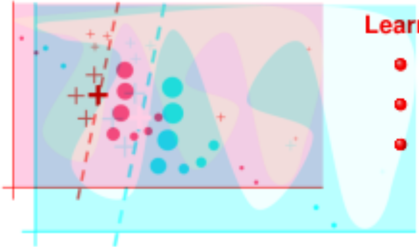
Interpretable Machine Learning

Introduction to Local Explanations



Learning goals

- Understand motivation for local explanations
- Develop an intuition for possible use-cases
- Know characteristics of local explanation methods

METHODOLOGICAL MOTIVATION

- Purpose of local explanations:
  - Insight into the driving factors for a **particular decision**
  - Understand the ML model's decisions in a **local neighborhood** of a given input
    (e.g., feature vector)

Interpretable Machine Learning – 1 / 9

METHODOLOGICAL MOTIVATION

- Purpose of local explanations:
  - Insight into the driving factors for a **particular decision**
  - Understand the ML model's decisions in a **local neighborhood** of a given input (e.g., feature vector)
- Local Methods can address questions such as:
  - **Why** did the model decide to predict $\hat{y}$ for input $\mathbf{x}$?
  - **How** does the model decide for observations that are similar to $\mathbf{x}$?
  - **What** would the ML model have decided if $\mathbf{x}$ its values in $\mathcal{X}$ were different?
  - **Where** (in which regions in $\mathcal{X}$) does the model fail?

SOCIAL MOTIVATION

- Explanations for laypersons must be tailored to the **explainee** (who receives the explanation)
  ⇝ **case specific**, **easy** for humans to understand, and **faithful** to the explained mechanism

SOCIAL MOTIVATION

- Explanations for laypersons must be tailored to the **explainee** (who receives the explanation)
  ⇝ **case specific**, **easy** for humans to understand, and **faithful** to the explained mechanism

Interpretable Machine Learning – 2 / 9

SOCIAL MOTIVATION

- Explanations for laypersons must be tailored to the **explainee** (who receives the explanation)
  $\rightsquigarrow$ **case specific**, **easy** for humans to understand, and **faithful** to the explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations

SOCIAL MOTIVATION

- Explanations for laypersons must be tailored to the **explainee** (who receives the explanation)
  $\rightsquigarrow$ **case specific**, **easy** for humans to understand, and **faithful** to the explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations

Interpretable Machine Learning – 2 / 9

SOCIAL MOTIVATION

- Explanations for laypersons must be tailored to the **explainee** (who receives the explanation)
  $\rightsquigarrow$ **case specific**, **easy** for humans to understand, and **faithful** to the explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations
- Local explanations cannot only increase **user trust**, but also help to detect **critical local biases** in algorithmic decision making

- Explanations for laypersons must be tailored to the **explainee** (who receives the explanation)
  $\rightsquigarrow$ **case specific**, **easy** for humans to understand, and **faithful** to the explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations
- Local explanations cannot only increase **user trust**, but also help to detect **critical local biases** in algorithmic decision making

Interpretable Machine Learning – 2 / 9

SOCIAL MOTIVATION

- Explanations for laypersons must be tailored to the **explainee** (who receives the explanation)
  - ⇝ **case specific**, **easy** for humans to understand, and **faithful** to the explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations
- Local explanations cannot only increase **user trust**, but also help to detect **critical local biases** in algorithmic decision making
- European citizens have the legally binding **right to explanation** as given in the General Data Protection Regulation (GDPR)
  - ⇝ Instead of explaining the entire (complex) model (with potential market secrets), explanations in a case-by-case usage is more reasonable

Interpretable Machine Learning – 2 / 9

GDPR: THE RIGHT TO EXPLANATION

"The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.

[. . .]

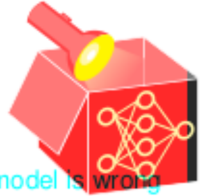
In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the **right** to obtain human intervention, to express his or her point of view, **to obtain an explanation of the decision reached after such assessment and to challenge the decision**. "

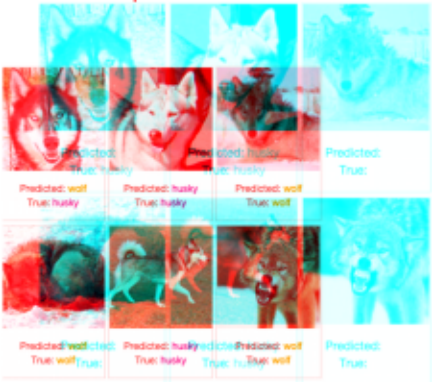
(Recital 71, GDPR)

Interpretable Machine Learning – 3 / 9

EXAMPLE: HUSKY OR WOLF?

- We trained a model to predict if an image shows a wolf or a husky
- Below the predictions on six test images are given
- Do you trust our predictor?





- Sometimes the ML model is wrong
- Can you guess the pattern the ML model learned to identify a wolf?

Source: [Sameer Singh 2018]

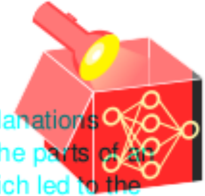
- We trained a model to predict if an image shows a wolf or a husky
- Below the predictions on six test images are given
- Do you trust our predictor?
- Sometimes the ML model is wrong
- Can you guess the pattern the ML model learned to identify a wolf?

Source: [Sameer Singh 2018]

Interpretable Machine Learning – 4 / 9

EXAMPLE: HUSKY OR WOLF? USING LIME

Source: [Sameer Singh 2018]



- Local explanations highlight the parts of an image which led to the prediction

⇝ our predictor is actually a snow detector

- Local explanations highlight the parts of an image which led to the prediction

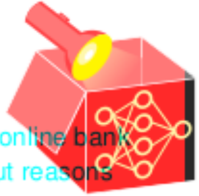
⇝ our predictor is actually a snow detector

Source: [Sameer Singh 2018]

Interpretable Machine Learning – 5 / 9

EXAMPLE: LOAN APPLICATION

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons



Source: [https://www.elte.hu]

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons

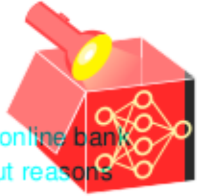
Source: [https://www.elte.hu]

EXAMPLE: LOAN APPLICATION

Source: [https://www.elte.hu]

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

  "If you were older than 21, your loan application would have been accepted."



- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

  "If you were older than 21, your loan application would have been accepted."

Source: [https://www.elte.hu]

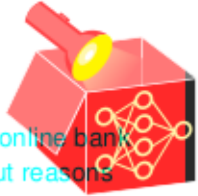
Interpretable Machine Learning – 6 / 9

EXAMPLE: LOAN APPLICATION

Source: [https://www.elte.hu]

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:
  "If you were older than 21, your loan application would have been accepted."

⇝ helps to understand the decision and to take actions for recourse (if req.)



EXAMPLE: LOAN APPLICATION

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:
  "If you were older than 21, your loan application would have been accepted."
  ⇝ helps to understand the decision and to take actions for recourse (if req.)

Source: [https://www.elte.hu]

Interpretable Machine Learning – 6 / 9

EXAMPLE: STOP OR RIGHT-OF-WAY?

- Imagine:
  - You work at a car company that develops image classifiers for autonomous driving
  - You show your model the following image (an adversarial example)

Interpretable Machine Learning – 7 / 9

EXAMPLE: STOP OR RIGHT-OF-WAY?

- Imagine:
  - You work at a car company that develops image classifiers for autonomous driving
  - You show your model the following image (an adversarial example)
  - Classifier is 99% sure it describes a right-of-way sign
- Would you entrust other peoples lives into the hands of this software?

Source: [Eykholt et. al 2018]

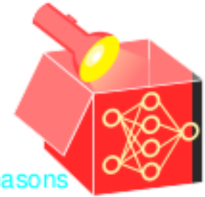
CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific prediction, local environment

- **Explanation scope:** Specific prediction, local environment

Interpretable Machine Learning – 8 / 9

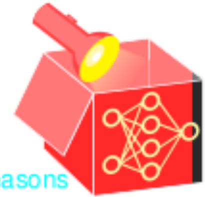
CHARACTERISTICS OF LOCAL EXPLANATIONS



- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons
  ⇝ very popular also for deep learning models

Interpretable Machine Learning – 8 / 9

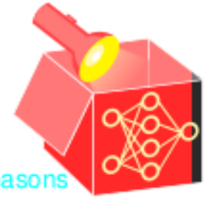
CHARACTERISTICS OF LOCAL EXPLANATIONS



- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons
  $\rightsquigarrow$ very popular also for deep learning models
- **Audience:** ML modelers and laypersons

Interpretable Machine Learning – 8 / 9

CHARACTERISTICS OF LOCAL EXPLANATIONS

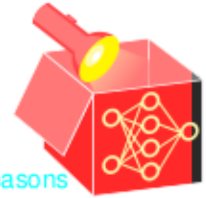


- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons
  $\rightsquigarrow$ very popular also for deep learning models
- **Audience:** ML modelers and laypersons
- **Data types:** Often agnostic, including tabular, image, text and audio data

Interpretable Machine Learning – 8 / 9

CHARACTERISTICS OF LOCAL EXPLANATIONS

CHARACTERISTICS OF LOCAL EXPLANATIONS

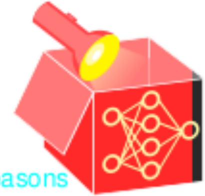


- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons
  $\rightsquigarrow$ very popular also for deep learning models
- **Audience:** ML modelers and laypersons
- **Data types:** Often agnostic, including tabular, image, text and audio data
- **Methods:** Many, most prominent are counterfactual explanations, shapley values, local interpretable model-agnostic explanations (LIME), adversarial examples, single ICE curve

- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons
  $\rightsquigarrow$ very popular also for deep learning models
- **Audience:** ML modelers and laypersons
- **Data types:** Often agnostic, including tabular, image, text and audio data
- **Methods:** Many, most prominent are counterfactual explanations, shapley values, local interpretable model-agnostic explanations (LIME), adversarial examples, single ICE curve

Interpretable Machine Learning – 8 / 9

CHARACTERISTICS OF LOCAL EXPLANATIONS



- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons
  ⇝ very popular also for deep learning models
- **Audience:** ML modelers and laypersons
- **Data types:** Often agnostic, including tabular, image, text and audio data
- **Methods:** Many, most prominent are counterfactual explanations, shapley values, local interpretable model-agnostic explanations (LIME), adversarial examples, single ICE curve
- **Special:** Due to audience, strong interactions with social sciences and strong connections to cognitive science and neurosciences due to data types

Interpretable Machine Learning – 8 / 9

CREDIT DATASET

- We illustrate local explanation methods on the German credit data See Kaggle
- 522 complete observations, 9 features containing credit and customer information
- Binary target "risk" indicates if a customer has a 'good' or 'bad' credit risk
- We merged categories with few observations

name	type	range
age	numeric	[19, 75]
sex	factor	{male, female}
job	factor	{0, 1, 2, 3}
housing	factor	{free, own, rent}
saving.accounts	factor	{little, moderate, rich}
checking.accounts	factor	{little, moderate, rich}
credit.amount	numeric	[276, 18424]
duration	numeric	[6, 72]
purpose	factor	{others, car, furniture, radio/TV}
risk	factor	{good, bad}

Interpretable Machine Learning – 9 / 9