Local surrogate interpretable model

Explainable and Trustworthy AI

Eliana Pastor

Stages of Explainability

- Explainability involves the entire AI development pipeline

Pre-modelling explainability

Before building the model
- Data exploration
- Data selection
- Feature engineering

Explainable modeling

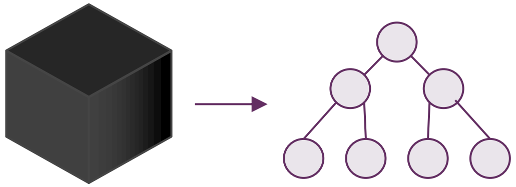
Build inherently interpretable models
- Manage the accuracy and interpretability trade-off

Post-modelling explainability

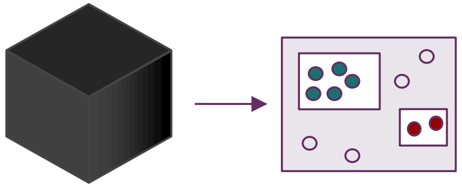
After model development
- Explaining predictions and behavior of trained models

Scope of Explainability

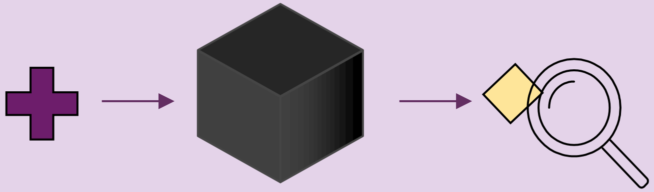
- *What do we explain?*



Global



Subgroup



Individual/local

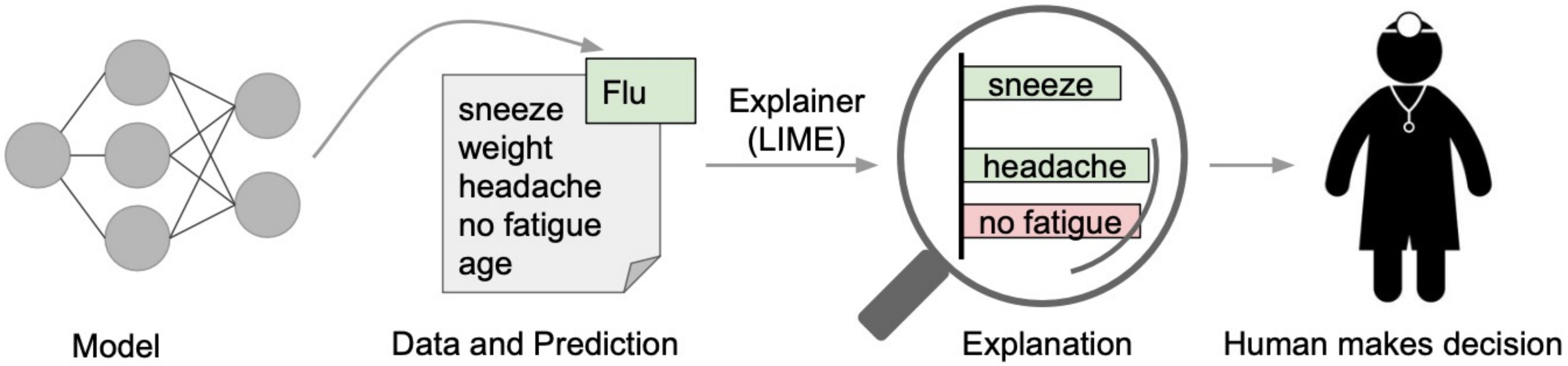
How the model globally works

How the model behaves in data subgroups

Explaining the reasons behind individual predictions

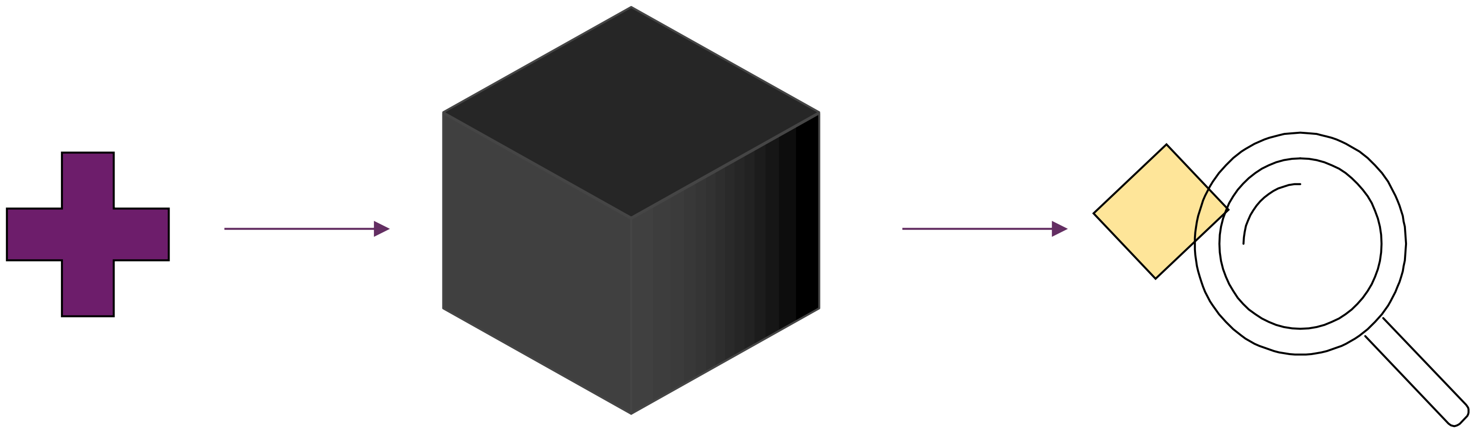
Explaining individual predictions

- Presenting **textual or visual artifacts** that **provide qualitative understanding** of the relationship between the instance’s components (e.g. words in text, patches in an image) and the model’s prediction.



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. KDD 2016.

Explaining individual prediction via model agnostic solutions

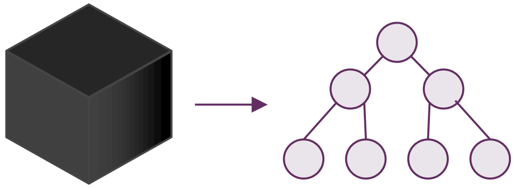


Methodology to derive explanations

- Local surrogate interpretable models
- Explaning by removing
- Gradient-based explanation methods
- Counterfactual methods

Local surrogate interpretable models

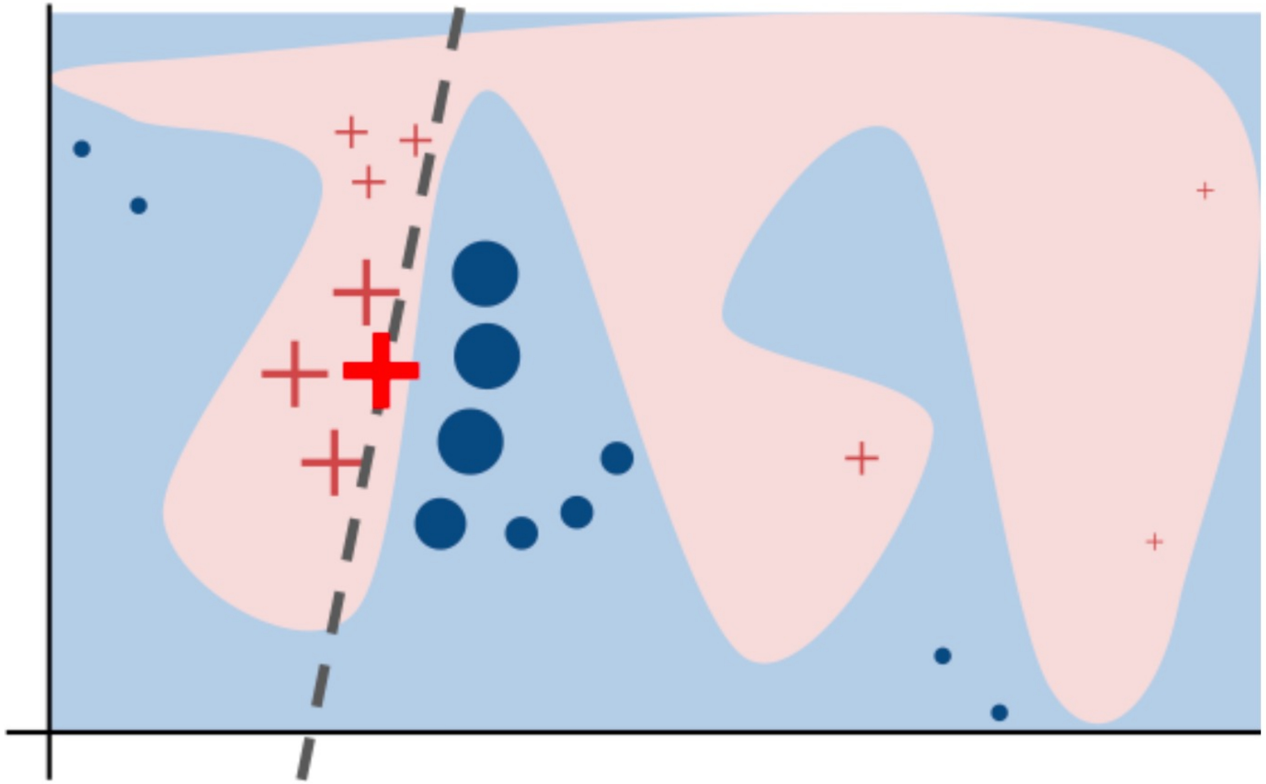
From global surrogate..



To local surrogate..

On the locality of the prediction

Local surrogate interpretable models



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. KDD 2016.

LIME
Local Interpretable Model-Agnostic Explanations

- Train a local intepretable model in the **locality of the prediction**
- **Interpretable model** use **interpretable representations**

- **Locality of the prediction**
  - Neighborhood of the instance → proximity
  - Generated via **perturbed samples**

- **Intepretable model**
  - E.g., linear model

- **Interpretable representations**
  - Representations interpretable for us a humans

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. KDD 2016.

Property of explanations

- **Interpretable**
  - Provide qualitative understanding, easy to interpret
  - **Features for explaining can be different from features for training**!
    - Notion of **interpretable data representation**

- **Locally faithful**
  - Correspond to **how the model behaves in the vicinity of the instance being explained**
  - Property of local fidelity
  - Local fidelity do not imply global fidelity!

LIME - Local surrogate - definition

$$explanation(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

- x an instance to explain and f is the model to explain
- G is the family of possible interpretable models
- π_x is proximity measure between x and instance perturbed z; define locality
- $\Omega(g)$ is the complexity of g (e.g., number of non zero weights in a linear model)

The explanation for instance x is the model g that minimizes loss L:

- $L(f, g, \pi_x)$ – how unfaithful g is to f in the locality given by π_x - **Local**
- $\Omega(g)$ – how the model is interpretable is kept low - **Interpretable**

LIME – High level steps

- Given x
  - Generate the neighborhood of x
  - Get the predictions of f for these local points
  - Weight the samples according to their proximity to x
  - Train a weighted, interpretable model on the neighborhood labeled dataset
  - Explain the prediction by interpreting the local model

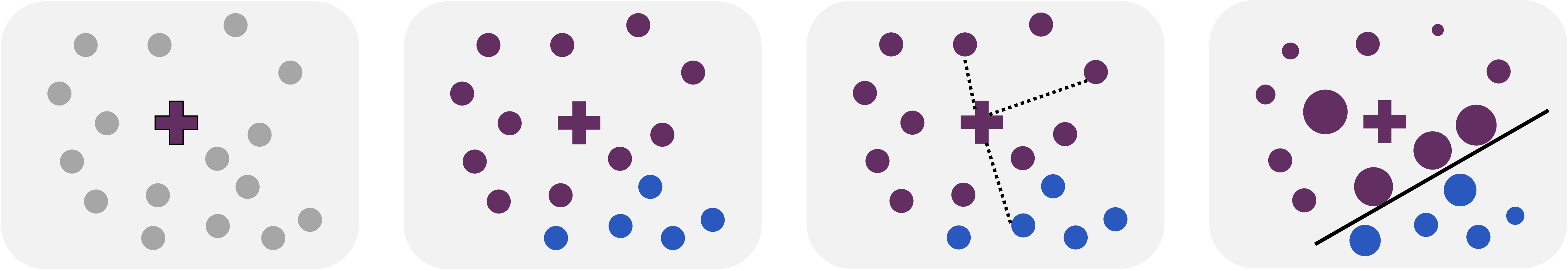
LIME – High level steps

Generate locality

Label with f

Weigh by proximity

Train a linear model



LIME – Points to address

- **(a) Interpretable representations**
  - The local model operates on ‘interpretable representations’
  - What is an interpretable representation?

- **(b) Locality of the prediction**
  - How to generate it?

- **(c) Interpretable model**
  - Which class of model to consider?

Interpretable data representation – (a)

- Explanations need to use a representation interpretable to humans
  - It can differ from the representation used by the model

Text

Input

Embeddings

Embeddings

Interpretable representation

Hello world

Interpretable data representation – (a)

- Explanations need to use a representation interpretable to humans

Images

Input

WxHxC

Interpretable representation

Super-pixel/patches

Image from: https://www.inovex.de/de/blog/lime-machine-learning-interpretability/

Interpretable data representation – (a)

- Explanations need to use a representation interpretable to humans

Tabular data

Already interpretable

gender=Female, age=30

gender=Female, age=30

Interpretable data representation – (a)

- Interpretable data representation are encoded as **binary vector** denoting the presence of absence of a (interpretable) feature

Text

hello	world
1	1

Image

Patch 1	Patch 2
1	1

Text - Locality of the prediction - (b)

- Neighbour samples are generated by randomly removing words from the input text
- Operating on the binarized interpretable representation
  - Feature values: 1 if the corresponding word is included and 0 if it has been removed

Welcome	to	the	Explainable	and	Trustworthy	AI	Course	Probability	Proximity
1	1	0	0	1	0	0	1	0.8	0.8
1	0	1	1	1	0	1	0	0.9	0.9
0	1	0	0	0	1	1	1	0.5	0.7

Text - Locality of the prediction - (b)

Welcome	to	the	Explainable	and	Trustworthy	AI	Course	Probability	Proximity
1	1	0	0	1	0	0	1	0.8	0.8
1	0	1	1	1	0	1	0	0.9	0.9
0	1	0	0	0	1	1	1	0.5	0.7

- **Prediction Probability**
  - Assigned by original model

- Original model works on original space
  - Concatenate words – removing the omitted ones
    - ‘Welcome to and Course’
  - Replace the omitted words with special tokes
    - ‘Welcome to [UNK] and [UNK] [UNK] Couse’

Text - Locality of the prediction - (b)

Welcome	to	the	Explainable	and	Trustworthy	AI	Course	Probability	Proximity
1	1	0	0	1	0	0	1	0.8	0.8
1	0	1	1	1	0	1	0	0.9	0.9
0	1	0	0	0	1	1	1	0.5	0.7

- **Proximity**
    - Between perturbed instance and original ones
    - Cosine similarity

Image – Locality of the prediction (b)

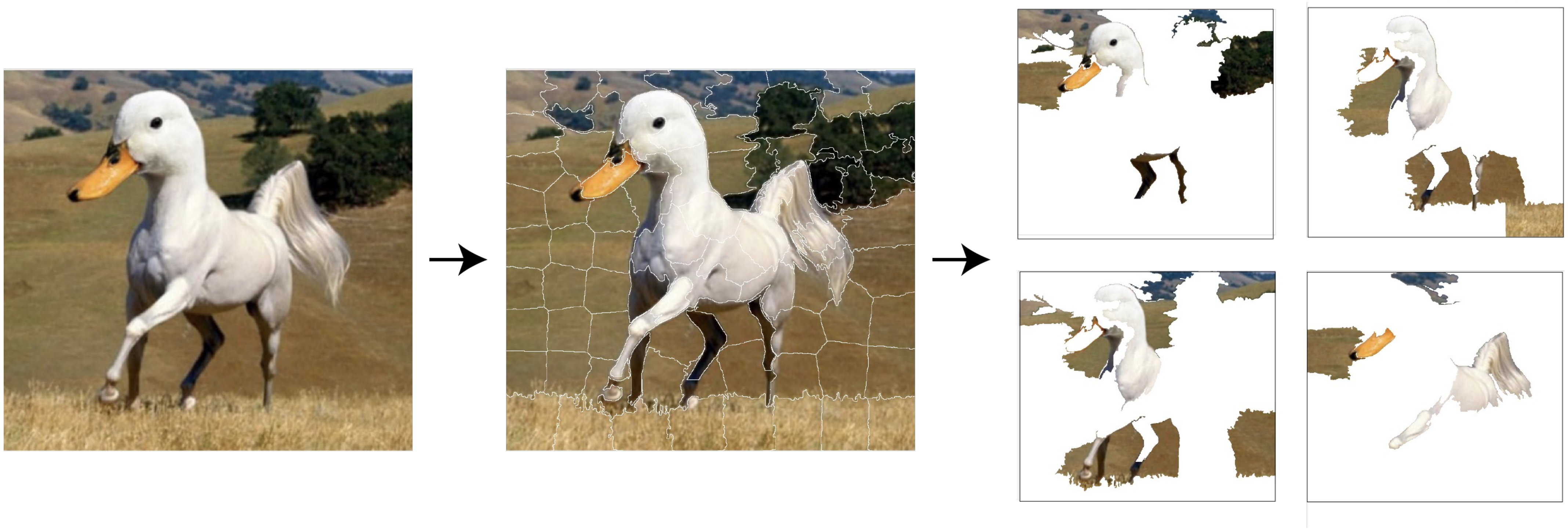
Intepretable representation via superpixels

Input

Interpretable representation

From https://ema.drwhy.ai/LIME.html

Image – Locality of the prediction (b)



From https://ema.drwhy.ai/LIME.html

Tabular – Locality of the prediction (b)

- For numerical features,
  - **perturb** them by sampling from a Normal(0,1)
  - inverse operation of mean-centering and scaling, according to the means and stds in the training data.

- For categorical features,
  - **perturb** by sampling according to the training distribution
  - Represent as a binary feature that is 1 when the value is the same to input to compute proximity

From gihub lime/lime_tabular.py doc

Interpretable model – Choice of g – (c)

- Train an interpretable model g on the generate samples, represented via interpretable representation

$$L(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z)\big(f(z) - g(z')\big)^2$$

- Linear model
  - LASSO to regularize – minimize the number of non-zero coefficient
  - Ridge - Linear least squares with l2 regularization in the code

- Parameter K to control the interpretability
  - E.g., Text: Limit the number of words
  - It applies a feature selection steps

Advantages of LIME

- Model agnostic
- Local explanations

- Interpretable representations
  - Distinction between representations used by the model and by the explanation
  - Different level of abstractions

- Provides feature attributions

- We can control the number of intepretable features
  - Shorter explanations = more interpretable

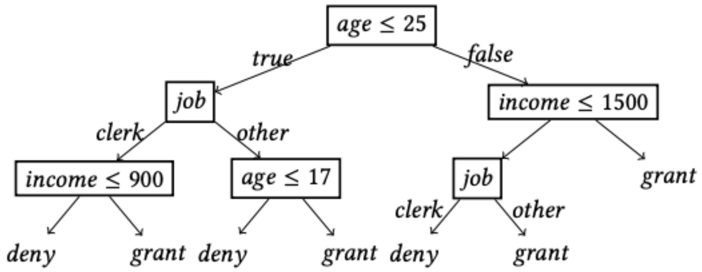
- Support multiple types of data (images, text, tabular)

Disadvantages of LIME

- Perturbated sample for the neighborhood may be unrealistic
  - Do not consider correlations
- Sensitive to the choice of perturbation method and the magnitude of perturbations
- Explanation instability – differ in multiple runs
  - Surrogate model relies on the random perturbations for the neighborhood
- Choice of the number and locality of the neighborhood
- Potential inconsistency
  - Explanations depend on the local neighborhood
  - Explanations for similar instances can differ, potentially leading to inconsistencies in interpretation
- Dependence on the interpretable representation

LORE
Local Rule-Based Explanations

- Local surrogate
    - Decision tree classifier



- Locality/Neighborhood
    - Based on genetic algorithm

Provide explanation as

- Decision path, i.e., local rule
- a set counterfactual rules*, i.e., the conditions should be changed to change the predicted class

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems.

Advantages and limitations of LORE

Advantages.

- Model agnostic
- Local explanations
- Provides local rules
- Provide conterfacual explanations

Limitations.

- Genetic neighboorhood could be more expensive to generate
- Generated samples may be unrealistic
- Focus on structured data

LACE

- Local surrogate
  - Associative classifier

- Locality/Neighborhood
  - Based on actual neighborhood

Provide explanation as

- Association rule, i.e., local rule
- Feature attributions as prediction difference *for individual features and local rules

Pastor, Eliana, and Elena Baralis. "Explaining black box models by means of local rules." *SAC* 2019.

Advantages and limitations of LORE

Advantages.

- Model agnostic
- Local explanations
- Provides local rules
- Provide prediction differences for individual features and local rules

Limitations.

- Require the actual training data to derive the neighborhood
- Neighbohood from the training data could be insufficient for the local behavior
- Focus on structured data

References

- Molnar, Christoph. *Interpretable machine learning* https://christophm.github.io/interpretable-ml-book/
- **[SUGGESTED]** Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. KDD 2016.
- Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems.
- Pastor, Eliana, and Elena Baralis. "Explaining black box models by means of local rules." SAC 2019.