Instance-based Method for Post-hoc Interpretability: a Local Approach

Thibault Laugel

LIP6 - Sorbonne Université

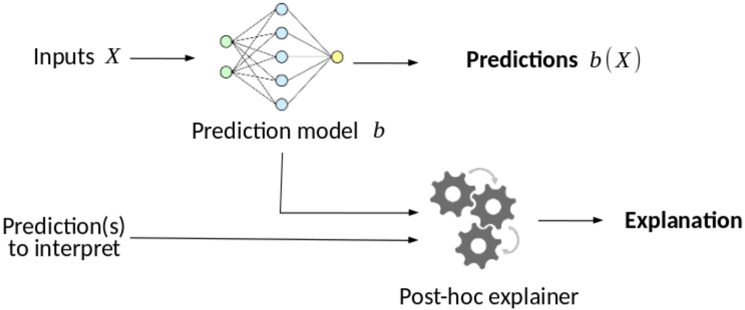
8 October 2018

Workshop on Machine Learning and Explainability

Research supported by the AXA Research Fund

Post-hoc Interpretability

Considered framework



Post-hoc Interpretability

State of the art

Several types of approaches exist in the litterature, such as:

- Sensitivity analysis

  e.g. Baehrens et al. 2010

- Rule extraction

  e.g. Wang et al. 2015, Turner 2016

- Surrogate model approaches

  e.g. Ribeiro et al. 2016 (LIME), Ljundberg et al. 2017 (SHAP)

- Instance-based approaches

  e.g. Kim et al. 2014, Kabra et al. 2015, Wachter et al. 2018

Instance-based Approaches (I)

Context

Principle

Using specific instances as explanations for the predictions of a model

- Arguments for instance-based approaches:
  - **Practical:** Using a 'raw' instance is in some cases better than forcing a specific form of explanation
  - **Legal:** Excessive disclosure of information about the inner workings of an automated system may reveal protected information
  - **Scientific:** Cognitive Sciences approaches relying on teaching through examples

Watson et al. 2008

Instance-based Approaches

State of the art

Different approaches using instances as explanations, such as:

- Prototype-based approaches

  e.g. Kim et al. 2014

- Influential neighbors

  e.g. Kabra et al. 2016

- Counterfactuals

  e.g. Wachter et al. 2018

Related Fields

Inverse Classification

- **Goal: manipulate an instance such that it is more likely to conform to a specific class**
- Several formulations, such as:
  - Find the smallest manipulation required

    Barbella et al. 2009
  - Increase the probability of belonging to another class

    Lash et al. 2016
- Related field: evasion attacks in adversarial learning

  Biggio et al. 2017

Inverse Classification for Interpretability

Problem definition

- **Inputs:**
  - Black-box classifier $b : \mathcal{X} \rightarrow \mathcal{Y} = \{-1, 1\}$
  - $x \in \mathcal{X}, b(x)$ the prediction to interpret
- **Goal:** Find the smallest change to apply to x to change $b(x)$
- With the following assumptions:
  - Feature representation is known
  - b can be used as an oracle to compute new predictions

Final Explanation

Final explanation = 'ennemy' associated to this smallest change

Inverse Classification Problem

Formalization

Proposed minimization problem:

$$e^* = \underset{e \in \mathcal{X}}{argmin}\{c(x, e) : b(e) \neq b(x)\}$$

With c a proposed cost function defined as:

$$c(x, e) = \underbrace{||x - e||_2}_{\textbf{proximity metrics}} + \underbrace{||x - e||_0}_{\textbf{sparsity metrics}}$$

Solving the Problem with *Growing Spheres*

General Idea

- Complex problem:
  - Cost function is discontinuous
  - No information about b
  - b is 'only' returning a class (no confidence score such as probability)

Solving the Problem with *Growing Spheres*

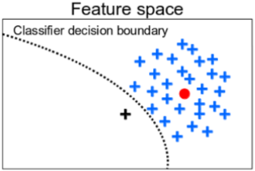
General Idea

- Complex problem:
  - Cost function is discontinuous
  - No information about b
  - b is 'only' returning a class (no confidence score such as probability)
- Proposition: solve sequentially the minimization problem:
  1. l_2 component: **Generation** step
  2. l_0 component: **Feature Selection** step

Solving the Problem with *Growing Spheres*

Implementation

1. **Generation** of instances uniformly in growing hyperspheres centered on x until an ennemy e is found

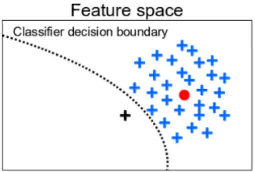


Step 1: Generation

Solving the Problem with *Growing Spheres*

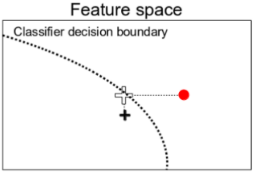
Implementation

1. **Generation** of instances uniformly in growing hyperspheres centered on x until an ennemy e is found



Step 1: Generation

2. **Feature Selection** performed by setting the coordinates of vector $x - e$ to 0 to make the explanation sparse



Step 2: Feature Selection

Possible Personnalization

Depending on the user needs and the prediction task, several elements can be modified, such as:

- The features that are used in the exploration
  - The user might be interested in some specific directions
  - E.g. Marketing model predicting if whether a user will buy a product or not: number of ads sent vs age of the customer
- The cost function used

Illustrative Results

Illustration on the Boston dataset

- **Boston Housing dataset**
- **Binary classification problem**:
  $\mathcal{Y} = \{\textit{expensive}, \textit{not expensive}\}$
  - expensive = median value higher than 26 000$
- **Representation**: 13 attributes.
  - Examples: number of rooms, age of the buildings...
- A **black-box classifier** is trained
  - In this case, a Random Forest algorithm
- We use *Growing Spheres* to **generate explanations** for individual predictions

Experimental Results

Illustration on the Boston dataset

Housing/class	Feature	Move
H1 Not Expensive	Average number of rooms per dwelling	+0.12
	Nitrogen oxides conc. (parts per 10 million)	-0.008
H2 Expensive	Average number of rooms per dwelling	-0.29
	Proportion of non-retail business acres per town	+0.90

Extension and link with surrogates models

- A possible requirement for an explanation could be its **robustness**:
  - Do two close instances have similar explanations?

    Alvarez-Melis et al. 2018
  - How can a local explanation be 'generalized'?
- Local surrogate models aim at approximating the local decision border of a black-box with an *interpretable* model

  Ribeiro et al. 2016 (LIME)

Performance metrics (I)

Proposed measure

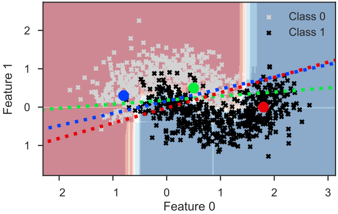
- **Local Fidelity**: measures the surrogate's local accuracy to the black-box model

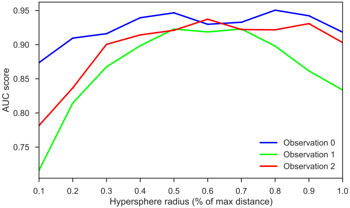
$$LocalFid(x, s_x) = Acc_{x_i \in \mathcal{V}_x}(b(x_i), s_x(x_i))$$

  - How well the surrogate mimics the black-box
  - Neighborhoods $\mathcal{V}_x$ can be modified
    - E.g. Hyperspheres of growing radius
  - A high fidelity in an a given neighborhood $\mathcal{V}_x$ means that the explanation can be generalized in this area

Performance metrics (II)

Measuring the quality of the local approximation



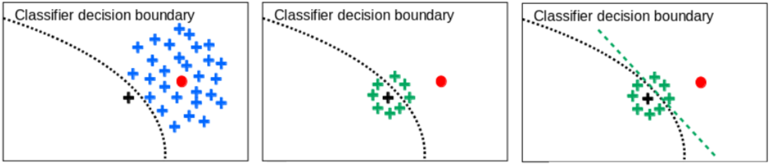


- The Local Fidelity measure captures the local behavior of the surrogate model

Local Surrogate Model (LS)

Principle

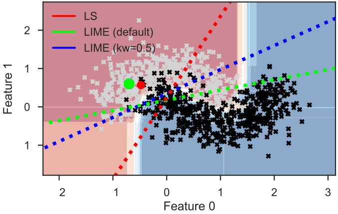
1. **Detection** of the black-box's **closest decision boundary**
2. Local **sampling in this area**
3. Fit of the surrogate
4. Extract explanations

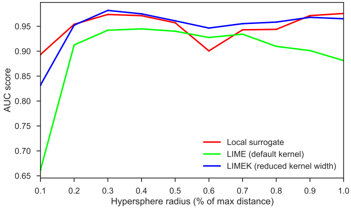


Step 0: Closest border detection **Step 1: Local sampling** **Step 2: Model training**

Preliminary Results (I)

- Experiment setup
  - Competitors: LS, LIME, LIME-K (reduced kernel width)





- LS has more intuitive frontier approximations
- Higher local fidelity for small hypersphere radius

Preliminary Results (II)

- Experiment setup
  - Competitors: LS, LIME, LIME-K (reduced kernel width)
  - Datasets: $^1/_2$-moons, cancer, credit, news, tennis (UCI)
  - Growing local fidelity metric for 5% radius, averaged over test set instances
- Avg. Local Fidelity (AUC): $+8\%$ over LIME ($^1/_2$-moons)
- UCI datasets: LS with $+9\%$ to $+18\%$

Conclusion and Perspectives

- The proposed approaches are:
  1. A post-hoc interpretability method using **instances to generate explanations** when **no information about the classifier nor any data is available**
  2. A surrogate model approach to generate more robust explanations by approximating the local decision border of the black-box
- Ongoing works:
  - Design heuristics for the hyperparameters tuning
  - Work on the notion of robustness
  - Work on explanation validation:
    - Define validation criteria
    - Have experiments with real users and industry experts

Instance-based Method for Post-hoc Interpretability: a Local Approach

Thibault Laugel

LIP6 - Sorbonne Université

8 October 2018

Workshop on Machine Learning and Explainability

Research supported by the AXA Research Fund