# Interpretable Maching Learning

## Chapter 02 Explanation

Hao ZHAN

haozhan1993@gmail.com

2020.5

# Table of Contents

# Table of Contents

# 1.Evaluation of Interpretability

Doshi-Velez and Kim (2017) propose three main levels for the evaluation of interpretability:

**Application level evaluation (real task)**

Put the explanation into the product and have it tested by the end user.

**Human level evaluation (simple task)**

The difference is that these experiments are not carried out with the domain experts, but with **laypersons**.

**Function level evaluation (proxy task)**

Does not require humans.

# Table of Contents

# 2.Properties of Explanations

An explanation usually relates the feature values of an instance to its model prediction in a humanly understandable way.
These properties can be used to judge how good an explanation method or explanation is (Robnik-Sikonja and Bohanec, 2018).

## (1) Properties of Explanation Methods

- **Expressive Power**

Expressive Power is the "language" or structure of the explanations the method is able to generate.

- **Translucency**

Translucency describes how much the explanation method relies on looking into the machine learning model, like its parameters.

# 2.Properties of Explanations

## (1) Properties of Explanation Methods

- **Portability**

Portability describes the range of machine learning models with which the explanation method can be used.

- **Algorithmic Complexity**

Algorithmic Complexity describes the computational complexity of the method that generates the explanation.

# 2.Properties of Explanations

## (2) Properties of Individual Explanations

- **Accuracy**

How well does an explanation predict unseen data?

- **Fidelity**

How well does the explanation approximate the prediction of the black box model?

- **Consistency**

How much does an explanation differ between models that have been trained on the same task and that produce similar predictions?

# 2.Properties of Explanations

## (2) Properties of Individual Explanations

- **Stability**

How similar are the explanations for similar instances?

- **Comprehensibility**

How well do humans understand the explanations?

- **Certainty**

Does the explanation reflect the certainty of the machine learning model?

# 2.Properties of Explanations

## (2) Properties of Individual Explanations

- **Degree of Importance**

How well does the explanation reflect the importance of features or parts of the explanation?

- **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?

- **Representativeness**

How many instances does an explanation cover?

# 2.Properties of Explanations

Properties of Explanation Methods:

- Expressive Power
- Translucency
- Portability
- AlgorithmicComplexity

# 2.Properties of Explanations

Properties of Individual Explanations:

- Accuracy
- Fidelity
- Consistency
- Stability
- Comprehensibility
- Certainty
- Degree of Importance
- Novelty
- Representativeness

# Table of Contents

# 3.Human-friendly Explanations

What‘s the "good" explantion? Humanities research can help us find out.

Miller (2017) has conducted a huge survey of publications on explanations.

The term **"explanationext"** refers to the social and cognitive process of explaining, but also to the product of these processes. The explainer can be a human being or a machine.

# 3.Human-friendly Explanations

**What Is a Good Explanation?**

## (1) Explanations are contrastive (Lipton,1990)

Humans usually do not ask why a certain prediction was made, but why this prediction was made instead of another prediction. People tend to think in counterfactual cases.

Humans do not want a complete explanation for a prediction, but want to compare what the differences were to another instance's prediction.

## (2) Explanations are selected

People are used to selecting one or two causes from a variety of possible causes as **THE** explanation.

Make the explanation very short, give only 1 to 3 reasons, even if the world is more complex.

# 3.Human-friendly Explanations

## (3) Explanations are social

Explanations are part of a conversation or interaction between the explainer and the receiver of the explanation.

Pay attention to the social environment of your machine learning application and the target audience.

## (4) Explanations focus on the abnormal

People focus more on abnormal causes to explain events (Kahnemann and Tversky, 1981). They consider these kinds of "abnormal" causes as good expla- nations.

Case: **Student's Fault** or **Teacher's Fault**?

The abnormal features should be included in an explanation.

# 3.Human-friendly Explanations

## (5) Explanations are truthful

Good explanations prove to be true in reality (i.e. in other situations). But disturbingly, this is not the most important factor for a "good" explanation.

The explanation should predict the event as truthfully as possible, which in machine learning is sometimes called fidelity.

## (6) Good explanations are consistent with prior beliefs of the explainee

Humans tend to ignore information that is inconsistent with their prior beliefs. This effect is called confirmation bias (Nickerson 1998)

Good explanations are consistent with prior beliefs. This is difficult to integrate into machine learning and would probably drastically compromise predictive performance.

# 3.Human-friendly Explanations

## (7) Good explanations are general and probable

A cause that can explain many events is very general and could be considered a good explanation.

Generality can easily be measured by the feature's support, which is the number of instances to which the explanation applies divided by the total number of instances.

# 3.Human-friendly Explanations

What Is a Good Explanation:

- Explanations are contrastive
- Explanations are selected
- Explanations are social
- Explanations focus on the abnormal
- Explanations are truthful
- Good explanations are consistent with prior beliefs of the explainee
- Good explanations are general and probable

# Refrence

Robnik-Sikonja, Marko, and Marko Bohanec. “Perturbation-Based Explanations of Prediction Models.” Human and Machine Learning. Springer, Cham, 2018. 159-175.

Lipton, Peter. “Contrastive Explanation.” Royal Institute of Philosophy Supplements 27 (1990): 247-266.

Kahneman, Daniel, and Amos Tversky. 1981. “The Simulation Heuristic.” STANFORD UNIV CA DEPT OF PSYCHOLOGY.

Nickerson, Raymond S. 1998. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” Review of General Psychology 2 (2). Educational Publishing Foundation: 175.

Thank you for your time!