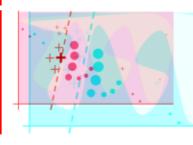
Interpretable Machine Learning

Introduction to Local Explanations





Learning goals

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

METHODOLOGICAL MOTIVATION

- Purpose of local explanations:

 - input., feature vector)

 Insight into the driving factors for a particular decision ion Understand the ML/moder's decisions in a local neighborhood of a given input (e.g., feature vector)

METHODOLOGICAL MOTIVATION

- Purpose of local explanations:
 - Insight into the driving factors for a particular decision ion
 - Underständ the ML/hoders decisions in a local neighborhood or a given input input., feature vector)
 - (e.g., feature vector)
- Local Methods can address questions such as:
- Local Méthodsican address que stions such as: for input x?
 - Why did the model decide to predictly for input x hat are similar to x?
 - How does the model decide for observations that are similar to x? ere different?
 - What would the ML mode have decided if x its values in X were different?
 - Where (in which regions in \(\mathcal{X} \)) does the model fail?

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

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

- Explanations for laypersons must be tailored to the explainee (who receives the sthe explanation) explanation) pecific, easy for humans to understand, and faithful to the explained mechanism case specific, easy for humans to understand, and faithful to the explained
 If algorithms make decisions in socially/safety critical domains, end users have a justified
- Tradecitions make decisions in socially/safety critical domains, end users have a justified mechanism interest in receiving explanations
- If algorithms make decisions in socially/safety critical domains, end users
 If algorithms make decisions in socially/safety critical domains, end users
 If algorithms make decisions in socially/safety critical domains, end users
- Local explanations cannot only increase user trust, but also help to detect critical local have a justified interest in receiving explanations biases in algorithmic decision making
- 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 s the explanation) explanation) pecific, easy for humans to understand, and faithful to the explained mechanism case specific, easy for humans to understand, and faithful to the explained
 If algorithms make decisions in socially/safety critical domains, end users have a justified
- If algorithms make decisions in socially/safety critical domains, end users have a justified mechanism interest in receiving explanations
- If algorithms make decisions in socially/safety critical domains, end users
 Local explanations cannot only increase user trust, but also help to detect critical local have a justified interest in receiving explanations places in algorithmic decision making
- Local explanations cannot only increase user trust, but also help to detect
 Local plases in algorithmic decision making
 Protection Hegulation (30) PR
- European citizens have the legally binding right to explanation as given in the Instead of explaining the entire (GDPR).

 General Data Protection Regulation (GDPR) in our reasonable.

 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

GDPR: THE RIGHT TO EXPLANATION

"The data subject should have the right not to be subject to a decision, which may include a "The data subject should have the right not to be subject to a decision, which may measure evaluating personal aspects relating to him or her which is based include a measure, evaluating personal aspects relating to him or her or similarly significantly and processing and which produces legal effects concerning him or her or similarly significantly affects 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 any human intervention.

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

decision reached after such assessment and to challenge the decision.

(Recital 71, GDPR (Recital 71, GDPR)

EXAMPLE: HUSKY OR WOLF??

- We trained a model to predict if an image shows a wolf or a husky
- 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?



Source: [Sameer Singh 2018]

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

EXAMPLE: HUSKY OR WOLF? USING LIME



EXAMPLE: LOAN APPLICATION



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

and are immediately rejected without reas

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

EXAMPLE: LOAN APPLICATION



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

Bank could e.g. provide a counterfactual without reason explanation using local explanation a counterfactual methods:
 explanation using local explanation methods:

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

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 ine: You apply for a loan at an online better.

Bank could e.g. provide a counterfactual without reason explanation using local explanation a counterfactual methods:
 explanation using local explanation methods:

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

accepted."

helps to understand the decision and to take

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

EXAMPLE: STOP OR RIGHT-OF-WAY?



Imagine:

- You work late a care company that at develops develops image classifiers formous driving autonomous driving del the following image
- You show your model the following image (an adversarial example)

EXAMPLE: STOP OR RIGHT-OF-WAY?



Imagine:

 You work at a car company that at develops develops image classifiers for mous driving autonomous driving del the following image

- You show your model the following
 - image (an adversarial example) ribes a
- Classifier is 99% sure it describes a
- Would you entrust other peoples lives into the
- Would you entrust other peoples lives into the hands of this software?



Source: [Eykholt et. al 2018]

Source: [Eykholt et. al 2018]

• Explanation scope: Specific prediction, local environmentnent



- Explanation scope: Specific prediction; local environment, ent
- Model classes: Model-agnostic by definition, model-specific for computational reason reasons y popular also for deep learning models
 → very popular also for deep learning models

- Explanation scope: Specific prediction; local environment, ent
- Model classes: Model-agnostic by definition, model-specific for computational reason reasons y popular also for deep learning models ~ Audie nce: ML modelers and laypersons
- Audience: ML modelers and laypersons

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

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

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

CREDIT DATASET

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

	name	type		range
name	age type	numeric	range	[19, 75]
age	senumeric	factor	[19, 75]	(male, female)
sex	jobfactor .			tle}(0, 1, 2, 3)
job	housifactor	factor	{0, 1, 2, 3	(free, own, rent)
housing saving acfactors		factor(free, own(litent) moderate, rich)		
saving.accountsking.afactorits		fac(tittle, moderate, rich) derate, rich)		
checking.accountslit.arfactor		num(tittle, moderate, rich), 18424]		
credit.amount	duraniumeric			4] [6, 72]
duration	purpnumeric			car, furniture, radio/TV)
purpose	rishumeric			e, radio/TV}ad}
risk	factor		good, bad	

