

Evaluation approaches for XAI methods

Szymon Bobek

Jagiellonian University
2023



<https://geist.re>

Properties of Explanation Mechanism

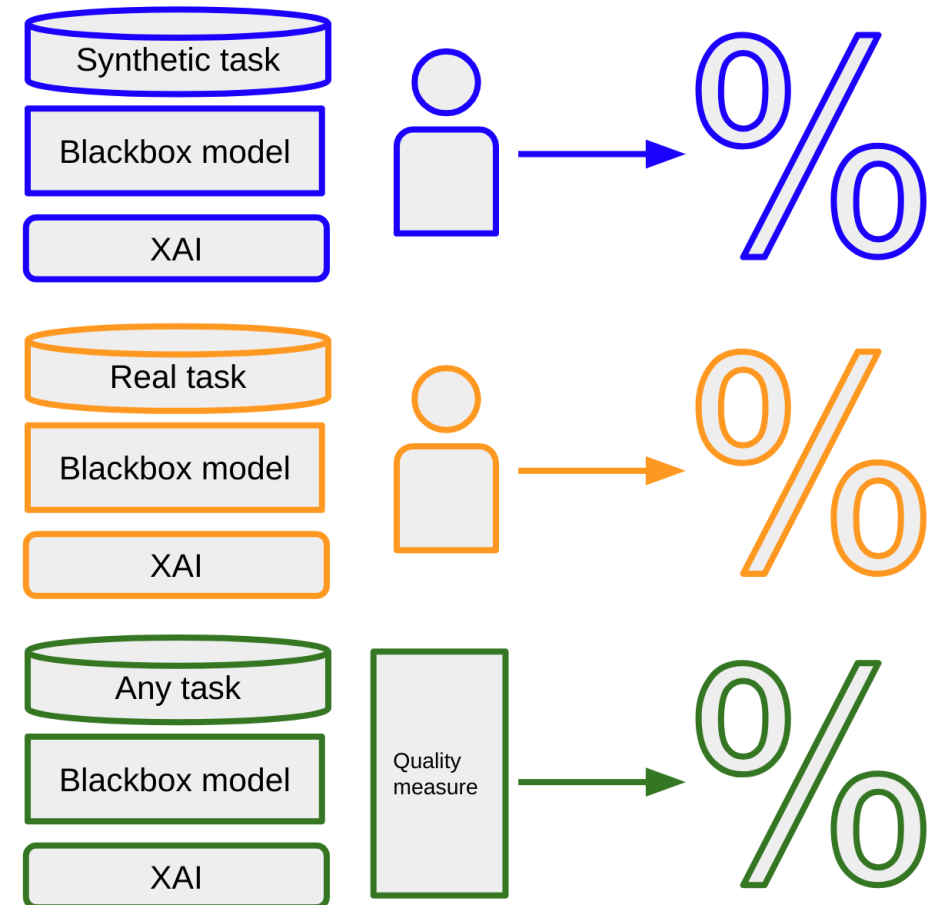
- **Expressive Power** is the “language” or structure of the explanations the method is able to generate. An explanation method could generate IF-THEN rules, decision trees, a weighted sum, natural language or something else.
- **Translucency** describes how much the explanation method relies on looking into the machine learning model, like its parameters. For example, explanation methods relying on intrinsically interpretable models like the linear regression model (model-specific) are highly translucent. Methods only relying on manipulating inputs and observing the predictions have zero translucency. Depending on the scenario, different levels of translucency might be desirable. The advantage of high translucency is that the method can rely on more information to generate explanations. The advantage of low translucency is that the explanation method is more portable.
- **Portability** describes the range of machine learning models with which the explanation method can be used. Methods with a low translucency have a higher portability because they treat the machine learning model as a black box. Surrogate models might be the explanation method with the highest portability. Methods that only work for e.g. recurrent neural networks have low portability.
- **Algorithmic Complexity** describes the computational complexity of the method that generates the explanation. This property is important to consider when computation time is a bottleneck in generating explanations.

Properties and evaluation metrics of explanations

- Types of evaluation approaches
 - **Human-grounded**
 - **Application-grounded**
 - **Functional**
- Popular Quality measures
 - Fidelity (local and global)
 - Stability
 - Consistency
 - Coverage
 - Certainty
 - Representativeness
 - Simplicity/Comprehensibility
 - Degree of Importance
 - Novelty
- Ready to use frameworks
 - Quantus



<https://github.com/understandable-machine-intelligence-lab/Quantus>



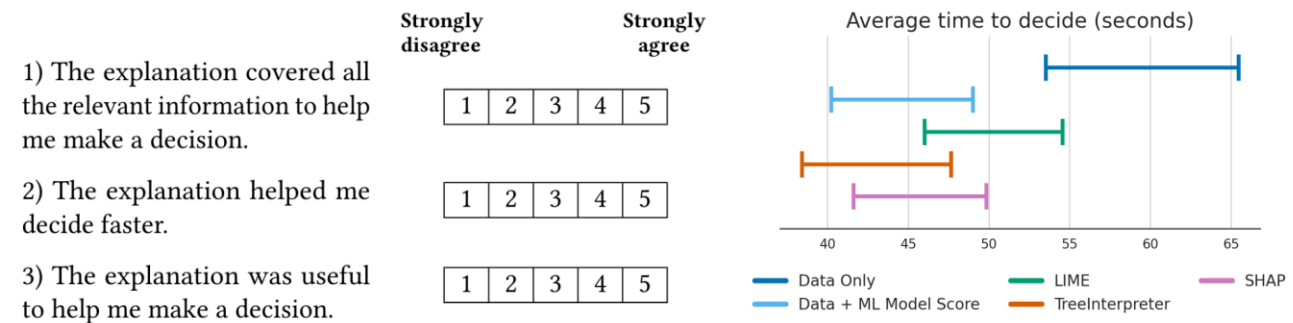


Evaluating factual explanations

Human grounded and task grounded evaluation

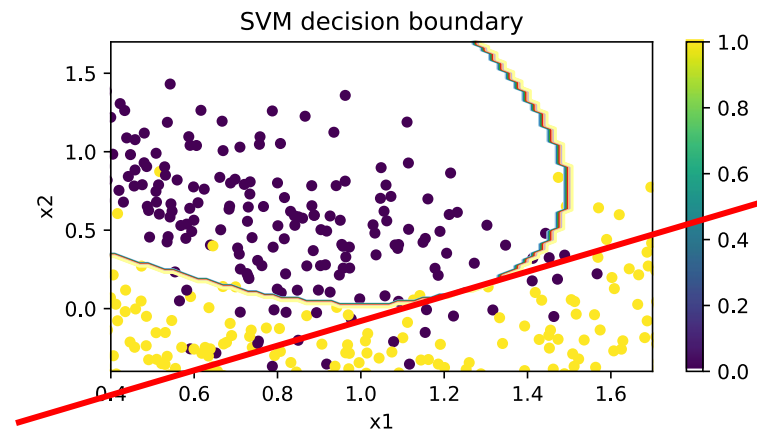
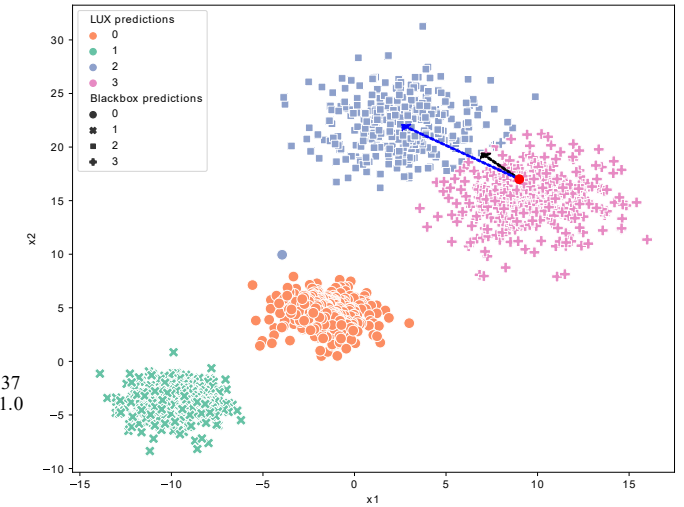
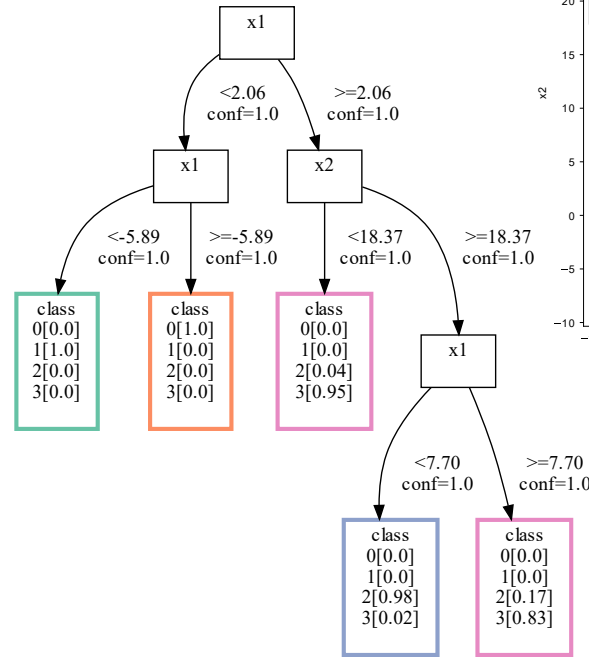
- The specialist has a task to accept/decline fraud warnings that the ML system is not certain about
- We set a set of objective evaluation criteria such as time-to-decision, accuracy
- We can also set of subjective evaluation criteria such as comprehensibility
- We compare the results with statistical tests such as Friedman, or Kruskal-Wallis

Group	Explainer	Sample Size	Metrics				Agreement
			accuracy (%)	recall (%)	FPR (%)	time (s)	
Data Only	-	200	62.00	35.87	15.74	59.50	0.41
Data + ML Model Score	-	200	51.02*	25.00	19.57	44.61*	-0.02
Data + ML Model Score + Explanations	LIME	300	58.59	27.03	10.07	50.29*	0.53
	TreeInterpreter	300	56.52	25.55	12.67	43.03*†	0.30
	SHAP	300	59.73	31.08	12.00	45.72*	0.15



Fidelity (local/global)

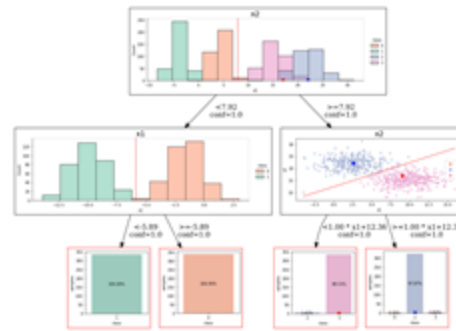
- How well does the explanation approximate the prediction of the black box model?
- High fidelity is one of the most important properties of an explanation, because an explanation with low fidelity is useless to explain the machine learning model.
- Accuracy and fidelity are closely related.
- If the black box model has high accuracy and the explanation has high fidelity, the explanation also has high accuracy.
- Some explanations offer only local fidelity



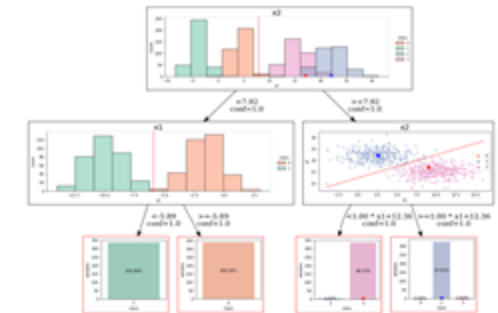
Stability

- How similar are the explanations for similar instances?
- High stability means that slight variations in the features of an instance do not substantially change the explanation (unless these slight variations also strongly change the prediction).
- A lack of stability can be the result of a high variance of the explanation method.
- In other words, the explanation method is strongly affected by slight changes of the feature values of the instance to be explained.
- A lack of stability can also be caused by non-deterministic components of the explanation method, such as a data sampling step, like the local surrogate method uses.
- Is high stability is always desirable?

Instance x_i

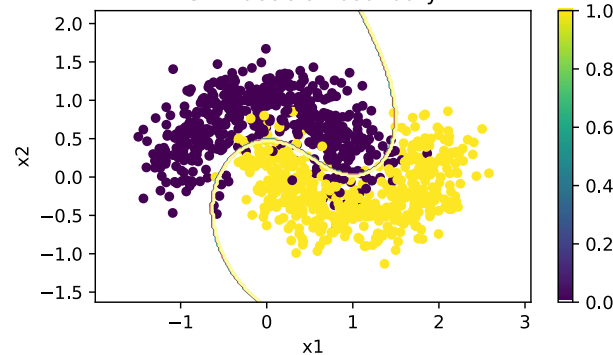


Instance x_j

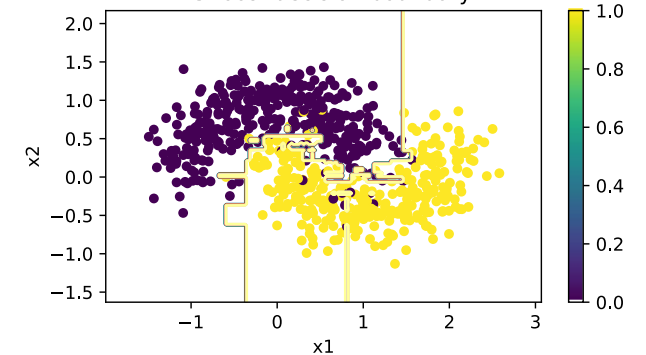


$$\hat{L}(\Phi^{e \rightarrow m}, X) = \max_{x_j \in N_\epsilon(x_i)} \frac{\|x_i - x_j\|_2}{\|\Phi_i^{e \rightarrow m} - \Phi_j^{e \rightarrow m}\|_2 + 1}$$

SVM decision boundary



XGBoost decision boundary



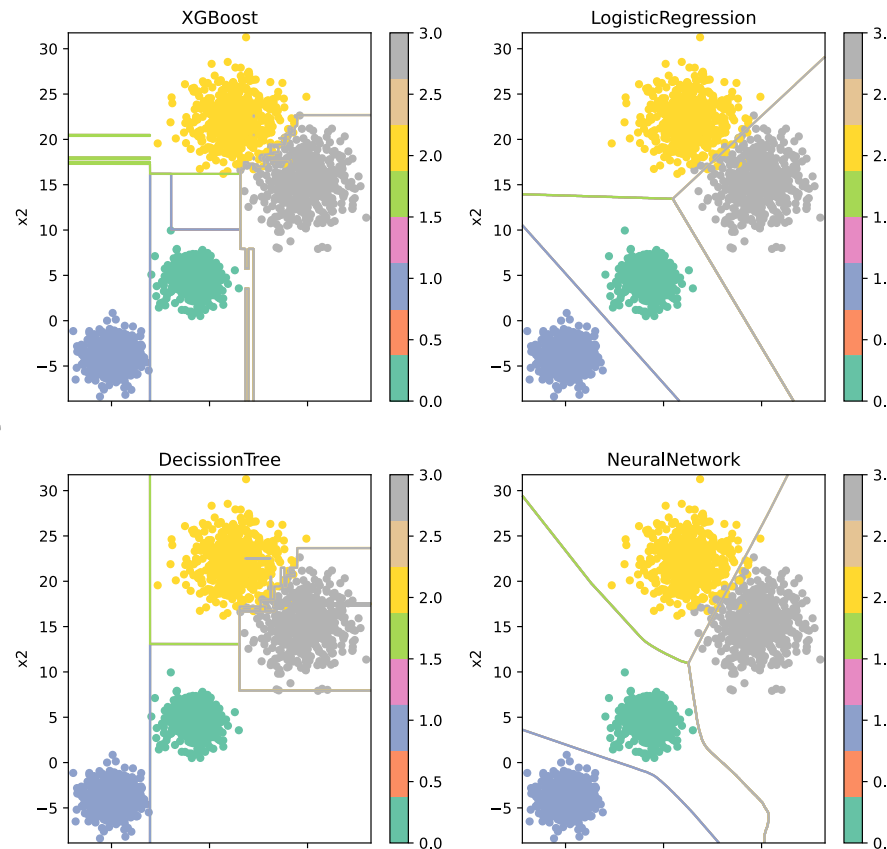
Consistency

- How much does an explanation differ between models that have been trained on the same task and that produce similar predictions?
- How much does an explanation differ between consecutive calls of the XAI model for the same instance
- How much does the explanations of different XAI methods differ for the same ML model?

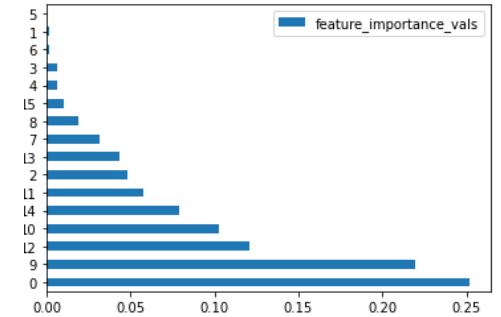
```
1: {'f_6': ['<=0.051216'], 'f_1': ['>-0.118408']}
2: {'f_6': ['<=0.034339'], 'f_3': ['<=0.231442'],
    'f_1': ['>-0.387023']}
3: {'f_6': ['<=-0.156712'], 'f_3': ['<=0.457593'],
    'f_1': ['>-0.619056']}
4: {'f_1': ['>-0.052240'], 'f_6': ['<=-0.101768']}
5: {'f_6': ['<=0.075657'], 'f_7': ['<=1.201282']}
```

```
1: {'f_6': ['<=-1.20169']}
2: {'f_6': ['<=0.152367'], 'f_3': ['<=1.206294']}
3: {'f_6': ['<=0.358263'], 'f_1': ['>-0.691120']}
4: {'f_6': ['<=-0.852889']}
5: {'f_6': ['<=-1.322858']}
```

EXPLAN



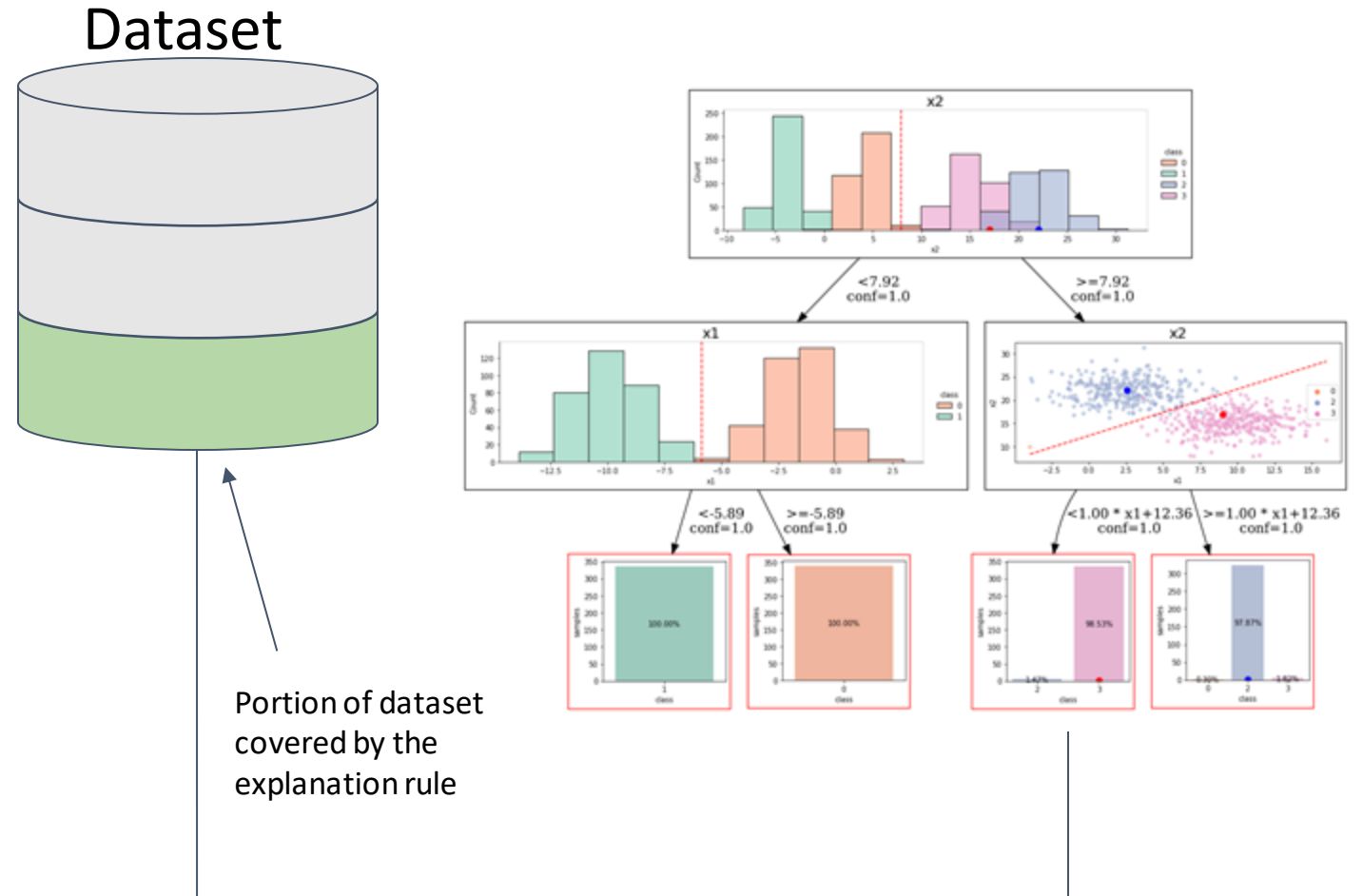
LORE



```
IF
  f_0 <-0.10939577221870422
AND
  f_14 >=-
0.5869548618793488 AND
  f_9 <0.1189308911561966
THEN
  class = 0 # 1.0
```

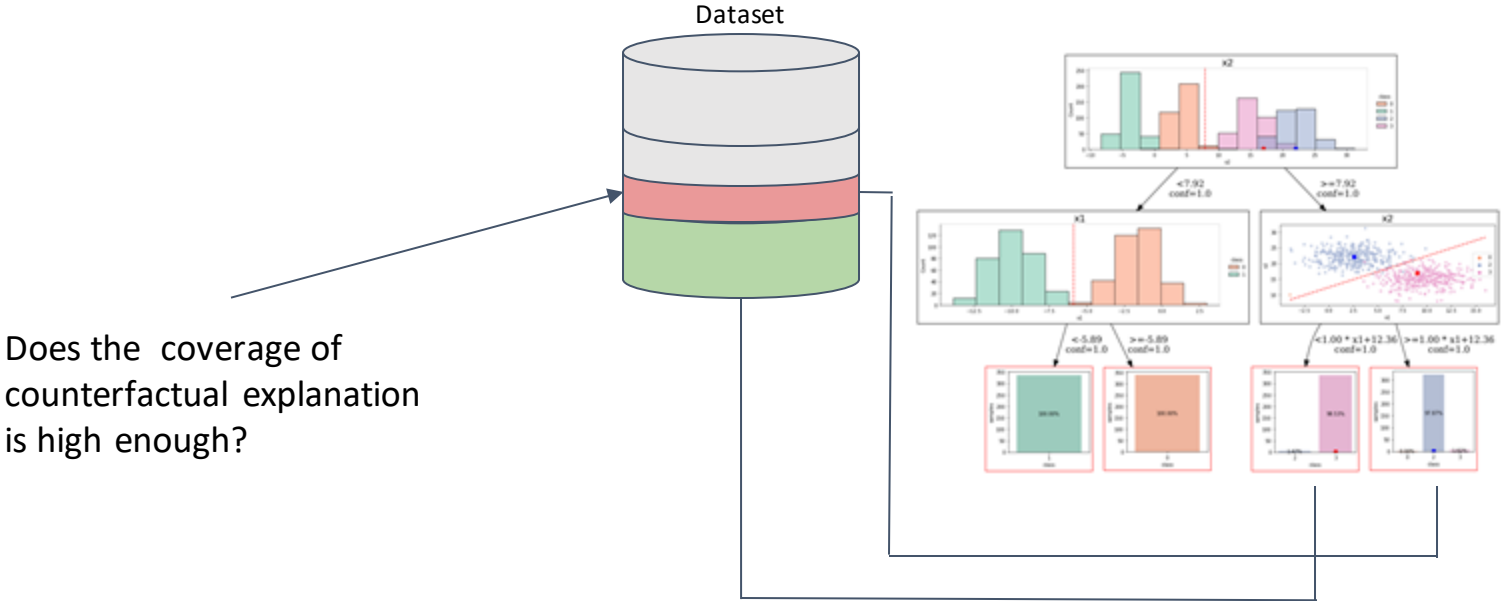
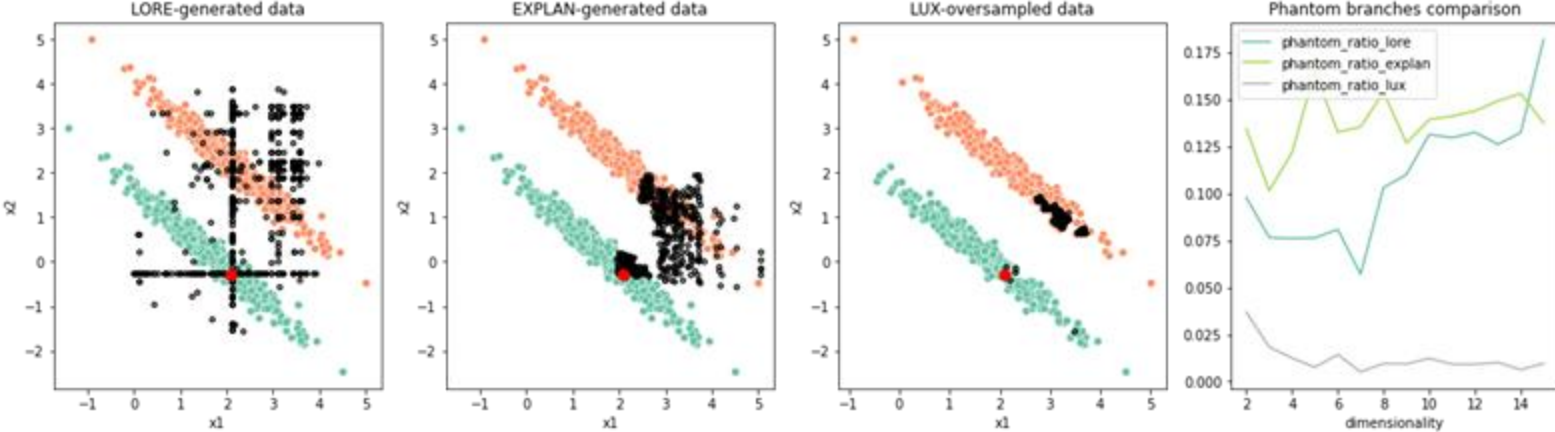

Coverage

- How general is the explanation?
- How many of the samples from a given set it can be triggered for
- Some explanations apply only for single instance (SHAP), but some are more robust (e.g. rule-based explanations)



Representativeness

- In case of example-based explanations, are the examples representative samples?
- In many situations XAI methods rely on generated data, which may produce unrealistic samples
- It is similar to Plausability metric from counterfactual's area



Portion of dataset covered by the explanation rule

Certainty

- Does the explanation reflect the certainty of the machine learning model?
- Many machine learning models only give predictions without a statement about the models confidence that the prediction is correct
- If the model predicts a 4% probability of cancer for one patient, is it as certain as the 4% probability that another patient, with different feature values, received?
- An explanation that includes the model's certainty is very useful
- How about the explanation itself?

Prediction probabilities



atheism

christian



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
 Subject: Another request for Darwin Fish
 Organization: University of New Mexico, Albuquerque
 Lines: 11

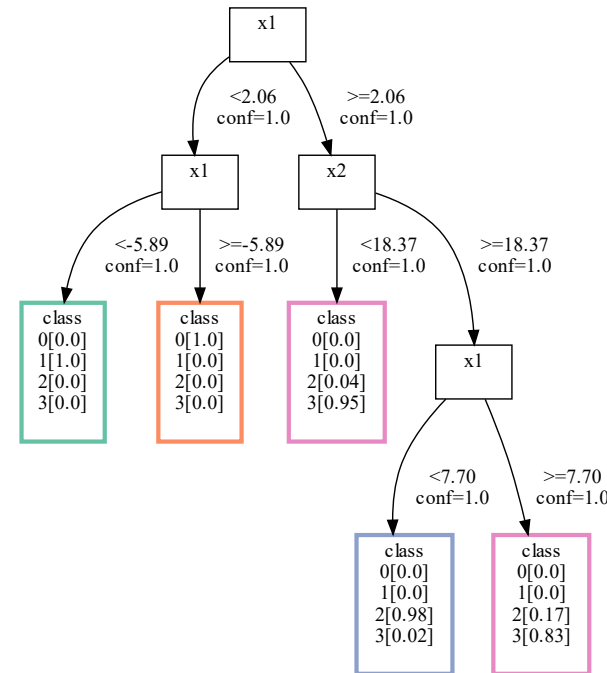
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.

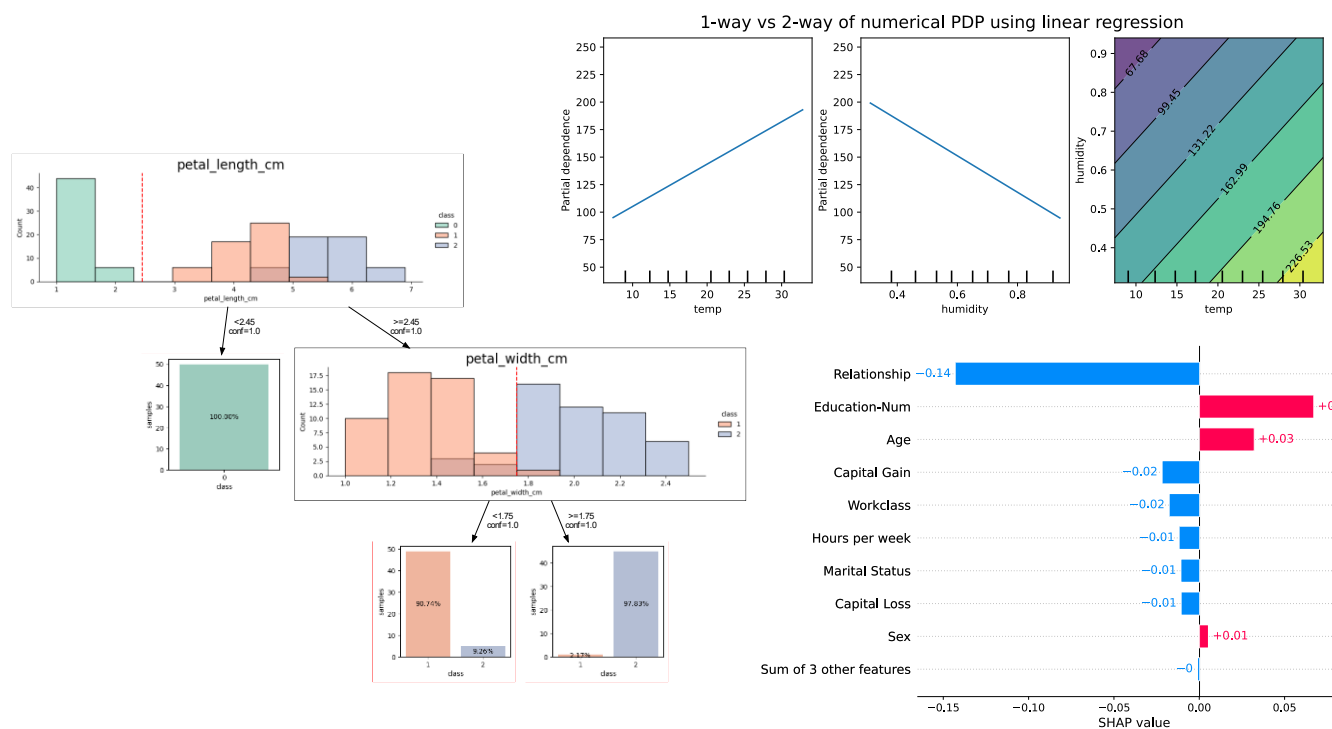
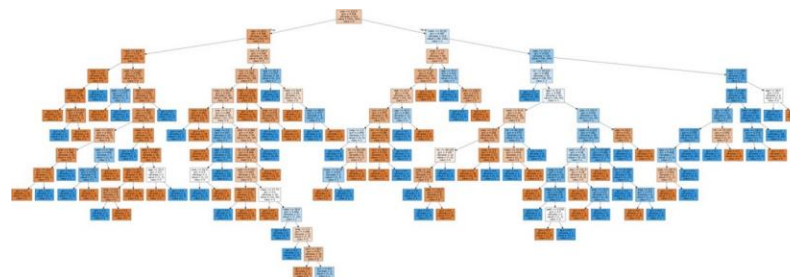
This is the same question I **have** and I **have** not seen an answer on the

net. If anyone has a contact please post on the net or email me.



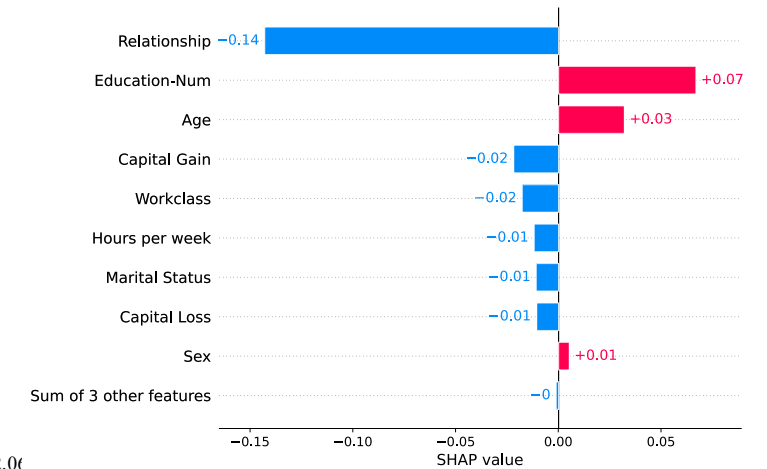
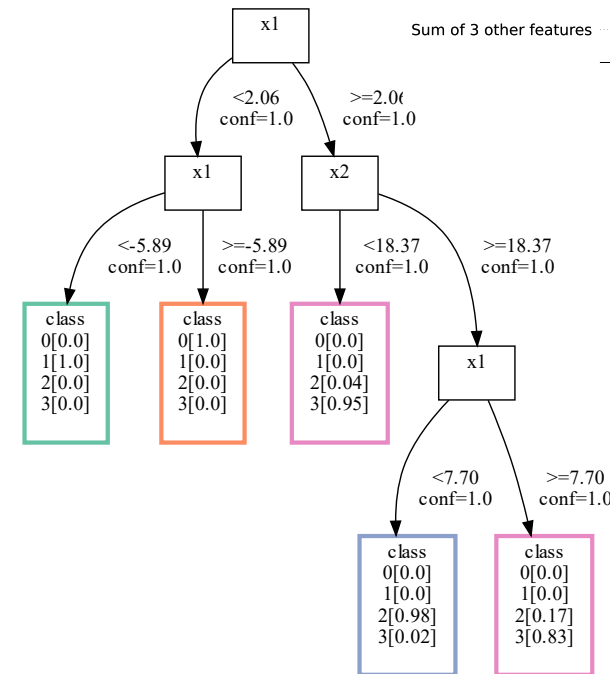
Simplicity/Comprehensability

- How well do humans understand the explanations?
- It is difficult to define and measure, but extremely important to get right.
- Many people agree that comprehensibility depends on the audience.
- Ideas for measuring comprehensibility include measuring the size of the explanation or testing how well people can predict the behavior of the machine learning model from the explanations.
- The comprehensibility of the features used in the explanation should also be considered. A complex transformation of features might be less comprehensible than the original features.



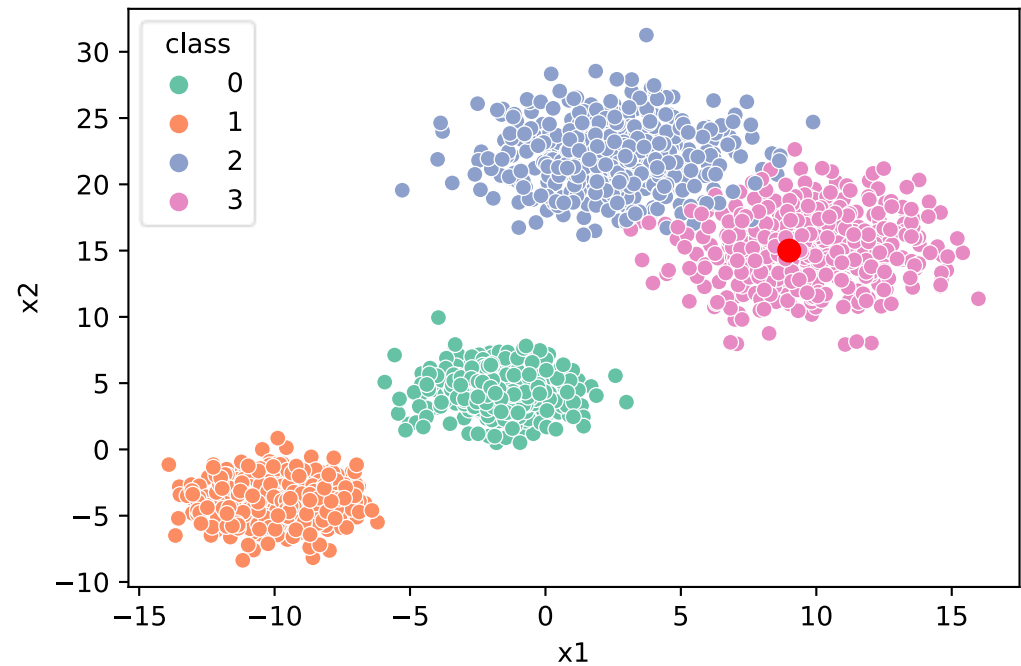
Degree of importance

- How well does the explanation reflect the importance of features or parts of the explanation?
- For example, if a decision rule is generated as an explanation for an individual prediction, is it clear which of the conditions of the rule was the most important?



Novelty

- Does the explanation reflect whether a data instance to be explained comes from a “new” region far removed from the distribution of training data?
- In such cases, the model may be inaccurate and the explanation may be useless.
- The concept of novelty is related to the concept of certainty.
- The higher the novelty, the more likely it is that the model will have low certainty due to lack of data.





Evaluating Counterfactuals

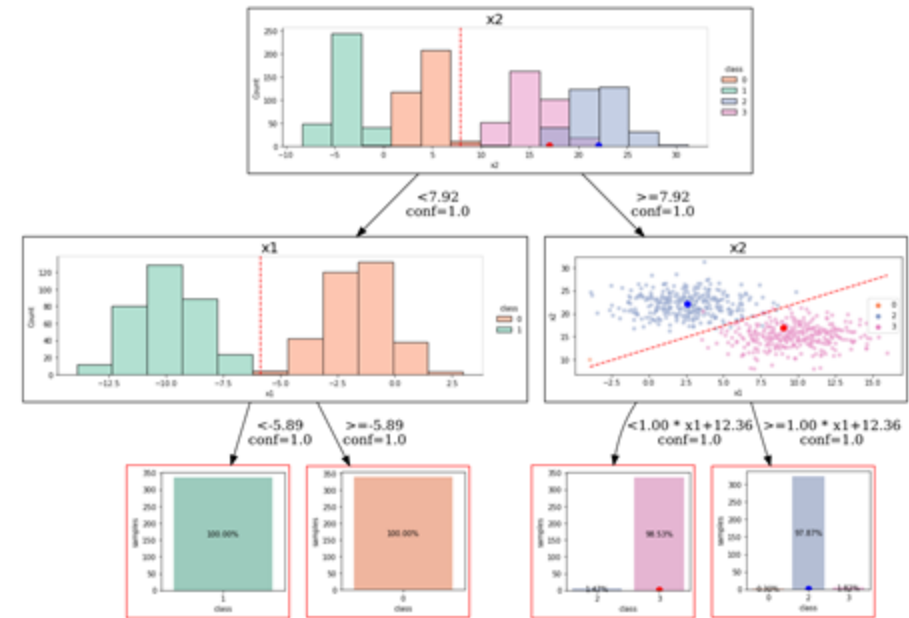
Size

- It measures the number of available counterfactuals.
- Indeed, the number of counterfactuals $|C|$ can be lower than k
- Therefore, size can be defined as

$$size = |C|/k$$

- The higher, the better

$$C = f_k(x, b, X)$$



Dissimilarity

- It measures the proximity between x and the counterfactuals in C . The lower the better.
- We measure it in two fashions

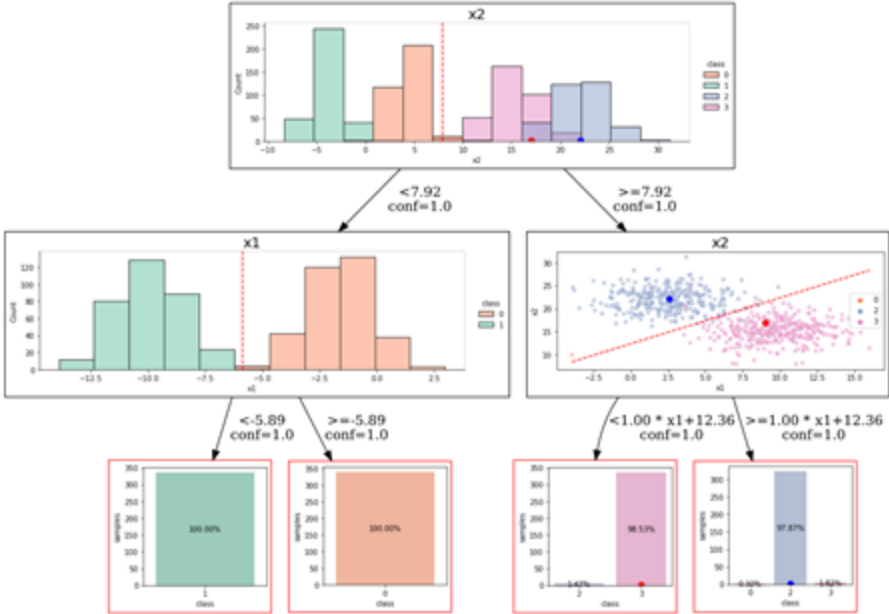
$$dis_{dist} = \frac{1}{|C|} \sum_{x' \in C} d(x, x')$$

Average distance between x and x'

$$dis_{count} = \frac{1}{|C|m} \sum_{x' \in C} \sum_{i=1}^m \mathbb{1}_{x'_i \neq x_i}$$

Average number of features changed

$$C = f_k(x, b, X)$$



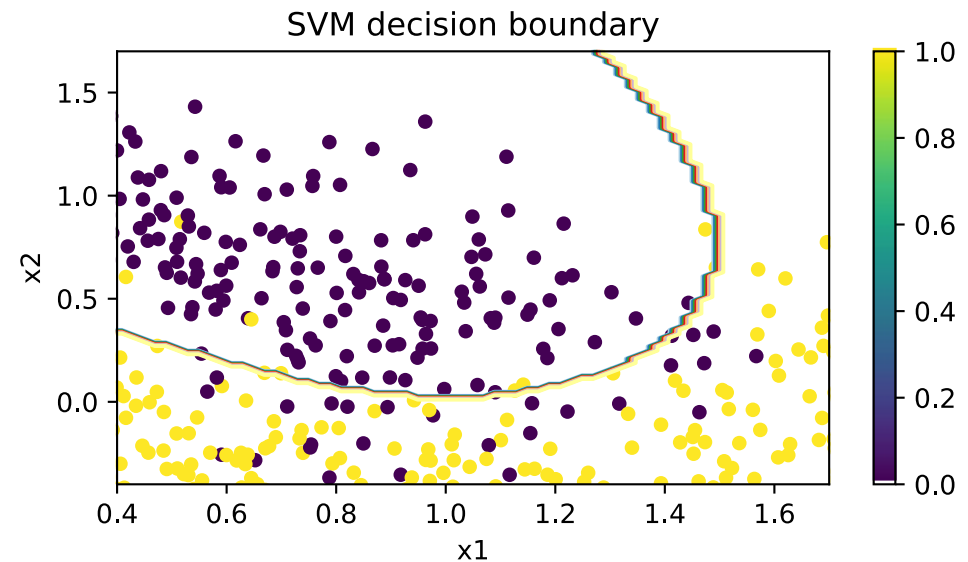
Diversity

- It accounts for a diverse set of counterfactuals, where different actions can be taken to change the decision
- Diversity can be measure either in feature-value domain or features domain
- The higher the better

$$div_{dist} = \frac{1}{|C|^2} \sum_{x' \in C} \sum_{x'' \in C} d(x', x'') \quad div_{count} = \frac{1}{|C|^2 m} \sum_{x' \in C} \sum_{x'' \in C} \sum_{i=1}^m \mathbb{1}_{x'_i \neq x''_i}$$

Average distance
between counterfactuals

Average number of different
features in counterfactuals



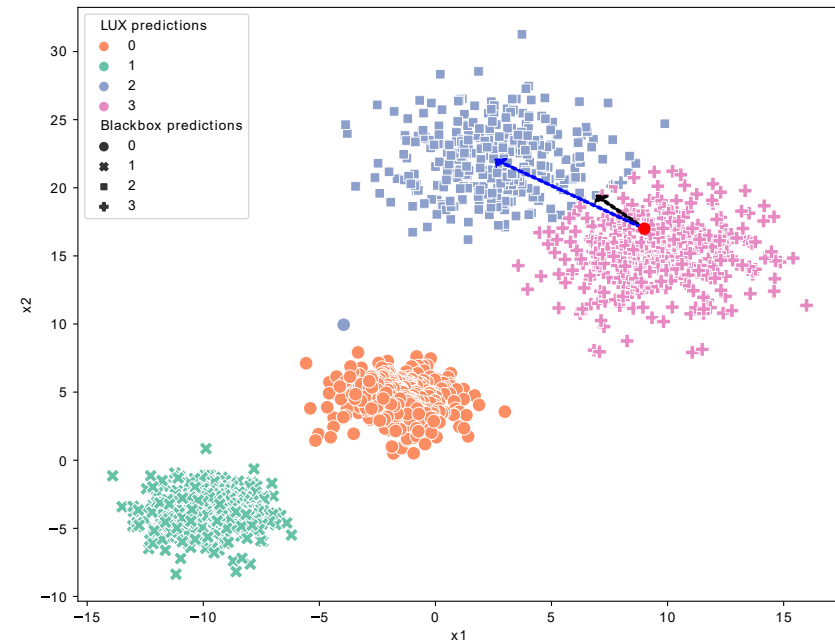
Instability

- It measures to which extent the counterfactuals C obtained for x are close to the counterfactuals \bar{C} obtained for \bar{x} . Here $x \in X$ and \bar{x} is the closest instance to x and \bar{C} receives the same black-box decision of x , i.e.

$$b(x) = b(\bar{x})$$

- The rationale is that similar instances get similar explanations
- The lower, the better

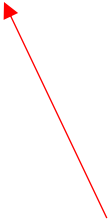
$$inst_{x,\bar{x}} = \frac{1}{1 + d(x, \bar{x})} \frac{1}{|C| |\bar{C}|} \sum_{x' \in C} \sum_{x'' \in \bar{C}} d(x', x'')$$



Actionability

- It measures the level of actionability of the counterfactuals C and accounts for the counterfactuals in C that can be realized
- The higher, the better
- There are frameworks from other fields that are trying to be adopted to explanations (e.g. PETAL for The Patient Education Materials Assessment Tool)

$$act = |\{x' \in C \mid a_A(x', x)\}|/k$$

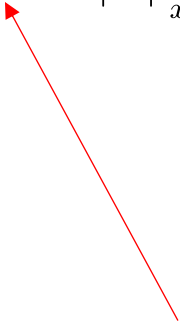


Is true, when x' is actionable with respect to list of features (e.g. cannot change age)

Implausability

- It measures the level of plausibility of the counterfactuals C .
- It accounts for how close are counterfactuals to the reference population X .
- It is the average distance of x'
- from the closest instance in the known set X .
- The lower the better.

$$impl = \frac{1}{|C|} \sum_{x' \in C} \min_{x \in X} d(x', x)$$



This is very simple metric, that does not account for the combination of unrealistic (yet implausible) features values (e.g. sex=male, pregnant=True)

Discriminative Power

- It measures the ability to distinguish through a naive approach between two different classes only using the counterfactuals in C .
- If 1NN is able to perfectly distinguish between class, then human will to.
- The higher, the better

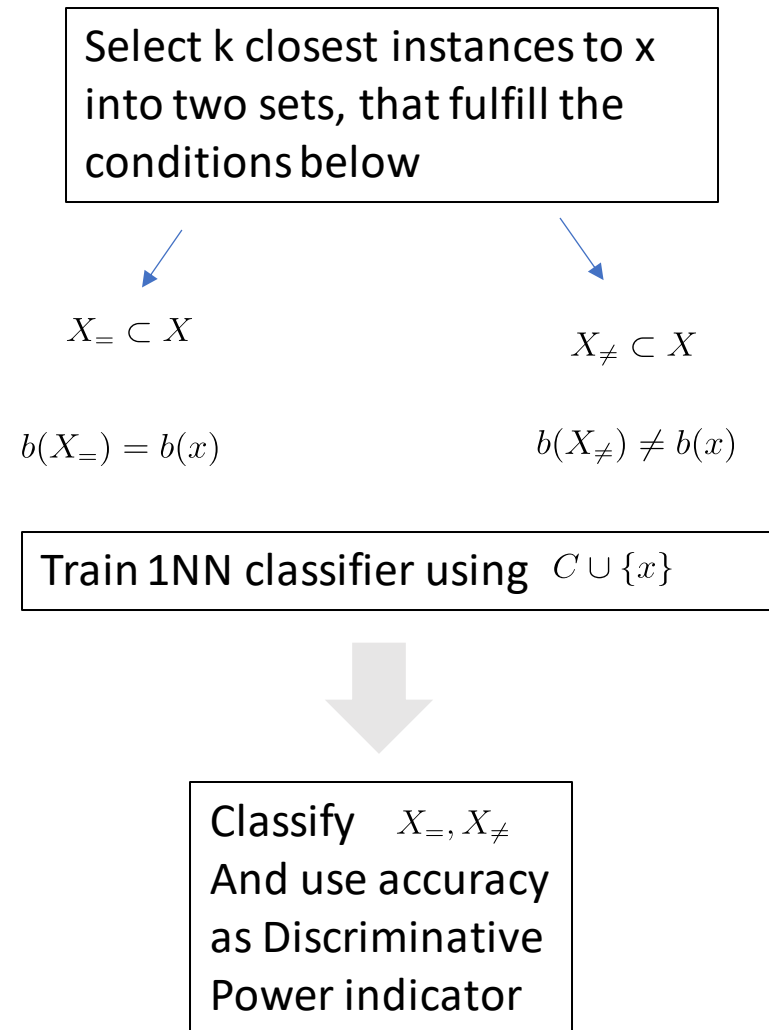




Image modality

Similar metrics, different interpretation

- **Faithfulness** quantifies to what extent explanations follow the predictive behaviour of the model (asserting that more important features play a larger role in model outcomes)
- **Robustness** measures to what extent explanations are stable when subject to slight perturbations of the input, assuming that model output approximately stayed the same
- **Localisation** tests if the explainable evidence is centred around a region of interest (RoI) which may be defined around an object by a bounding box, a segmentation mask or, a cell within a grid
- **Complexity** captures to what extent explanations are concise i.e., that few features are used to explain a model prediction
- **Randomisation** tests to what extent explanations deteriorate as inputs to the evaluation problem e.g., model parameters are increasingly randomised

Thank you for your attention!



JAGIELLONIAN UNIVERSITY
IN KRAKÓW



<https://geist.re>