Interpretable Machine Learning

Hendra Bunyamin Maranatha Christian University September 30, 2023

Table of Content

1. What is Interpretability?
2. Metode *Model-Agnostic*
3. Metode *Global Model-Agnostic*
4. Partial Dependence Plot

Prerequisites

- Model *machine learning* seperti linear regression.
- Statistika $\Rightarrow$ peluang bersyarat & distribusi marginal.

Github repository

The repository

Table of Content

1. What is Interpretability?
2. Metode *Model-Agnostic*
3. Metode *Global Model-Agnostic*
4. Partial Dependence Plot

What is Interpretability?

- Interpretability is the degree to which a human can understand the cause of a decision (Miller, 2019).

- Interpretability is the degree to which a human can consistently predict the model’s result (Kim et al., 2016).

The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made.

- Christoph Molnar (a statistician, a machine learner)

A model is better interpretable than another model if its decisions are easier for a human to comprehend than decisions from the other model.

- Christoph Molnar

Taksonomi Teknik Interpretability

- Berbagai taksonomi teknik *Interpretability* dapat dibaca di Molnar (2022).

- Fokus kepada taksonomi berdasarkan **model-specific** atau **model-agnostic**?

Teknik Interpretasi Model yang Spesifik (*Not Limited*)†

Algorithm	**Linear**	**Interaction**	**Task**
Linear regression	✓	✗	regr
Logistic regression	✗	✗	class
Decision trees	✗	✓	class, regr
RuleFit	✓	✓	class, regr
Naïve-Bayes	✗	✗	class
k-nearest neighbors	✗	✗	class, regr

Table of Content

1. What is Interpretability?
2. Metode *Model-Agnostic*
3. Metode *Global Model-Agnostic*
4. Partial Dependence Plot

Metode Model-Agnostic

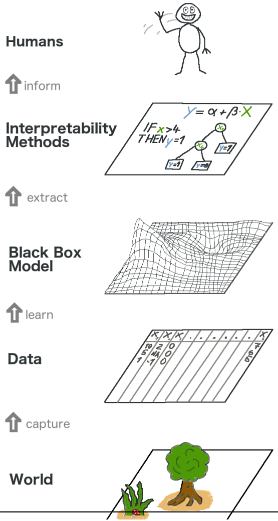
- Memisahkan penjelasan dari model machine learning mempunyai beberapa keuntungan (Ribeiro et al., 2016).
- Keuntungan terbesar metode ini adalah **fleksibilitas**nya.
- Pengembang model machine learning bebas menggunakan model machine learning apa saja.

Aspek yang Diinginkan[†]

Aspek yang diinginkan dari penjelasan model-agnostic (Ribeiro et al., 2016) adalah

- Model flexibility,
- Explanation flexibility, and
- Representation flexibility.

High Level Look (Molnar, 2022)[†]



Humans

inform

Interpretability Methods

extract

Black Box Model

learn

Data

capture

World

Table of Content

1. What is Interpretability?

2. Metode *Model-Agnostic*

3. Metode *Global Model-Agnostic*

4. Partial Dependence Plot

Global Methods

Global Methods

- Global methods menjelaskan **the average behavior** of a machine learning model.

Global Methods

- Global methods menjelaskan **the average behavior** of a machine learning model.
- Global methods $\approx$ **expected values** based on the distribution of the data.

Global Methods

- Global methods menjelaskan **the average behavior** of a machine learning model.
- Global methods $\approx$ **expected values** based on the distribution of the data.
- Contoh: $\hat{f}(x_1, x_2, x_3) =$ fungsi prediksi dengan 3 fitur.
  Untuk melihat efek x_1 pada fungsi prediksi, maka

$$\hat{g}(x_1) = \sum_{x_2} \sum_{x_3} \hat{f}(x_1, x_2, x_3).$$

Table of Content

1. What is Interpretability?

2. Metode *Model-Agnostic*

3. Metode *Global Model-Agnostic*

4. Partial Dependence Plot

Partial Dependence Plot (PDP)

- PDP menunjukkan **efek marginal satu atau dua fitur** pada hasil prediksi sebuah model machine learning (Friedman, 2001).

References I

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*.