für Regression oder Clustering verwendet werden. Sie bietet sich daher für verschiedene Methoden wie neuronale Netze, Random Forests, logistische Regression oder Support Vector Machines (SVM) an.

Einschränkungen

- Eingeschränkte Interpretierbarkeit bei komplexen Modellen.
- Schwierigkeit bei der Verarbeitung kategorischer oder diskreter Merkmale.
- Empfindlichkeit gegenüber dem Ausmaß der Störung der Eingabemerkmale.

Paper
https://bit.ly/3YyaXf5

Implementation
https://bit.ly/3KBqZiv

Anchors

goKI

High-Precision Model-Agnostic Explanations

Lokal | Modell-agnostisch | Regel-basiert | Post-hoc Erklärung

Vertrauen | Evaluieren | Legitimieren | Lernen

Methode

Die Anchor-Methode sucht nach interpretierbaren Bedingungen (Anchors), die die Entscheidungsgrenze um einen bestimmten Datenpunkt herum ausreichend beschreiben. Die generierten Regeln haben die Form "WENN Bedingung DANN Vorhersage" und zielen darauf ab, die Bedingungen zu erfassen, unter denen die Vorhersage des Modells konsistent ist. Die sich daraus ergebenden Anchors liefern verständliche und von Menschen überprüfbare Erklärungen für die Modellvorhersagen.

Beispiel mit Text

IF Education = Bachelors **AND** Relationship = Husband **AND** Hours per week > 45.00 **AND** Race = White **AND** Country = United-States **AND** 28.00 < Age <= 37.00 **THEN** Salary > 50k

Beispiel mit Bild

Anwendungsbereich

Anchors können aufgrund ihrer Modellunabhängigkeit auf eine Vielzahl von linearen und nichtlinearen Modellen angewendet werden. Solange das Modell Vorhersagewahrscheinlichkeiten oder Konfidenzwerte für jede Klasse liefert, kann es verwendet werden.

Einschränkungen

- Bietet möglicherweise kein umfassendes, globales Verständnis des Verhaltens des Modells.
- Empfindlich gegenüber Störungen in den Daten.
- Anchor Regeln sind von Natur aus vereinfacht und können die Feinheiten des zugrundeliegenden Modells nicht vollständig erfassen.

Paper
https://bit.ly/45lpm0e

Implementation
https://bit.ly/3qlzNll

t-SNE

goKI

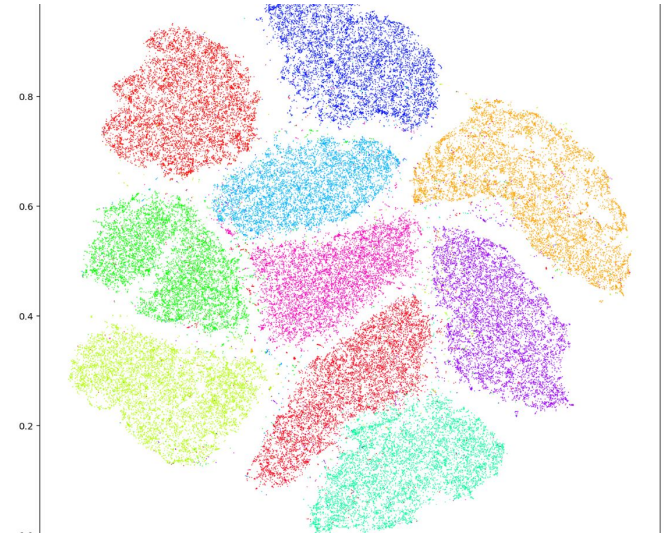
t-Stochastic Neighbor Embedding

Global | Modell-agnostisch | Evaluieren

Statistisch | Datensatz Erklärung | Lernen

Methode

Das Hauptziel von t-SNE besteht darin, komplexe, hochdimensionale Datensätze in einem Raum mit niedriger Dimensionalität abzubilden. Dies sind klassischerweise zwei oder drei Dimensionen, da sich diese graphisch darstellen lassen. In möglichen Darstellungen werden ähnliche Datenpunkte näher beieinander und sich unterscheidende Punkte weiter voneinander entfernt gruppiert. Dies trägt dazu bei, zugrunde liegende Muster und Cluster in den Daten aufzudecken, die im ursprünglichen, hochdimensionalen Raum möglicherweise nicht sofort ersichtlich gewesen wären.



Anwendungsbereich

t-SNE kann auf eine Breite von Modellen und Datensätzen angewendet werden, da es sich in erster Linie um eine non-lineare Technik zur Reduktion von Dimensionalität und Datenvisualisierung handelt, die über klassische ML-Modelle hinausgeht. Es kann für (Un-) Supervised Learning, NLP oder Feature Engineering verwendet werden.

Einschränkungen

- Keine eigenständige XAI-Methode, sollte in Kombination mit anderen XAI-Methoden verwendet werden.
- Liefert nur statische, zwei- oder dreidimensionale Datendarstellungen.
- Starke Abhängigkeit der Ergebnisse von den Eingabeparametern.

Paper
https://bit.ly/43ZN6WI

Implementation
https://bit.ly/3ONAk9p

MC

goKI

Model Cards Dokumentation

Global | Modell-agnostisch | Vertrauen

Redaktionell | Modell Erklärung | Legitimieren

Dokumentation – Model Cards

goKI

Methode

Model Cards, fassen die wesentlichen Informationen über ein ML-Modell kompakt und übersichtlich zusammen. Sie können Informationen zum Zweck, Leistung, Einschränkungen und möglichen Problemen enthalten. Eine gängige Anwendung ist die Erstellung als Markdown-Dateien mit zusätzlichen YAML Abschnitten als Metadaten, die eine Gruppierung über Modelle hinweg ermöglichen.

Paper
https://bit.ly/443nvw9

Beispiel

- **Modellarchitektur:** Architektur, einschließlich der Art des Modells (z. B. Deep Neural Network, Random Forest) und seiner Komponenten.
- **Verwendungszweck:** Zweck und Anwendung des Modells sowie etwaige Einschränkungen.
- **Metriken:** Angaben zur Leistung des Modells in Bezug auf verschiedene Bewertungsmetriken wie Genauigkeit, Präzision oder F1-Score.
- **Daten:** Informationen zu den Datensätzen, die zum Trainieren und Validieren des Modells verwendet wurden, einschließlich der Größe, der Quelle und möglicher Bias in den Daten.
- **Fairness und Bias:** Möglicher Bias oder Fairnessprobleme im Zusammenhang mit dem Modell, insbesondere im Hinblick auf verschiedene demografische Gruppen.
- **Sicherheit:** Informationen über die Robustheit des Modells gegenüber gegnerischen Angriffen und seine allgemeine Sicherheit.
- **Einsatz:** Hinweise zum verantwortungsvollen Einsatz unter Berücksichtigung potenzieller Risiken und geeigneter Anwendungsfälle.
- **Versionsgeschichte :** Aufzeichnungen