LRP

goKI

Layer-wise relevance propagation

Local

Model specific

Feature Importance

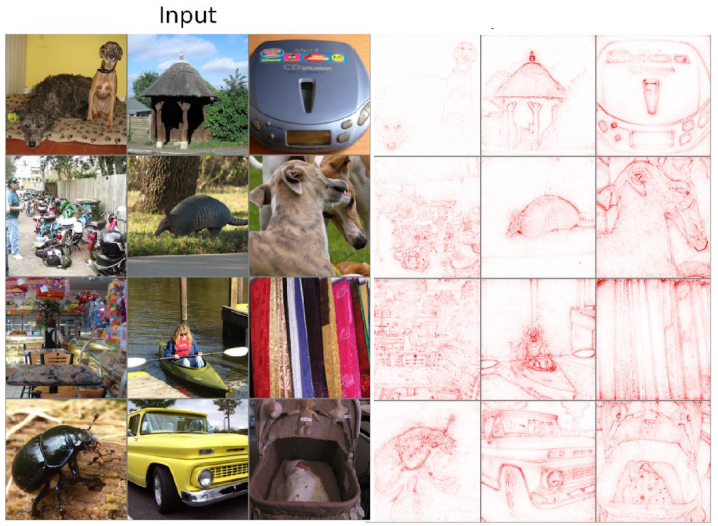
Ad-Hoc Explainer

Evaluate AI

Learn from AI

About

Propagating the prediction backwards in the neural network, using a set of purposely designed propagation rules. The LRP method is based on a relevance conservation principle and leverages the structure of the model to decompose its prediction. Initially developed by Fraunhofer Institute, LRP is one of the most used XAI methods with numerous applications in research and industry.



Application

Designed for deep convolutional neural networks (CNNs) and kernel machines (SVMs). The method has since been extended to other types of AI models such as Recurrent Neural Networks (RNNs), One-Class Support Vector Machines, and K-Means clustering.

Limitations

- Sensitivity to network architecture and configuration.
- Different parameter settings may produce different results.
- May require more expert knowledge compared to other methods.

Paper
https://bit.ly/47l2Kin

Implementation
https://bit.ly/3Qvl3eQ

SHAP

goKI

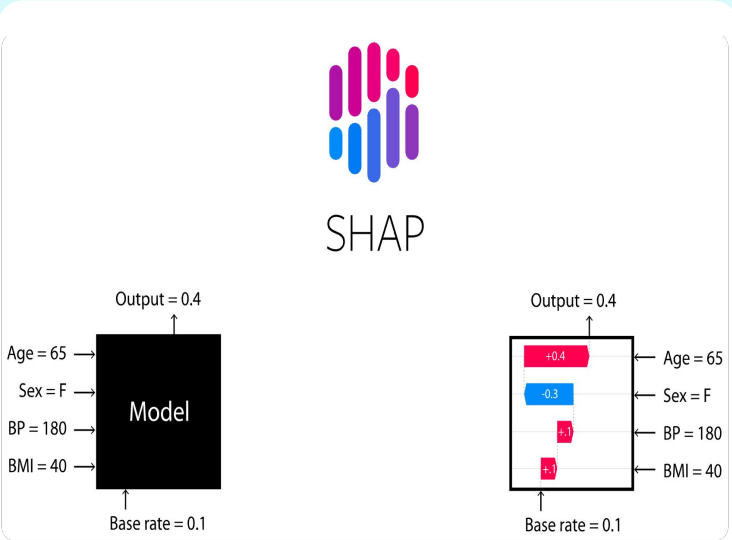
SHapley Additive exPlanations

Local/Global | Model agnostic | Evaluate AI

Mixed Approach | Post-Hoc Explainer | Learn from AI

About

Based on cooperative game theory, SHAP assigns each feature a unique attribution value, measuring its contribution to the prediction outcome. It provides a principled framework for quantifying feature importance, enabling users to understand the factors driving model predictions. This method has gained popularity due to its theoretical grounding, interpretability, and ability to handle complex models effectively.



Application

SHAP is recommended for a wide range of models, including but not limited to linear models, tree-based models (such as decision trees and random forests), support vector machines (SVMs), neural networks, and ensemble models.

Limitations

- Computationally expensive.
- Challenging interpretation in the presence of correlated features or interactions.
- Potential limitations in capturing all aspects of model behaviour, e.g. in cases where the model assumptions or data distribution are violated.

Paper
https://bit.ly/3s5anJF

Implementation
https://bit.ly/3Yp8nle

LIME

goKI

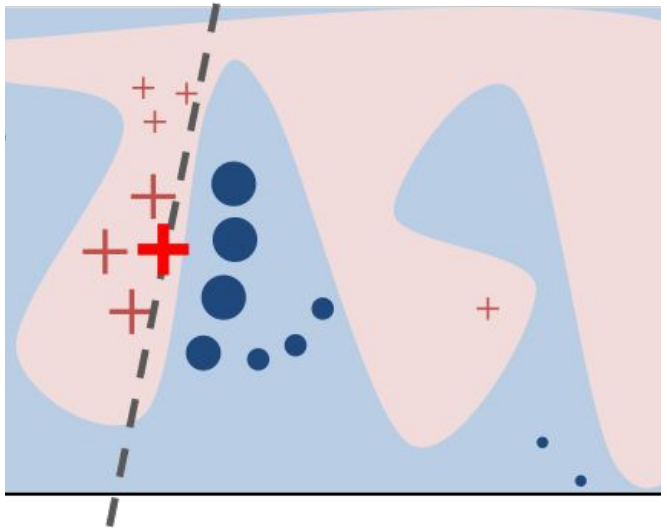
Local Interpretable Model-Agnostic Explanations

Local | Model agnostic | Surrogate Models | Post-hoc Explainer

Accept AI | Improve AI

About

By generating locally faithful explanations, LIME provides insight into individual predictions by approximating complex model behaviour with a set of interpretable rules. Through a process called perturbation sampling, LIME perturbs input instances and assesses the resulting impact on the model's output, allowing for the identification of influential features and their contributions to predictions.



Application

LIME is able to explain predictions of any model architectures, such as random forests, neural networks, and more. In comparison to SHAP, LIME is often preferred for datasets with many features, as LIME's perturbation-based sampling approach helps to reduce complexity.

Limitations

- Computationally expensive
- Potentially missing out on important systemic patterns or biases present in the data.
- Potentially misleading explanations by sampling bias.
- Simplifications or approximations may not capture the full complexity of certain models or their interactions.

Paper
https://bit.ly/450sD5l

Implementation
https://bit.ly/3YxpJms

ProtoSeg

goKI

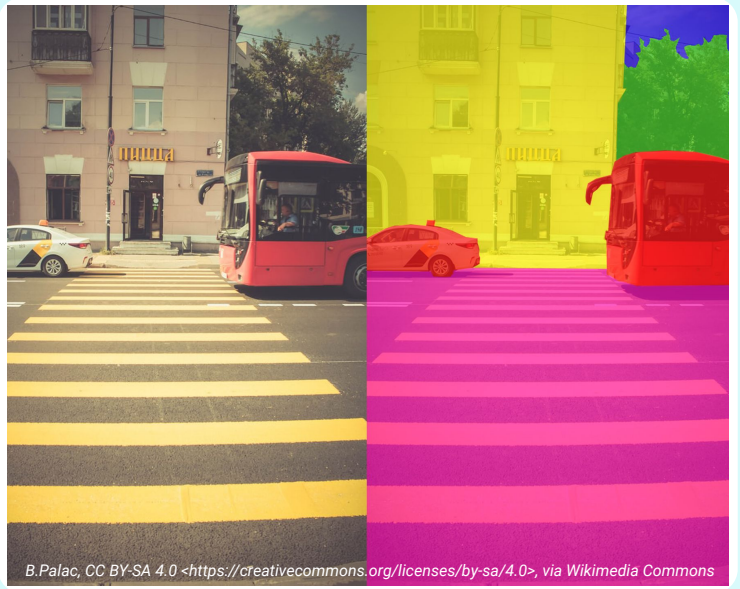
Interpretable Semantic Segmentation with Prototypical Parts

Local | Model specific | Accept AI

Prototypes | Interpretable Model | Learn from AI

About

ProtoSeg is a model for interpretable segmentation that uses prototypes as explanation method. ProtoSeg learns prototypes for each class and uses them to generate and explain segmentation with patches (cases) from the training set. It consists of a backbone network, prototype layer, and a fully connected layer. While the backbone network segments the image, the prototype and fully connected layers connect the segments with similar image patches.



Application

ProtoSeg is an interpretable model, which means it produces explainable results by design. It is therefore both a model and an explainability method. Roughly speaking, ProtoSeg is a convolutional neural network (CNN), so ProtoSeg is an XAI method for CNN.

Limitations

- Specific explainability method only for segmentation tasks.
- Impaired precision compared to segmentation models without integrated explainability.

Paper
https://bit.ly/3Oug4sh

Implementation
https://bit.ly/3DP7DCE

CEM

goKI

Contrastive Explanation Method

Local | Model specific

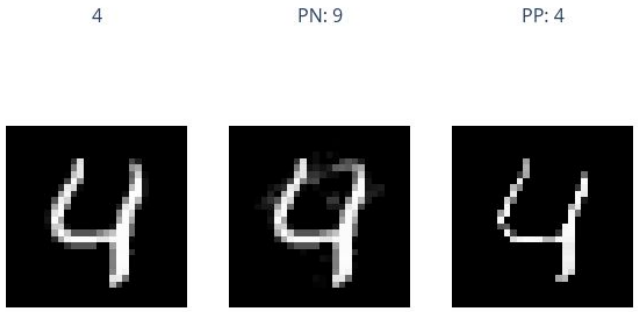
Counterfactuals | Post-hoc explainer

Accept AI

Learn from AI

About

CEM is an explanation method that provides contrastive explanations justifying the classification of an input by a black box classifier such as a deep neural network. It divides explanatory elements into pertinent positives (PP) and pertinent negatives (PN). PP is a factor whose presence is minimally sufficient and PN whose absence is necessary in justifying a certain classification.



Application

CEM is an explanation method for classification tasks but can also be used for regression or clustering. It can therefore be applied to various statistical methods such as neural networks, random forests, logistic regression or support vector machines (SVM).

Limitations

- Limited interpretability for complex models.
- Difficulty in handling categorical or discrete features.
- Sensitivity to perturbation magnitude of input features.

Paper
https://bit.ly/3YyaXf5

Implementation
https://bit.ly/3KBqZiv

Anchors

goKI

High-Precision Model Agnostic Explanations

Local | Model agnostic | Rule based | Post-hox explainer

Accept AI | Evaluate AI | Justify AI | Learn from AI

About

The Anchors method searches for interpretable conditions (rules) that sufficiently describe the decision boundary around a given data point. The generated rules take the form "IF condition THEN prediction," and they aim to capture the conditions under which the model's prediction is consistent. The resulting anchors provide understandable and human-verifiable explanations for model predictions.

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

Application

Anchors can be applied to a wide range of linear and non-linear models due to their model-agnostic nature. As long as the model provides prediction probabilities or confidence scores for each class, it can be used.

Limitations

- May not provide a comprehensive global understanding of the model's behaviour.
- Sensitive to data perturbations.
- Anchor rules are inherently simplified and may not fully capture the intricacies of the underlying model.

Paper
https://bit.ly/45lpm0e

Implementation
https://bit.ly/3qlzNll

t-SNE

goKI

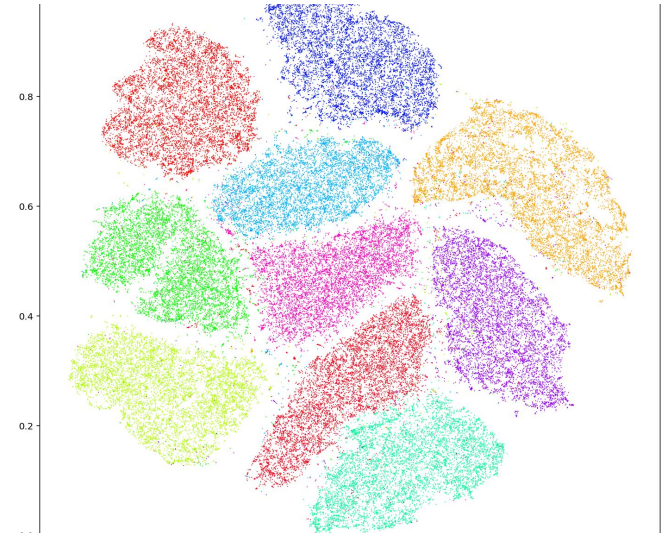
t-Stochastic Neighbor Embedding

Global | Model agnostic | Evaluate AI

Statistics | Dataset Explainer | Learn from AI

About

The main goal of t-SNE is to map data points from their original high-dimensional space to a lower-dimensional space, such that similar data points are represented closer together, while dissimilar points are positioned further apart. This helps to reveal underlying patterns and clusters in the data that might not be immediately apparent in the original high-dimensional space.



Application

t-SNE can be applied to a wide range of models and datasets, as it is primarily a dimensionality reduction and data visualization technique which go beyond classical ML models. It can be used for (Un-) Supervised Learning, NLP or Feature Engineering.

Limitations

- No stand-alone method, should be used in combination with other XAI methods.
- Only provides static, two-dimensional data representations.
- Sensitive to its perplexity parameter, which can lead to differing results.

Paper
https://bit.ly/43ZN6WI

Implementation
https://bit.ly/3ONAk9p

Documentation – Model Cards

goKI

About

Model cards, for example used in Huggingface, are documents that provide essential information about a machine learning model, particularly regarding its intended use, performance, limitations, and potential biases. A common application is to create them as Markdown files with additional YAML sections as metadata, which enable filtering or grouping on a selection of models.

Paper
https://bit.ly/443nvw9

- **Model architecture:** Architecture, including the type of model (e.g., deep neural network, random forest) and its components (layers, nodes, etc.).
- **Intended use:** Information about the primary purpose and application of the model, along with any limitations or specific scenarios where the model is most effective.
- **Performance metrics:** Details about the model's performance on various evaluation metrics, such as accuracy, precision, recall, F1 score, etc.
- **Data:** Information about the datasets used to train and validate the model, including the size, source, and potential biases in the data.
- **Fairness and bias considerations:** Discussion of any potential bias or fairness issues related to the model, particularly in terms of different demographic groups.
- **Robustness and security:** Information about the model's robustness against adversarial attacks and its overall security.
- **Deployment considerations:** Guidance on how to deploy and use the model responsibly, taking into account potential risks and appropriate use cases.
- **Model version history:** Records

Documentation – Fact Sheets

goKI

About

Documentation in a structured layout can be a great tool for ensuring justification of AI. It should contain all information related to data engineering, training, testing and tuning. The information should provide insights about quality, responsibility, sources and inspections for unwanted behaviour. This slide presents an extendible list of exemplary questions such documentation could contain.

- **Which data** was used for training and validation?
- Is **personal or health data** used for training, validation or operation?
- **How and where** was the data acquired?
- From **which distribution** was it sampled?
- **Who labeled** the data?
- What is their **expertise for labeling** the data (e.g. student, assistant or doctor)?
- **Who is responsible** for the model and the different steps (data collection, training, evaluation etc.)?
- Is the application area of the model prone to any **discrimination or bias**?
- Was the model **tested for bias**, if yes how? Was any **bias detected** and what where the counter measures taken?