UNIVERSITÄT MANNHEIM

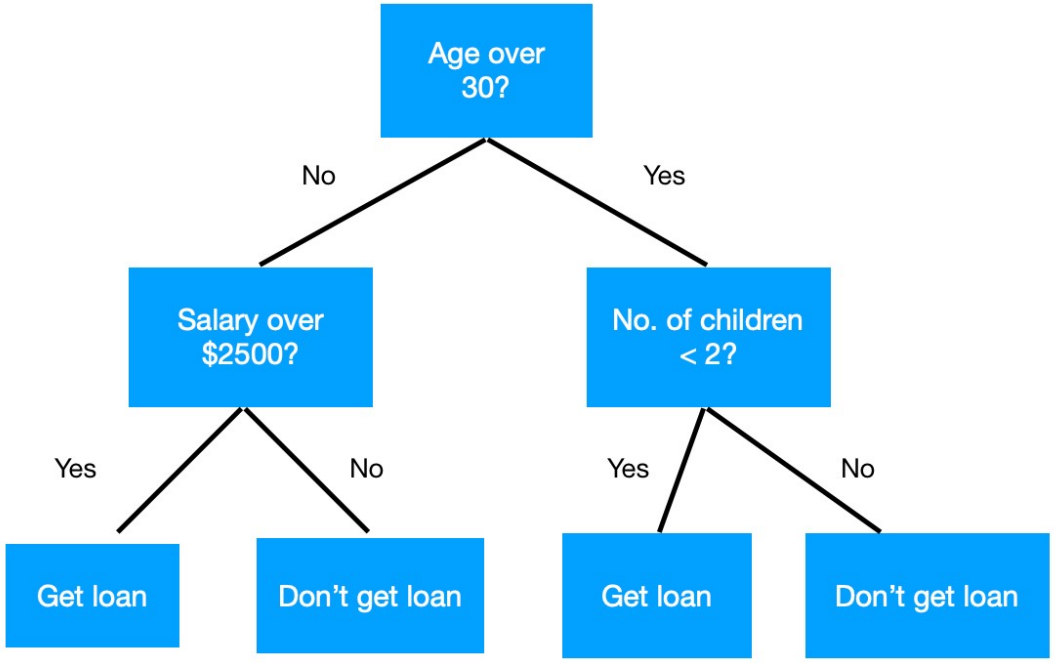
Explainable Artificial Intelligence

Seminar Kick Off

Heiko Paulheim

Introduction

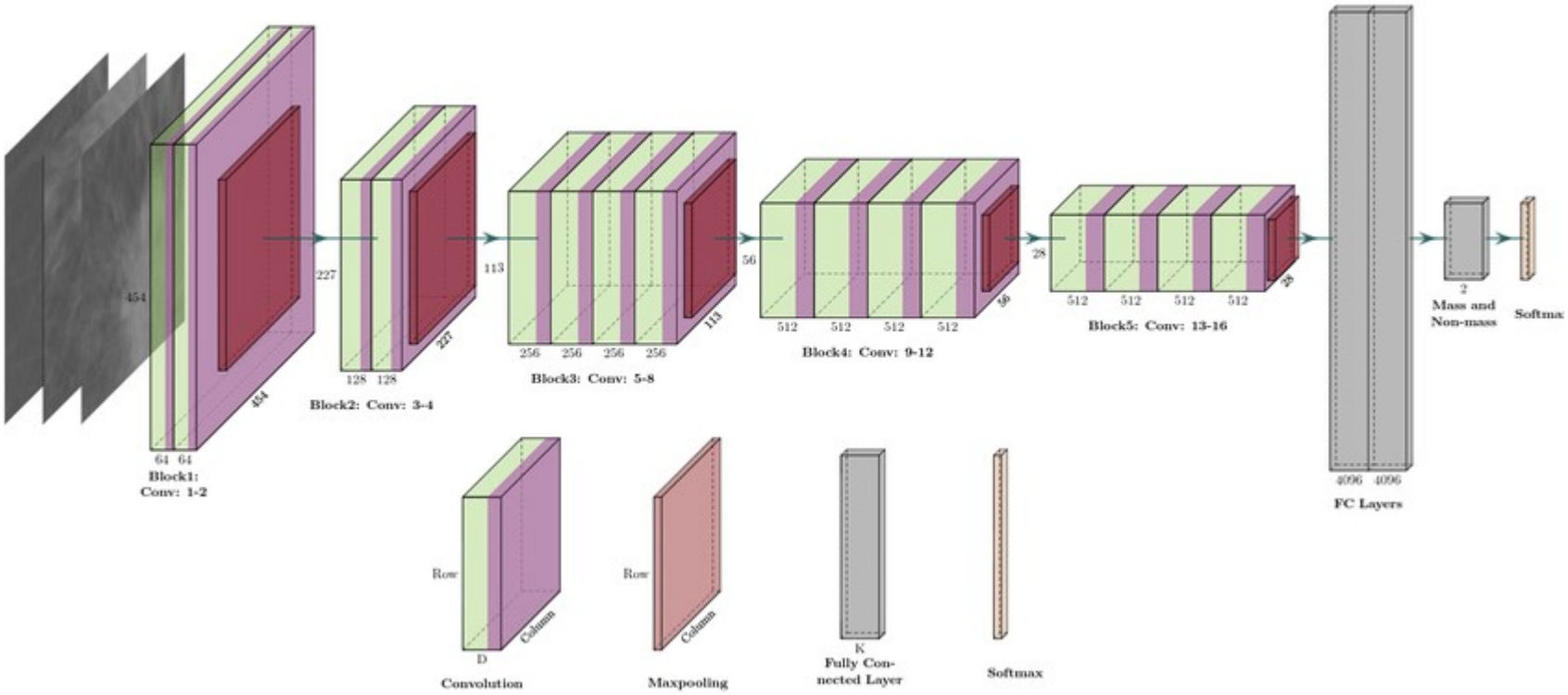
- Artificial Intelligence / Machine Learning in the early days:
  - Explainable by nature



https://eloquentarduino.github.io/2020/10/decision-tree-random-forest-and-xgboost-on-arduino/

Introduction

- Artificial Intelligence / Machine Learning nowadays:
  - Not very transparent



https://www.researchgate.net/publication/334388209_Automatic_Mass_Detection_in_Breast_Using_Deep_Convolutional_Neural_Network_and_SVM_Classifier

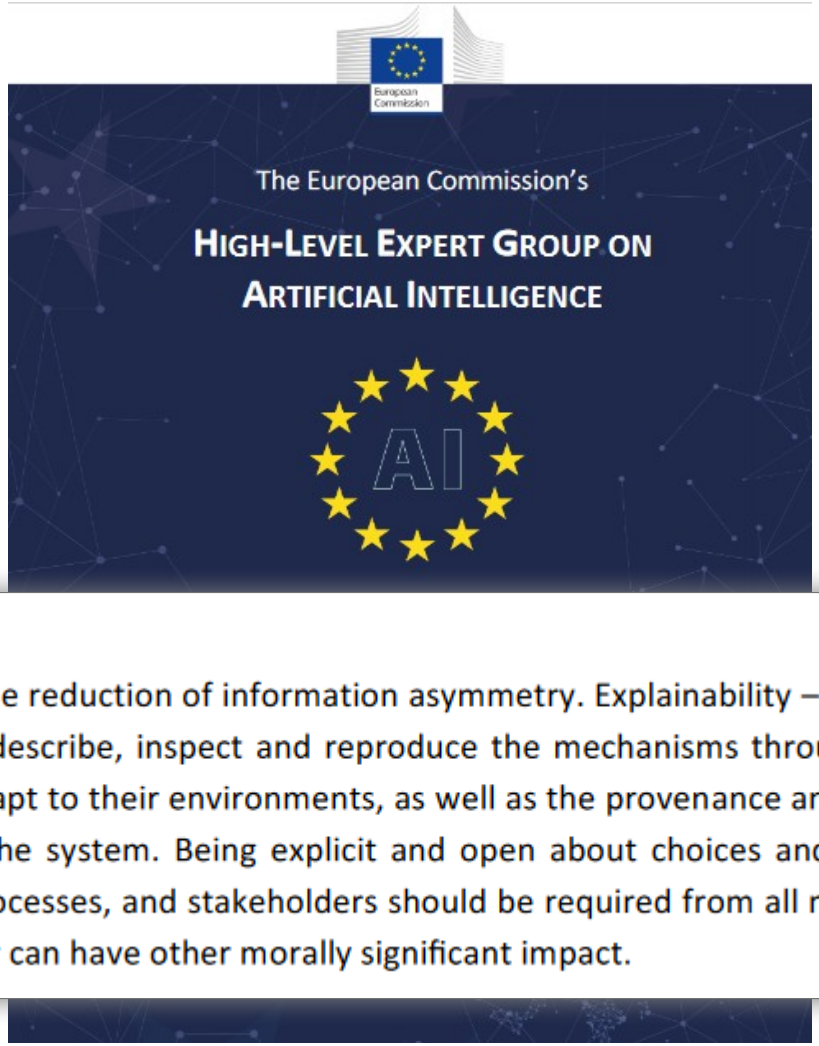
Why Bother?

- The Principle of Explicability: "Operate transparently"

Transparency is key to building and maintaining citizen's trust in the developers of AI systems and AI systems themselves. Both technological and business model transparency matter from an ethical standpoint. Technological transparency implies that AI systems be auditable,[14] comprehensible and intelligible by human beings at varying levels of comprehension and expertise. Business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems.

Explicability[15] is a precondition for achieving informed consent from individuals interacting with AI systems and in order to ensure that the principle of explicability and non-maleficence are achieved the requirement of informed consent should be sought. Explicability also requires accountability measures be put in place. Individuals and groups may request evidence of the baseline parameters and instructions given as inputs for AI decision making (the discovery or prediction sought by an AI system or the factors involved in the discovery or prediction made) by the organisations and developers of an AI system, the technology implementers, or another party in the supply chain.

Why Bother?

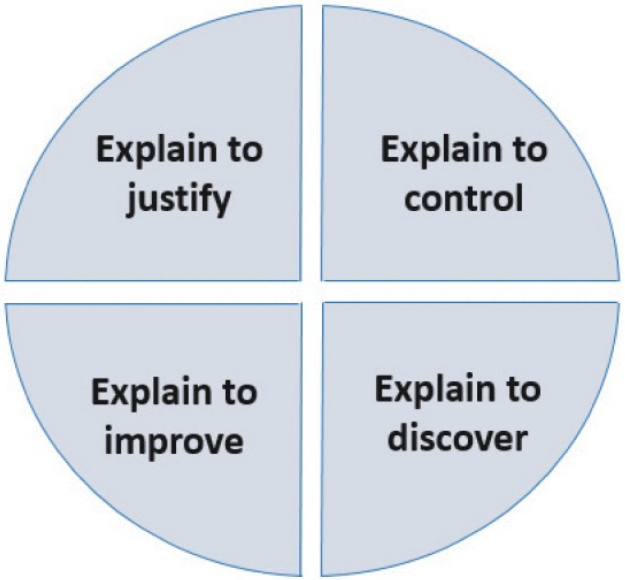


10. Transparency

Transparency concerns the reduction of information asymmetry. Explainability – as a form of transparency – entails the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environments, as well as the provenance and dynamics of the data that is used and created by the system. Being explicit and open about choices and decisions concerning data sources, development processes, and stakeholders should be required from all models that use human data or affect human beings or can have other morally significant impact.

Introduction

- Who is the target for Explainable AI?
  - End users
  - Developers
  - …



https://ieeexplore.ieee.org/document/8466590

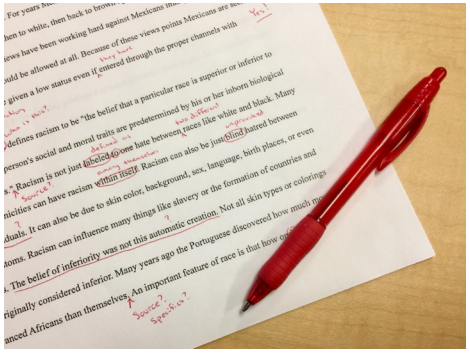
Introduction

- Intrinsic vs. post-hoc
  - e.g., decision trees are intrinsically interpretable
  - post-hoc: take a trained model (e.g., neural net) and create explanations
- Global vs. local
  - global: explain the entire model at once
  - local: explain a decision made by the model
- Model-specific vs. model-agnostic
  - some approaches work for a particular model type (e.g., CNN)
  - others are applicable to any ML model

https://ieeexplore.ieee.org/document/8466590

Organization

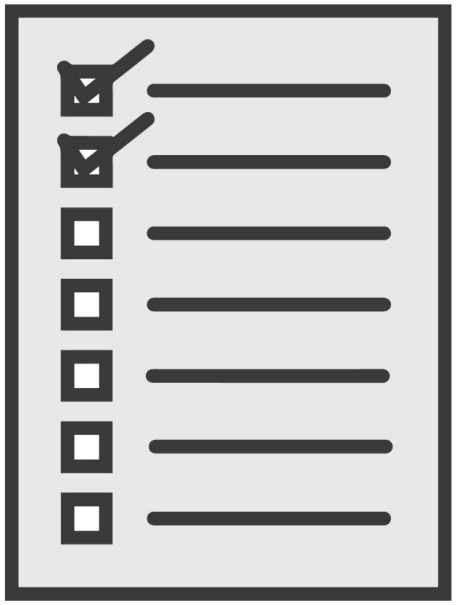
- Requirements
  - Familiarize yourself with a particular approach to XAI
  - Present it in the seminar
  - Write a seminar paper
  - Review others' seminar papers
    - it is a good idea to also read the main papers for the topics you review



- First step
  - Pick topics of interest
  - Send a ranked list to Ms. Bianca Lermer by the end of the week

Organization

- We will use a process called “peer review”
  - widely used (and discussed) in science
  - you will review your fellow students’ seminar papers
- Timeline
  - Prepare a draft until March 30th
  - You will get two seminar papers to review
  - Submit your reviews until April 14th
- Seminar (i.e. ,presentations, discussions)
  - April 28th, May 5th, 12th, 19th
  - We are planning an on campus seminar
- Final seminar paper submission: June 18th



Preparing Your Seminar Paper

- Conduct initial literature research
  - get your hands on some core papers on the topic
- Read follow up papers about the topic
  - which follow-up approaches/refinements have been developed?
- Find papers citing the original paper
  - who says what about it?
  - who uses it, and for which purposes?
- Anything else you want to discuss

Preparing a Review

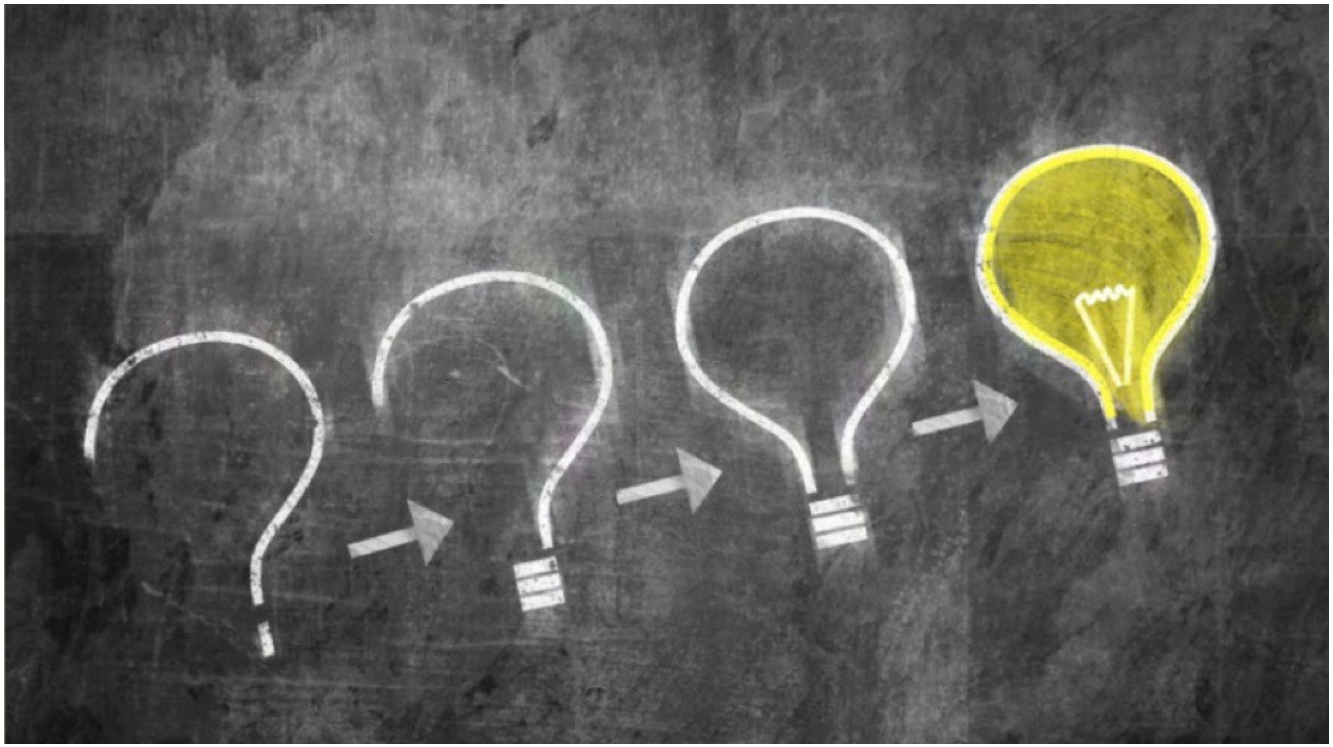
- 1st rule: be constructive!
- What you should point at
  - can you follow easily? is there information missing at any point?
  - are all claims well supported?
  - do you have any questions not answered?
  - aspects underrepresented
- What you should not do
  - provide general criticism ("don't like the paper")
  - correct every spelling mistake
  - rewrite the seminar paper

Preparing the Presentation

- You don’t have to start at zero
  - everybody should be familiar with XAI
- Focus on key aspects
  - which approach is taken in the study you present?
  - how was it evaluated?
  - what are its strengths and weaknesses?
- Be illustrative
  - use examples
  - try it out yourself, if possible
- Be entertaining
  - it’s (partly) up to you whether we are having fun here ;-)

Let's Talk about Topics

- Note: these are suggestions
  - you are free to set your own focus
  - and formulate questions which you find interesting



Topics

TABLE 2. Summary of explainability techniques.

Techniques	References	Intrinsic/Post-hoc	Global/Local	Model-specific/ Model-agnostic
Decision trees	[139], [140], [141], [142], [143]	I	G	SP
Rule lists	[66], [143], [144], [145], [146]	I	G	SP
LIME	[84], [85], [102], [147]	H	L	AG
Shapely explanations	[101]	H	L	AG
Saliency map	[87], [88], [89], [90], [91], [96], [97]	H	L	AG
Activation maximization	[82], [83]	H	G	AG
Surrogate models	[106], [107], [84]	H	G/L	AG
Partial Dependence Plot (PDP)	[108], [51], [110]	H	G/L	AG
Individual Conditional Expectation (ACI)	[112], [113]	H	L	AG
Rule extraction	[74], [114], [115], [116], [117], [118]	H	G/L	AG
Decomposition	[93], [94], [95]	H	L	AG
Model distillation	[49], [123], [124], [125], [126], [127]	H	G	AG
Sensitive analysis	[129], [130]	H	G/L	AG
Layer-wise Relevance Propagation (LRP)	[131]	H	G/L	AG
Feature importance	[113], [132], [86]	H	G/L	AG
Prototype and criticism	[133], [134], [135], [136]	H	G/L	AG
Counterfactuals explanations	[137]	H	L	AG

I: Intrinsic, H: Post-hoc, G: Global, L: Local, SP: Model-specific, AG: Model-agnostic

https://ieeexplore.ieee.org/document/8466590

Topics

- Local interpretable model-agnostic explanations (LIME)

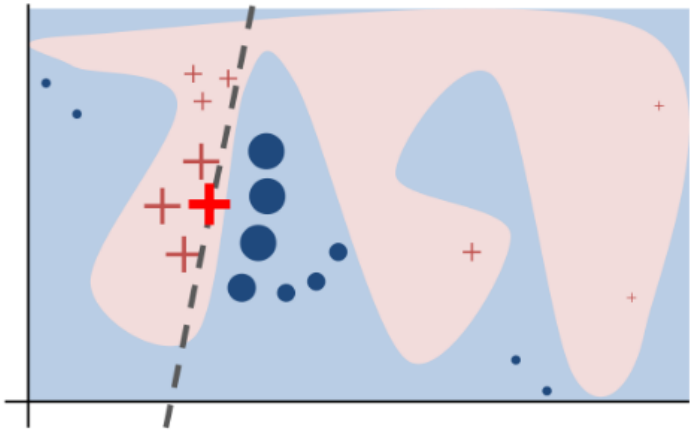
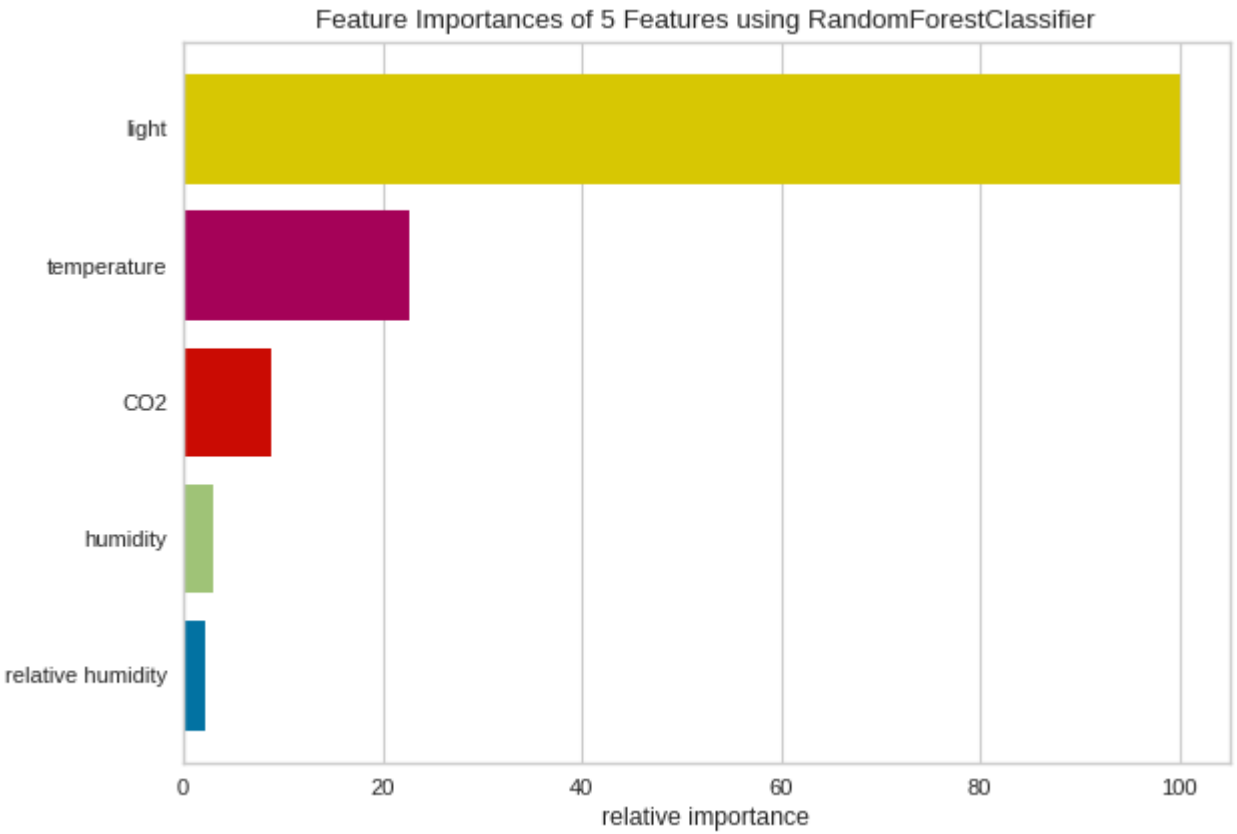


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

https://arxiv.org/pdf/1602.04938.pdf

Topics

- Feature importance



https://www.scikit-yb.org/en/latest/api/model_selection/importances.html

Topics

- SHAP (Shapley Values)

https://github.com/slundberg/shap

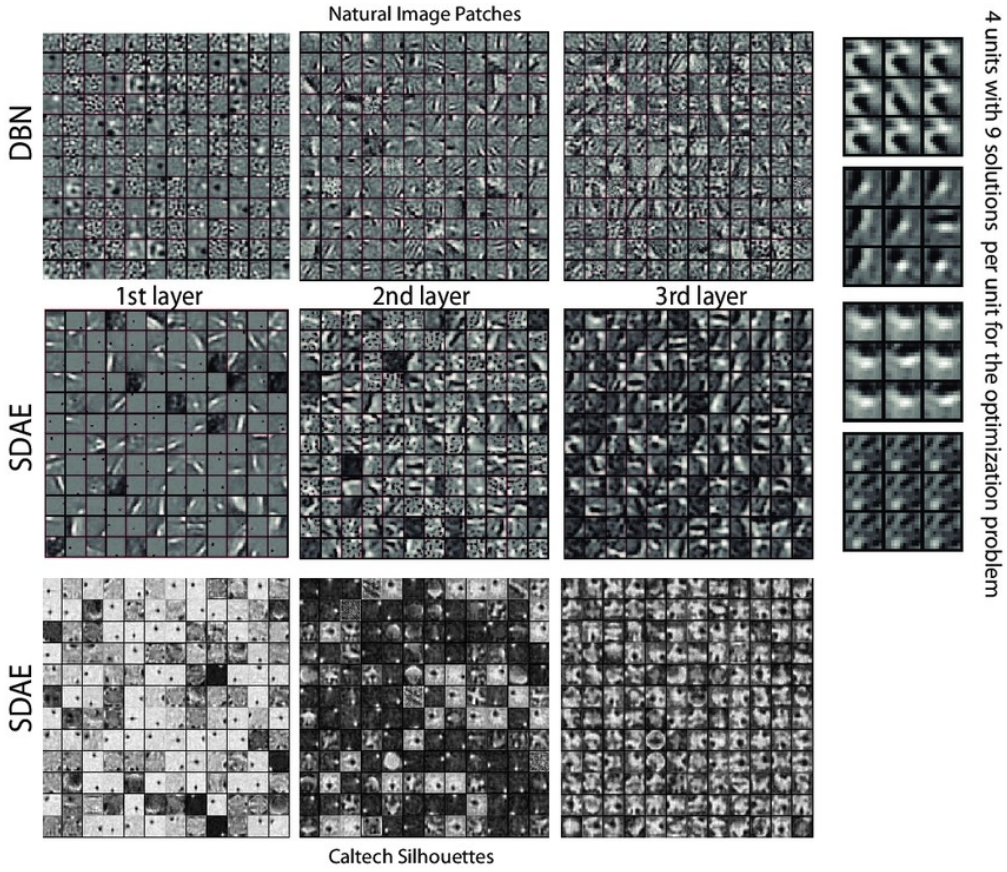
Topics

- Saliency Maps

https://www.geeksforgeeks.org/what-is-saliency-map/

Topics

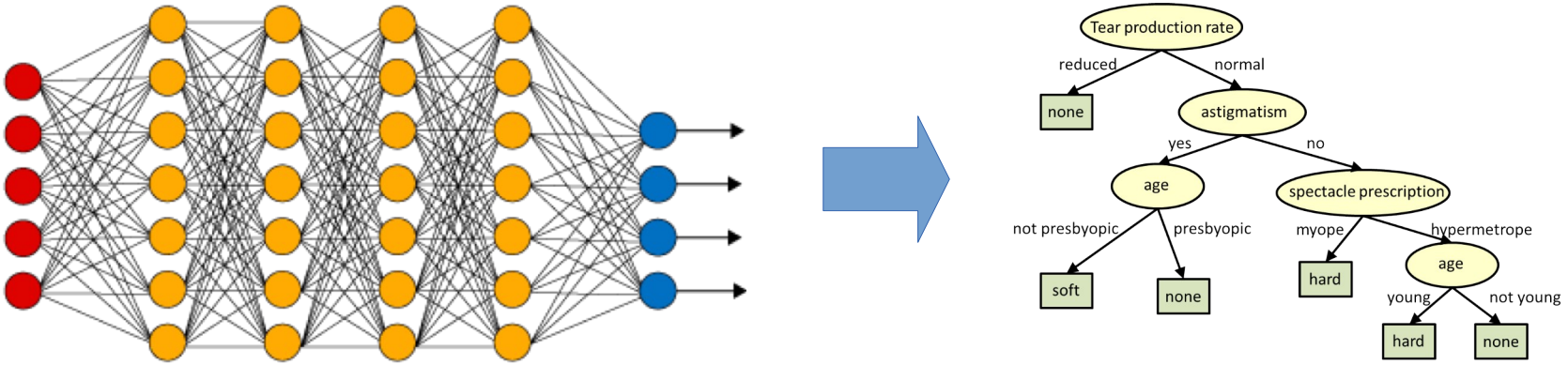
- Activation Maximization



http://www.iro.umontreal.ca/~lisa/pointeurs/invariances_techreport.pdf

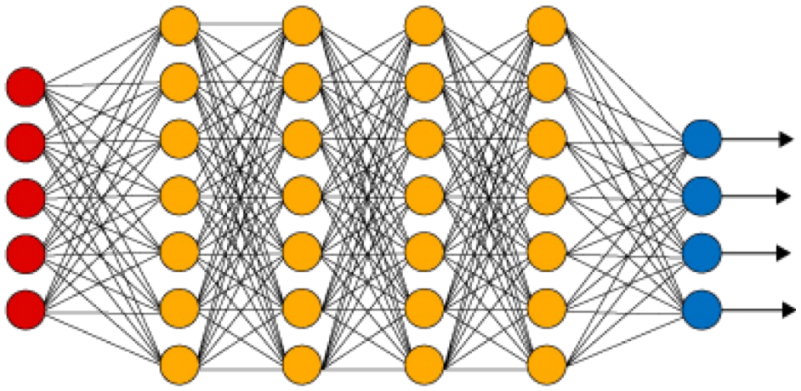
Topics

- Surrogate Models



Topics

- Rule Extraction from Neural Networks



Rule 1:
Exp(CFD) ∈ [1.80,7.20] ∧ Exp(AMACR) > 4.04
→ class 1

Rule 2:
Exp(NELL2) > 4.32 ∧ Exp(ATP8A1) > 3.22
→ class 2

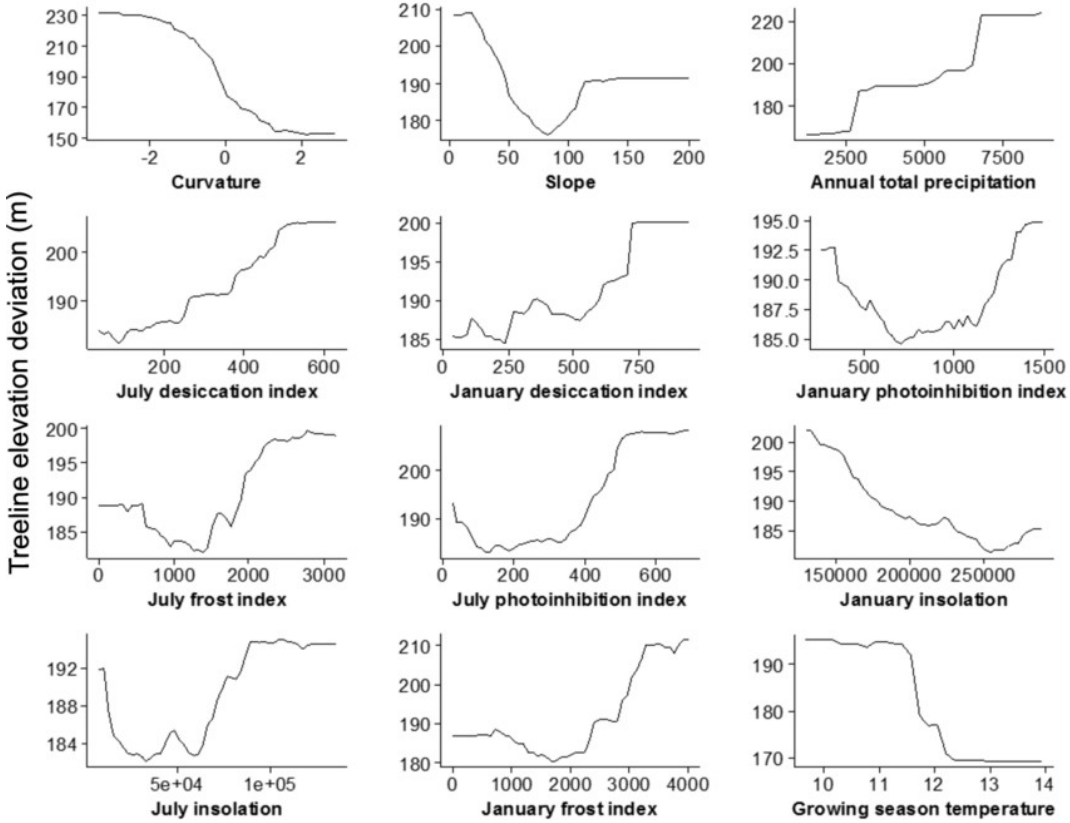
Rule 3:
Exp(HBB) ∈ [5.57,8.88]
→ class 1

Rule 4:
Exp(SUMO3) > 2.09
→ class 2

Rule 5:
Otherwise → class 1

Topics

- Partial Dependency Plots



https://www.researchgate.net/publication/283071368_Local-scale_topoclimate_effects_on_treeline_elevations_A_country-wide_investigation_of_New_Zealand%27s_southern_beech_treelines

Topics

- Individual Conditional Expectation Plots

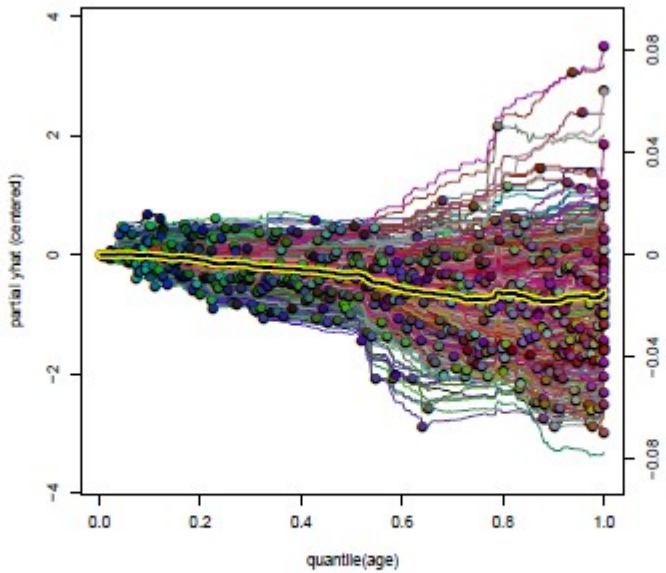
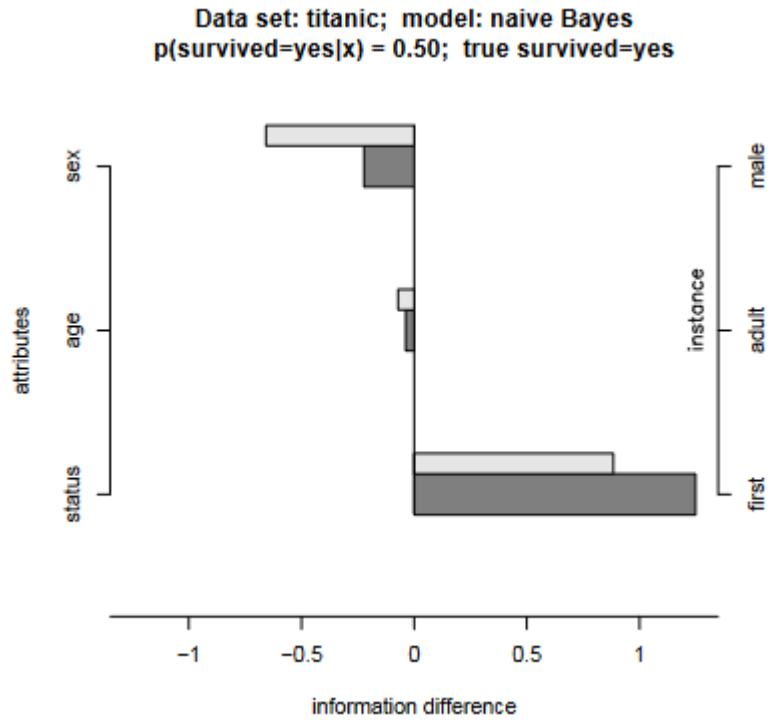


Figure 3: c-ICE plot for age with x^* set to the minimum value of age. The right vertical axis displays changes in $\hat{f}$ over the baseline as a fraction of y's observed range.

https://www.researchgate.net/publication/257028373_Peeking_Inside_the_Black_Box_Visualizing_Statistical_Learning_With_Plots_of_Individual_Conditional_Expectation

Topics

- Decomposition



https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4407709

Topics

- Model Distillation

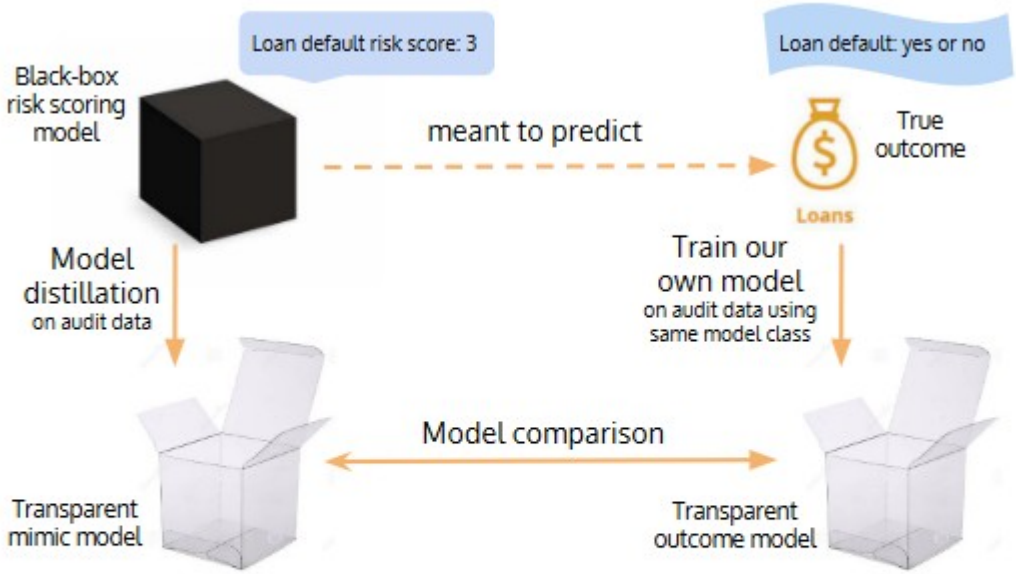
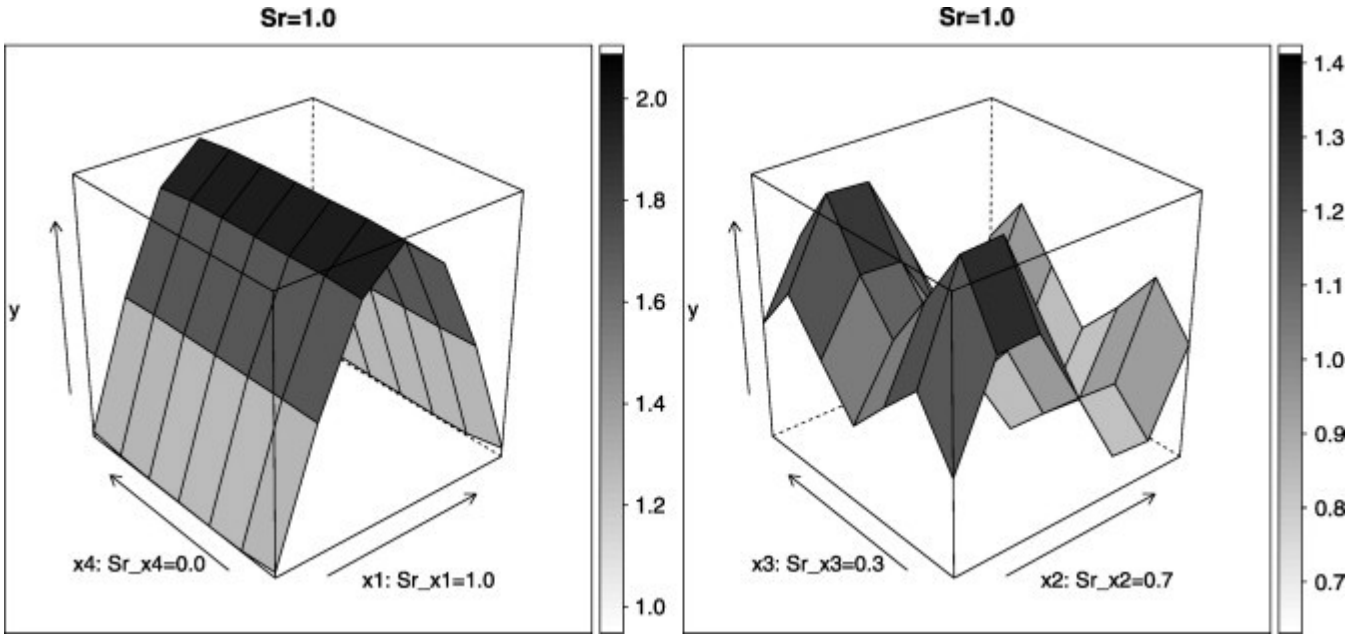


Figure 1: Auditing a loan risk scoring model by training transparent models on data labeled with the risk scores and with ground-truth outcomes for loan defaults.

https://arxiv.org/pdf/1710.06169.pdf

Topics

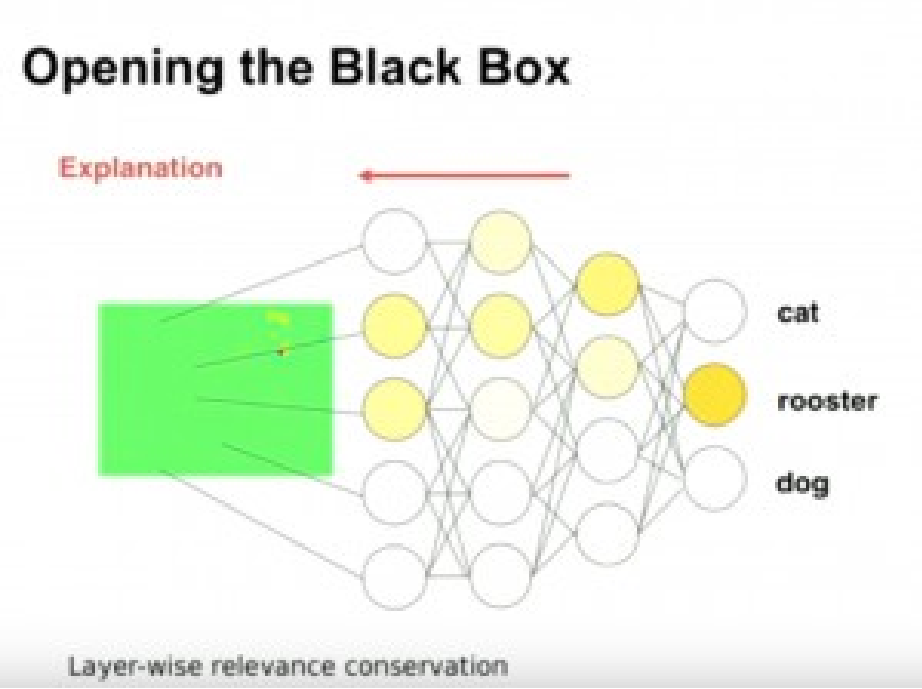
- Sensitivity Analysis



https://www.sciencedirect.com/science/article/pii/S0020025512007098

Topics

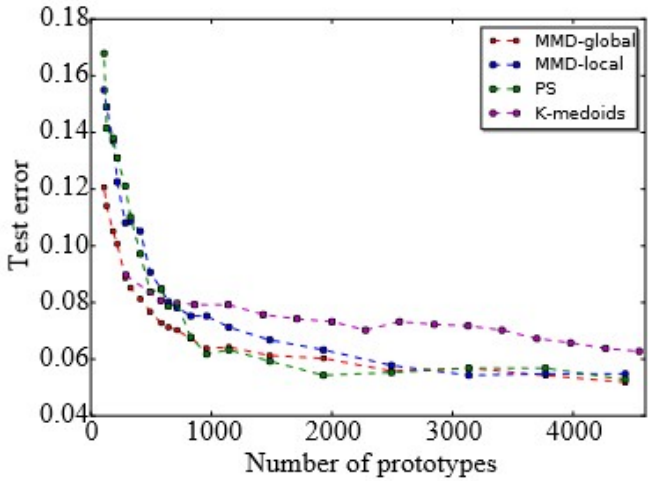
- Layer-wise Relevance Propagation



https://odsc.medium.com/layer-wise-relevance-propagation-means-more-interpretable-deep-learning-219ff5158914

Topics

- Prototypes and criticisms



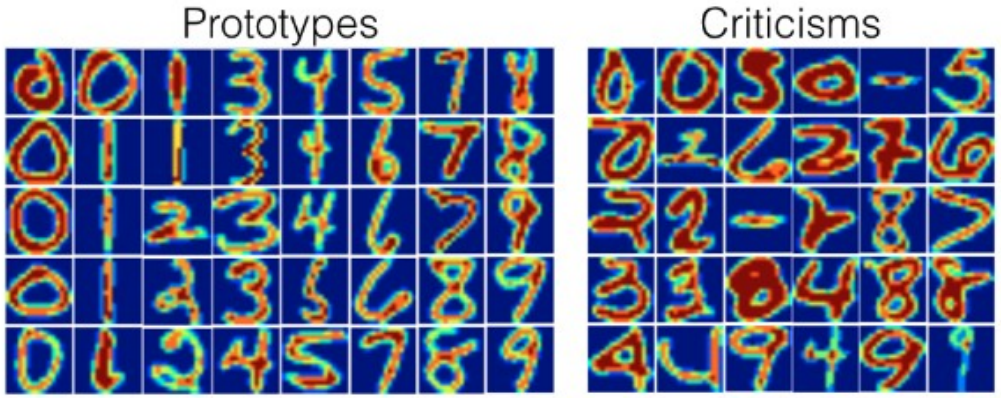
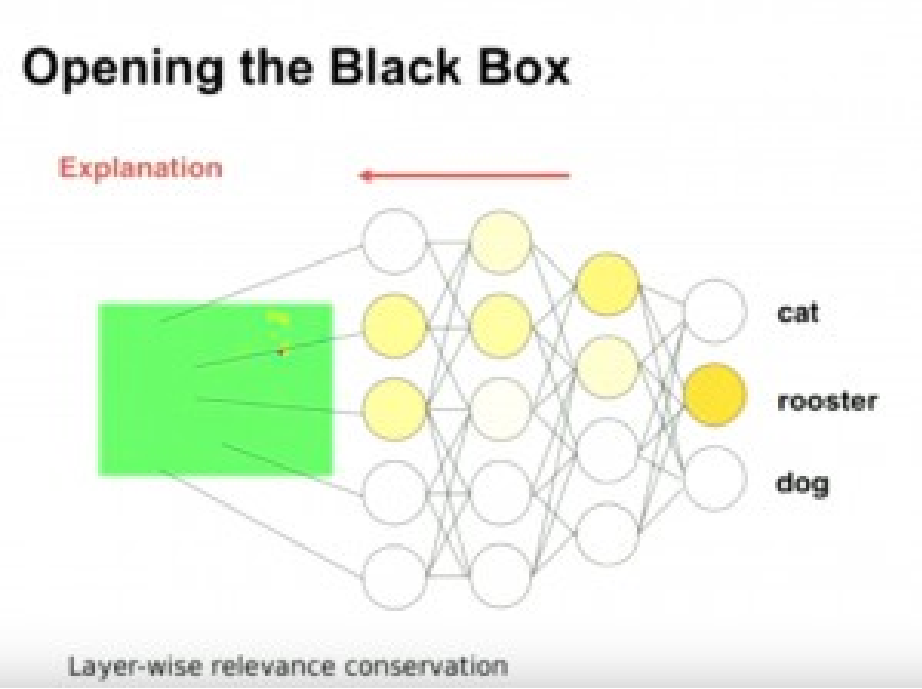


Figure 1: Classification error vs. number of prototypes $m = |S|$. `MMD-critic` shows comparable (or improved) performance as compared to other models (left). Random subset of prototypes and criticism from the USPS dataset (right).

https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

Topics

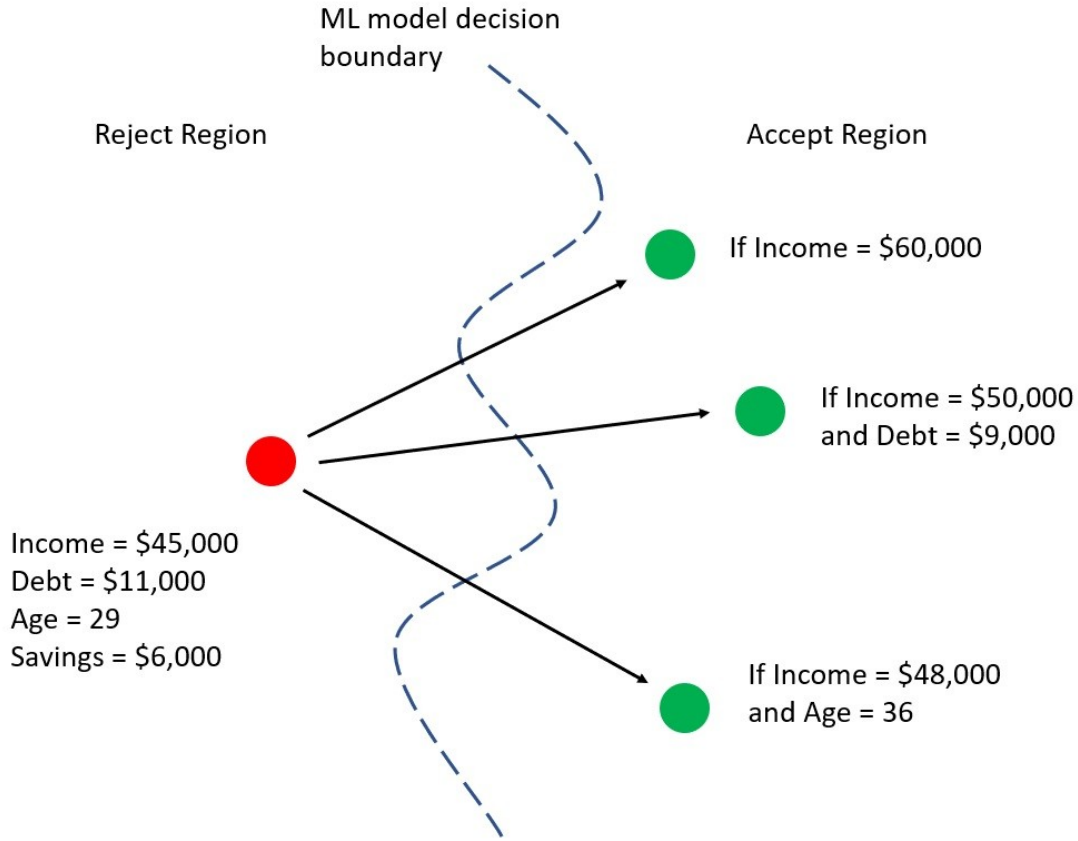
- Layer-wise Relevance Propagation



https://odsc.medium.com/layer-wise-relevance-propagation-means-more-interpretable-deep-learning-219ff5158914

Topics

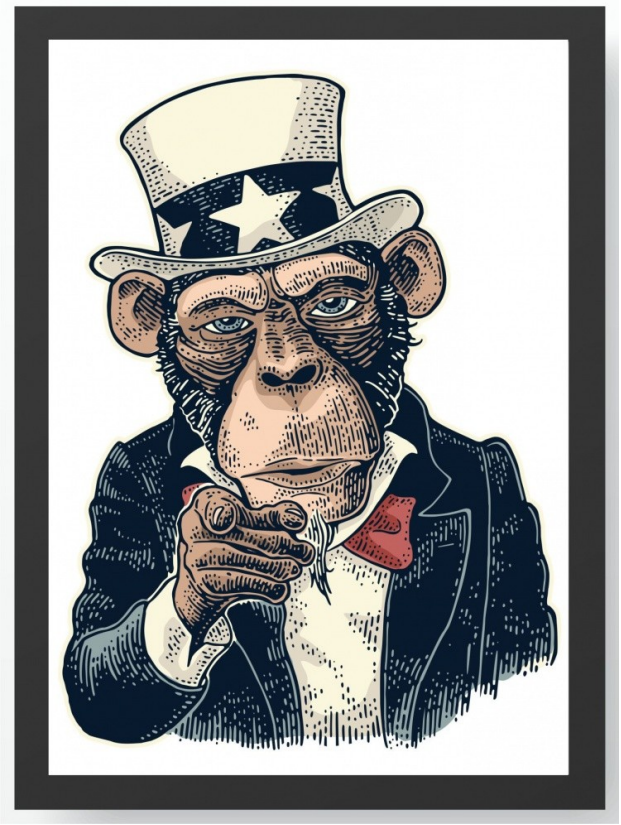
- Counterfactual explanations



https://pureai.com/articles/2020/03/13/open-source-counterfactuals.aspx

Now it’s up to you...

- ...pick your favorites
- Maybe you have other ideas? They are welcome!



Questions?