

Contrastive model explanations, an overview

Wojciech Sobala, Data Scientist at IBM

Motivation

1. The data subject shall have the **right not to be subject to a decision based solely on automated processing**, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

=> Article: 4

2. Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

=> Dossier: Legitimate Interests (Data Subject), Opening Clause

(c) is based on the data subject's explicit consent.

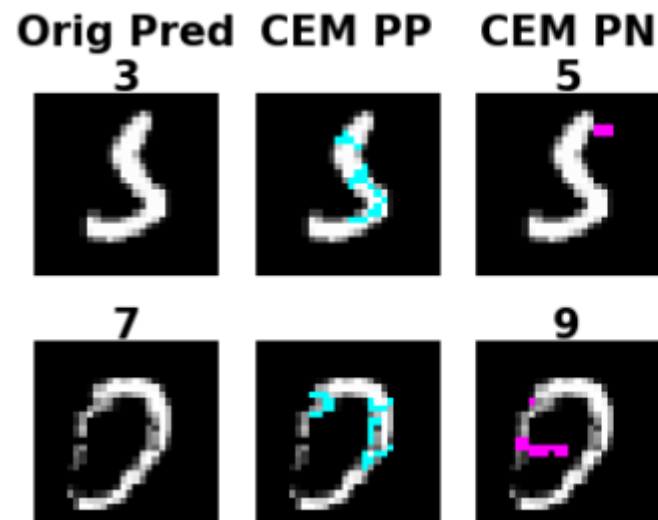
=> Dossier: Consent

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the **right to obtain human intervention on the part of the controller**, to express his or her point of view and **to contest the decision**.

Methods to explain model

- Global
 - Partial Dependence Plots,
 - Feature Importance,
 - Feature Interaction,
 - Accumulated Local Effects,
 - Global Surrogate Models
- Individual predictions
 - Local Surrogate Models,
 - Shapley Value Explanations,
 - Counterfactual Explanations

Contrastive Explanations Method (CEM)



Finding pertinent negatives

$$\min_{\boldsymbol{\delta} \in \mathcal{X}/\mathbf{x}_0} c \cdot f_{\kappa}^{\text{neg}}(\mathbf{x}_0, \boldsymbol{\delta}) + \beta \|\boldsymbol{\delta}\|_1 + \|\boldsymbol{\delta}\|_2^2 + \gamma \|\mathbf{x}_0 + \boldsymbol{\delta} - \text{AE}(\mathbf{x}_0 + \boldsymbol{\delta})\|_2^2.$$

where:

$$f_{\kappa}^{\text{neg}}(\mathbf{x}_0, \boldsymbol{\delta}) = \max\{[\text{Pred}(\mathbf{x}_0 + \boldsymbol{\delta})]_{t_0} - \max_{i \neq t_0} [\text{Pred}(\mathbf{x}_0 + \boldsymbol{\delta})]_i, -\kappa\}$$

$\text{AE}(\mathbf{x})$ the reconstructed example of \mathbf{x}

$$\boldsymbol{\delta} \in \mathcal{X}/\mathbf{x}_0$$

Finding pertinent positives

$$\min_{\boldsymbol{\delta} \in \mathcal{X} \cap \mathbf{x}_0} c \cdot f_{\kappa}^{\text{pos}}(\mathbf{x}_0, \boldsymbol{\delta}) + \beta \|\boldsymbol{\delta}\|_1 + \|\boldsymbol{\delta}\|_2^2 + \gamma \|\boldsymbol{\delta} - \text{AE}(\boldsymbol{\delta})\|_2^2,$$

where:

$$f_{\kappa}^{\text{pos}}(\mathbf{x}_0, \boldsymbol{\delta}) = \max\{\max_{i \neq t_0} [\text{Pred}(\boldsymbol{\delta})]_i - [\text{Pred}(\boldsymbol{\delta})]_{t_0}, -\kappa\}.$$

perturbation $\boldsymbol{\delta} \in \mathcal{X} \cap \mathbf{x}_0$ such that after removing it from \mathbf{x}_0

$$\arg \max_i [\text{Pred}(\mathbf{x}_0)]_i = \arg \max_i [\text{Pred}(\boldsymbol{\delta})]_i$$

Handwritten Digits



Procurement Fraud

Method	PP % Match	PN % Match
CEM	90.3	94.7
LIME	86.6	N/A
LRP	88.2	N/A

Explanations acceptable by experts

Experiment details

As to the implementation of the projected FISTA for finding pertinent negatives and pertinent positives, we set the regularization coefficients $\beta = 0.1$, and $\gamma = \{0, 100\}$. The parameter c is set to 0.1 initially, and is searched for 9 times guided by run-time information. In each search, if f_κ never reaches 0, then in the next search, c is multiplied by 10, otherwise it is averaged with the current value for the next search. For each search in c , we run $I = 1000$ iterations using the SGD solver provided by TensorFlow. The initial learning rate is set to be 0.01 with a square-root decaying step size. The best perturbation among all searches is used as the pertinent positive/negative for the respective optimization problems.

References

Explanations based on the Missing: Towards
Contrastive Explanations with Pertinent
Negatives*

Amit Dhurandhar^{†1}, Pin-Yu Chen^{†1}, Ronny Luss¹, Chun-Chen Tu²,
Paishun Ting², Karthikeyan Shanmugam¹ and Payel Das¹

October 30, 2018

Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, Payel Das


(Submitted on 21 Feb 2018 (v1), last revised 29 Oct 2018 (this version, v2))

In this paper we propose a novel method that provides contrastive explanations justifying the classification of an input by a black box classifier such as a deep neural network. Given an input we find what should be %necessarily and minimally and sufficiently present (viz. important object pixels in an image) to justify its classification and analogously what should be minimally and necessarily \emph{absent} (viz. certain background pixels). We argue that such explanations are natural for humans and are used commonly in domains such as health care and criminology. What is minimally but critically \emph{absent} is an important part of an explanation, which to the best of our knowledge, has not been explicitly identified by current explanation methods that explain predictions of neural networks. We validate our approach on three real datasets obtained from diverse domains; namely, a handwritten digits dataset MNIST, a large procurement fraud dataset and a brain activity strength dataset. In all three cases, we witness the power of our approach in generating precise explanations that are also easy for human experts to understand and evaluate.

Subjects: **Artificial Intelligence (cs.AI)**; Computer Vision and Pattern Recognition (cs.CV); Machine Learning (cs.LG)

Report number: accepted to NIPS 2018

Try the Bibliographic Explorer
(can be disabled at any time)
[Enable](#) [Don't show again](#)

 [cs.AI] for this version)

Bibliographic data

[\[Enable Bibex \(What is Bibex?\)\]](#)

Submission history

From: Amit Dhurandhar [\[view email\]](#)
[v1] Wed, 21 Feb 2018 15:51:38 UTC (539 KB)

[v2] Mon, 29 Oct 2018 16:08:36 UTC (975 KB)

[Which authors of this paper are endorsers?](#) | [Disable MathJax \(What is MathJax?\)](#)

Download:

- [PDF](#)
 - [Other formats](#)
- (license)

Current browse context:

cs.AI
[< prev](#) | [next >](#)
[new](#) | [recent](#) | [1802](#)

Change to browse by:

cs
cs.CV
cs.LG

References & Citations

- [NASA ADS](#)

DBLP - CS Bibliography

[listing](#) | [bibtex](#)
Amit Dhurandhar
Pin-Yu Chen
Ronny Luss
Chun-Chen Tu
Pai-Shun Ting
...

Google Scholar

Bookmark (what is this?)

