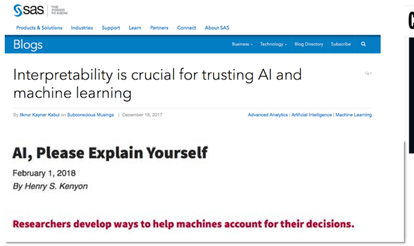
Interpreting Neural Network Judgments via Minimal, Stable, and Symbolic Corrections

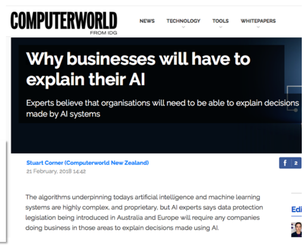
Xin Zhang, Armando Solar-Lezama, and Rishabh Singh

MIT CSAIL

Google

Motivation



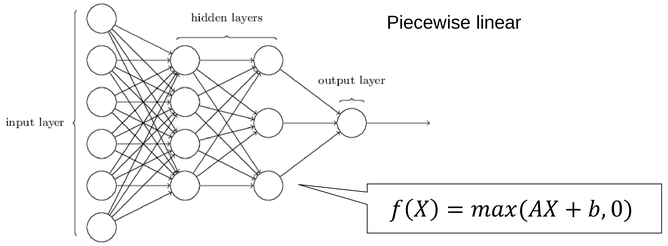


General Data Protection Regulation
(enacted 2016, taking effect 2018)

- Grants consumers a "right to explanation"
- Fine for violations: up to 10 million EURO

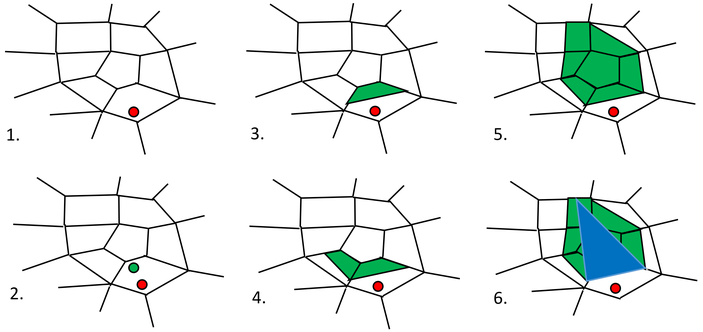
Algorithm

We focus on neural networks with ReLU activations:



Piecewise linear

Our algorithm is inspired by concolic testing:



Problem Definition

Judgement Problem: special binary classification problem where one result is more desirable than the other.

Corrections as interpretations:

Increase your salary | Improve your credit score | Lower your current debt
Ask for less money | Look at a different type of house | …

Desired properties of corrections:

Minimal

✓ Your application would be approved if you increase the monthly income by $100

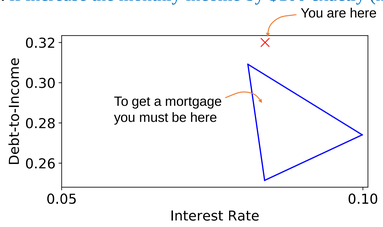
✗ … if you become a millionaire living in Montana

Stable

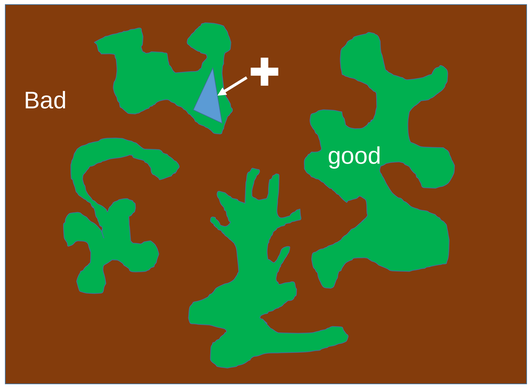
✓ … if you increase the monthly income by $100 (+/-10 is fine)

✗ … if increase the monthly income by $100 exactly (not even $99 or $101)

Symbolic

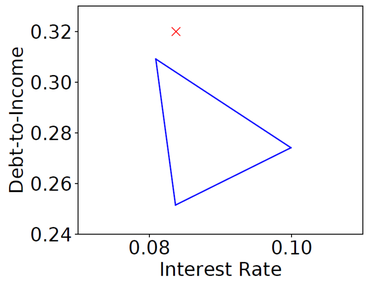


Our corrections are sound underapproximations that are local to the input



Experiment Results

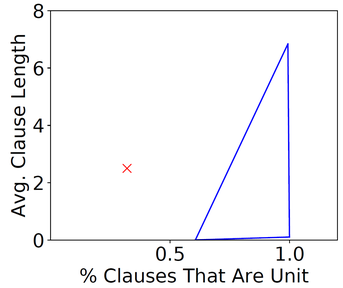
Mortgage Underwriting



Network structure: 5-layer feedforward network with 1000 neurons
Dataset: applications and performance of 36 million single-family loans
Explanations: triangles (best 2 features out of 5)

% inputs that our approach was able to generate corrections to: 85%
Average runtime: 20 minutes

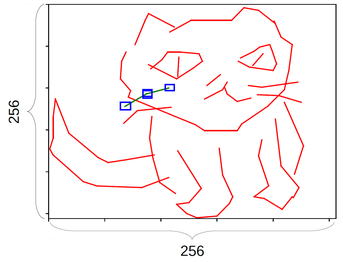
Solver Performance Prediction



Network structure: 5-layer feedforward network with 800 neurons
Dataset: statistics of 6k first-order theorems and whether they can be solved by a solver
Explanations: triangles (best 2 features out of 5)

% inputs that our approach was able to generate corrections to: 81%
Average runtime: 2 minutes

Drawing Tutorial



Network structure: convolutional network with 4096 neurons
Dataset: 0.12 million variants of a canonical cat drawing and 0.12 million other cat drawings
Explanations: boxes (upto 20 dimensions)

% inputs that our approach was able to generate corrections to: 75%
Average runtime: 13 minutes