Automated Reasoning for Explainable AI

Joao Marques-Silva

ANITI, Univ. Toulouse, France

January 16, 2020

A disclaimer

A disclaimer – recent & not so recent work…

- SAT Solving (Clause learning, UIPs, …)
- Quantification & CEGAR (QBF, QMaxSAT, etc.)
- Function Synthesis (Min DNF cover, …)
- Inconsistency (MUS, MCS, etc.)
- Certification of Reasoners
- Model Checking, Synthesizing Invariants, ATPG, Reconfiguration
- Optimization (MaxSAT, MinSAT, PBO, WBO, etc.)
- Propositional Encodings, Backbones, Autarkies, Minimal models, etc.
- Enumeration (MUSes, MCSes, etc.)
- Proof Systems (DRMaxSAT, etc.)
- Primes, Abduction, DLs, etc.

A disclaimer – new area of research, since 2018...

SAT Solving (Clause learning, UIPs, ...)	Quantification & CEGAR (QBF, QMaxSAT, etc.)	Function Synthesis (Min DNF cover, ...)
Inconsistency (MUS, MCS, etc.)	Certification of Reasoners	Model Checking, Synthesizing Invariants, ATPG, Reconfiguration
Optimization (MaxSAT, MinSAT, PBO, WBO, etc.)	Propositional Encodings, Backbones, Autarkies, Minimal models, etc.	Enumeration (MUSes, MCSes, etc.)
Proof Systems (DRMaxSAT, etc.)	Primes, Abduction, DLs, etc.	Explainability & Interpretability in ML

A disclaimer – new area of research, since 2018...

SAT Solving (Clause learning, UIPs, ...)

Quantification & CEGAR (QBF, QMaxSAT, etc.)

Function Synthesis (Min DNF cover, ...)

Inconsistency (MUS, MCS, etc.)

Certification of Reasoners

Model Checking, Synthesizing Invariants, ATPG, Reconfiguration

Optimization (MaxSAT, MinSAT, PBO, WBO, etc.)

Propositional Encodings, Backbones, Autarkies, Minimal models, etc.

Enumerati
(MUSe

Many initial ideas...

Comments welcome!

Proof Systems (DRMaxSAT, etc.)

Primes, Abduction, DLs, etc.

Explainability & Interpretability in ML

Recent & ongoing ML successes

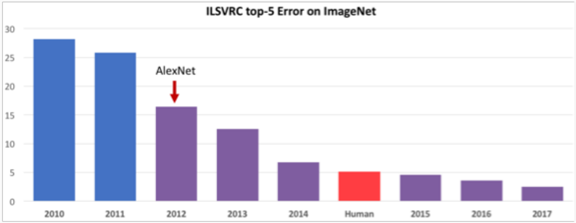
https://en.wikipedia.org/wiki/Waymo

DeepMind

AlphaGo

AlphaGo Zero & **Alpha Zero**

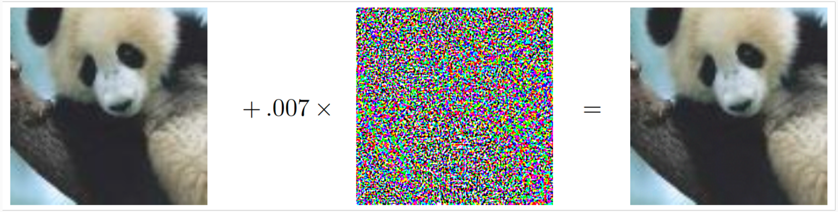
Image & Speech Recognition



http://gradientscience.org/intro_adversarial/

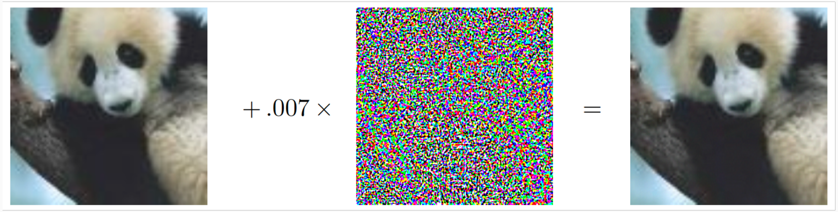
https://fr.wikipedia.org/wiki/Pepper_(robot)

But ML models are brittle — adversarial examples



Goodfellow et al., ICLR'15

But ML models are brittle — adversarial examples



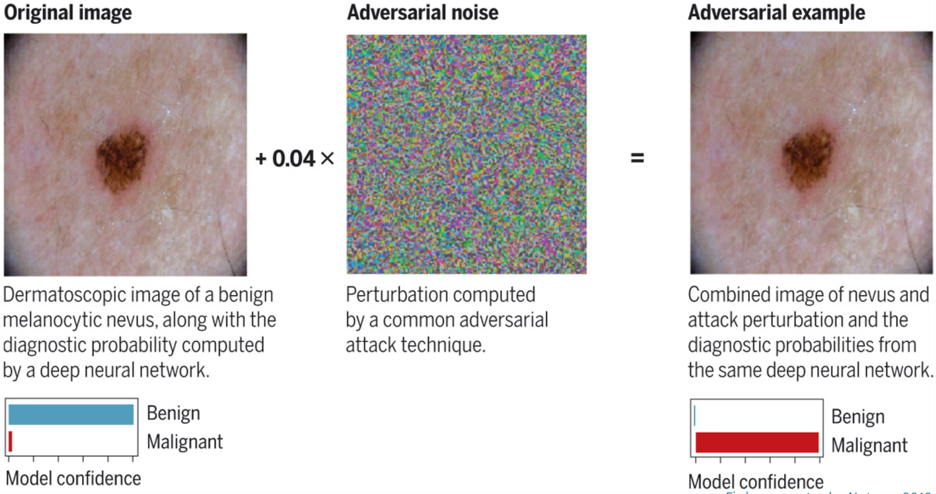
Goodfellow et al., ICLR'15

Eykholt et al'18

Aung et al'17

But ML models are brittle — adversarial examples

Adversarial examples can be very problematic



Finlayson et al., Nature 2019

Also, some ML models are interpretable

decision|rule lists|sets
decision trees; ...

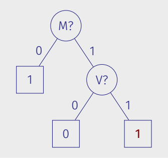
Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

Also, some ML models are interpretable

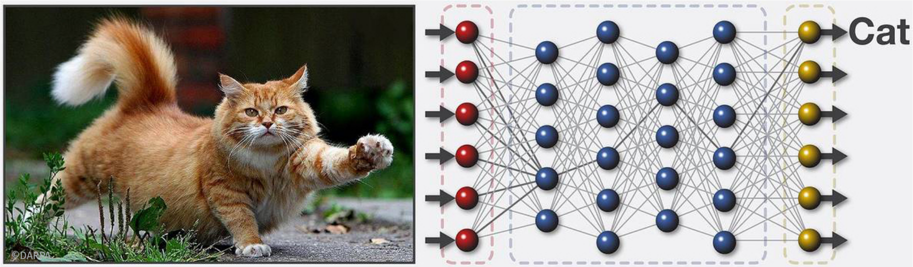
decision|rule lists|sets
decision trees; ...

if ¬Meeting **then** Hike
if ¬Vacation **then** ¬Hike

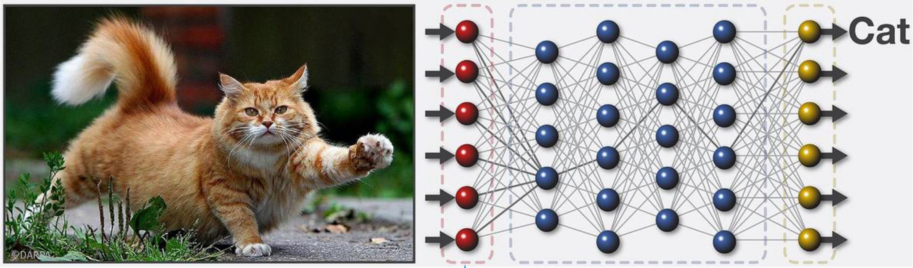
Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1



But other ML models are not (interpretable)...

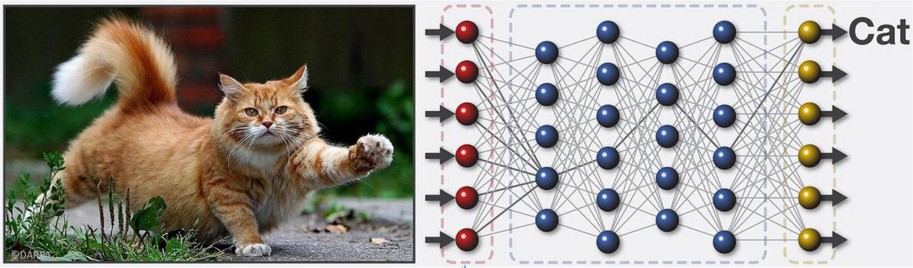


But other ML models are not (interpretable)...



Why does the NN predict a cat?

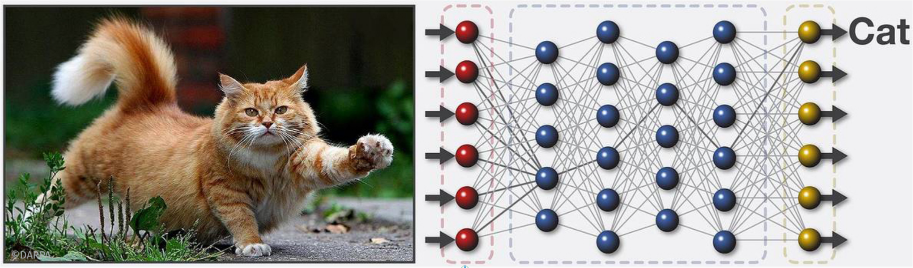
But other ML models are not (interpretable)...



Why does the NN predict a cat?

Which features matter?

But other ML models are not (interpretable)...



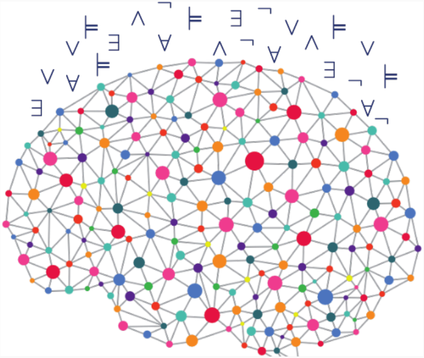
Why does the NN predict a cat?

Which features matter?

Are there general explanations??

ML meets AR

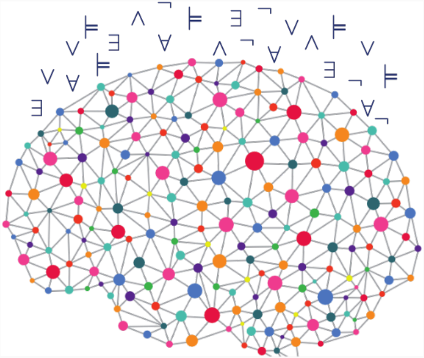
ML meets AR



"Combining machine learning with logic is the challenge of the day"

M. Vardi, MLmFM'18 Summit

ML meets AR – a challenge for the next decade?

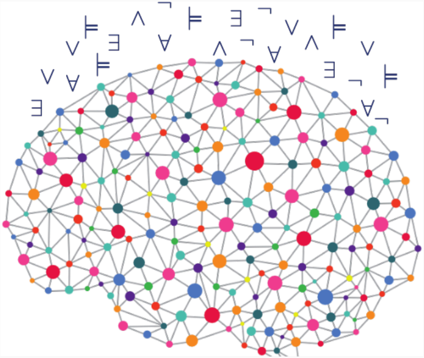


"Combining machine learning with logic is the challenge of the day"

M. Vardi, MLmFM'18 Summit

Exploit ML $\xrightarrow{\text{heuristics; portfolios; abstractions; tactics; …}}$ **Improve AR (Efficiency)**

ML meets AR – a challenge for the next decade?



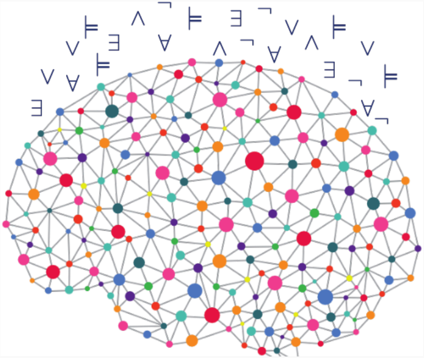
"Combining machine learning with logic is the challenge of the day"

M. Vardi, MLmFM'18 Summit

Exploit ML → heuristics; portfolios; abstractions; tactics; … → **Improve AR (Efficiency)**

Exploit AR → verification; synthesis; explanations; … → **Improve ML (Robustness)**

ML meets AR – a challenge for the next decade?



> “Combining machine learning with logic is the challenge of the day”
>
> M. Vardi, MLmFM’18 Summit

Exploit ML → **Improve AR (Efficiency)**

heuristics; portfolios; abstractions; tactics; …

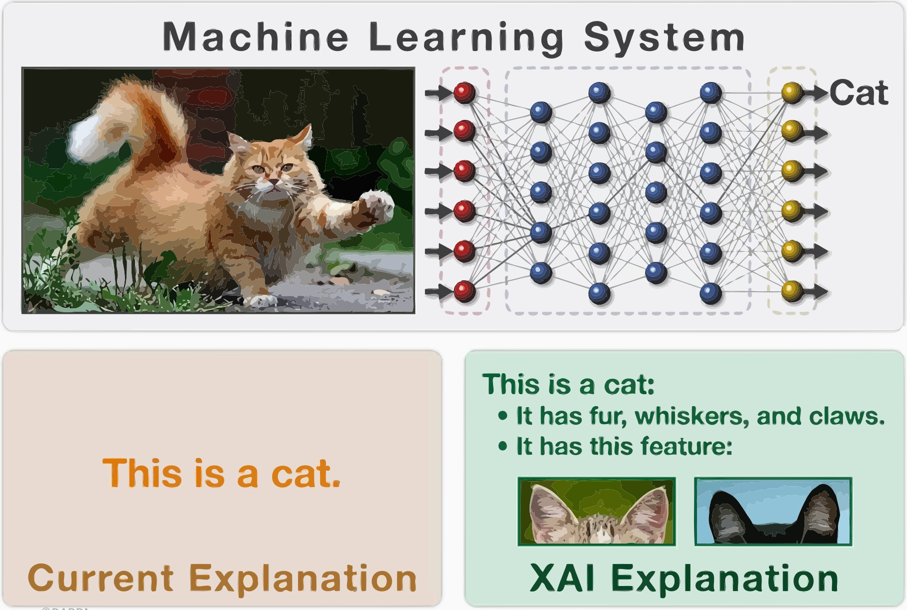
simplify system design

Exploit AR → **Improve ML (Robustness)**

verification; synthesis; explanations; …

build trust; debug; aid decision making

What is eXplainable AI (XAI)?



©DARPA

Why XAI?

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

Why XAI?

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman,[1*] Seth Flaxman,[2]

Why XAI?

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman,[1*] Seth Flaxman,[2]

■ *We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine-learning algorithms. Slated to take effect as law across the European Union in 2018, it will place restrictions on automated individual decision making (that is, algorithms that make decisions based on user-level predictors) that "significantly affect" users. When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them. We argue that while this law may pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks that avoid discrimination and enable explanation.*

Why XAI?

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman,[1*] Seth Flaxman,[2]

POLICY \ US & WORLD \ TECH \ TheVerge.com

A new bill would force companies to check their algorithms for bias

By Adi Robertson | @thedextriarchy | Apr 10, 2019, 3:52pm EDT

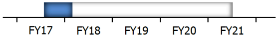
Algorithmic Accountability Act

■ *We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine-learning algorithms. Slated to take effect as law across the European Union in 2018, it will place restrictions on automated individual decision making (that is, algorithms that make decisions based on user-level predictors) that "significantly affect" users. When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them. We argue that while this law may pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks that avoid discrimination and enable explanation.*

Explainable Artificial Intelligence (XAI)

XAI

EXPLAINABLE ARTIFICIAL INTELLIGENCE



David Gunning
DARPA/I2O
Program Update November 2017

DARPA

Why XAI?

REGULATION (EU) 2016/679 ... THE COUNCIL

on the p... ...ata and on the free
movem... ...tion Regulation)

In order to trust deployed AI systems, we must not only improve their robustness,[5] but also develop ways to make their reasoning intelligible. Intelligibility will help us spot AI that makes mistakes due to distributional drift or incomplete representations of goals and features. Intelligibility will also facilitate control by humans in increasingly common collaborative human/AI teams. Furthermore, intelligibility will help humans learn from AI. Finally, there are legal reasons to want intelligible AI, including the European GDPR and a growing need to assign liability when AI errs.

Weld & Bansal, CACM, Jun'19

European Union regulation ...
and a "righ...

Bryce Goodma...

TheVerge.com

...mpanies to check their

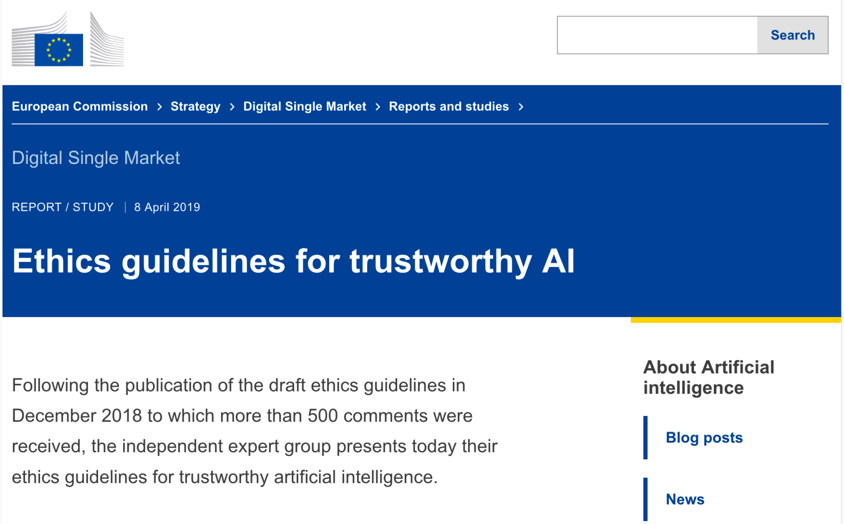
Algorithmic Accountability Act

■ We summarize the potential im... that the European Union's new Gen... Data Protection Regulation will have ... the routine use of machine-learni... algorithms. Slated to take effect as la... across the European Union in 2018, ... will place restrictions on automate... individual decision making (that is, algorithms that make decisions based on user-level predictors) that "significantly affect" users. When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them. We argue that while this law may pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks that avoid discrimination and enable explanation.

...ence (XAI)

...ber 2017

DARPA



Search

European Commission > Strategy > Digital Single Market > Reports and studies >

Digital Single Market

REPORT / STUDY | 8 April 2019

Ethics guidelines for trustworthy AI

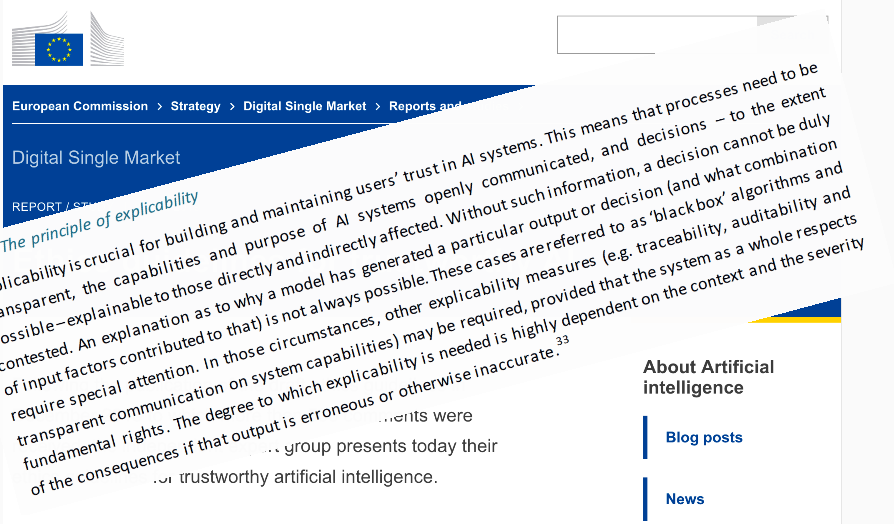
Following the publication of the draft ethics guidelines in December 2018 to which more than 500 comments were received, the independent expert group presents today their ethics guidelines for trustworthy artificial intelligence.

About Artificial intelligence

Blog posts

News

XAI & the principle of explicability



The principle of explicability

- Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.[33]

XAI & the principle of explicability

European Commission > Strategy > Digital Single Market > Reports and

Digital Single Market

REPORT / ST

- *The principle of explicability*

  Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.[33]

ments were

group presents today their

for trustworthy artificial intelligence.

About Artificial intelligence

Blog posts

News

& tens of recent papers!

Today's talk

Goals: Overview our work at intersection of AR & ML

- Part #1: Learning interpretable models (brief) [IPNM18, NIPM18]
  - Additional detail in our IJCAI'18 & IJCAR'18 papers
- Part #2: Rigorous explanations for black-box models [INM19a]
  - Additional detail in our AAAI'19 paper
- Part #3: Assessing heuristic explanations (brief) [NSM+19, INM19b]
  - Additional detail in our SAT'19 & CoRR'19 papers
- Part #4: Relating explanations with adversarial examples (brief) [INMS19]
  - Additional detail in our NIPS'19 paper

Part 1

Learning Interpretable ML Models

Outline

Background

Decision Sets

A Word on Decision Trees

Classification problems I

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Training data (or **examples**/instances): $\mathcal{E} = \{e_1, \ldots, e_M\}$

Classification problems I

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Training data (or **examples**/instances): $\mathcal{E} = \{e_1, \ldots, e_M\}$
- Binary **features**: $\mathcal{F} = \{f_1, \ldots, f_K\}$
  - Literals: f_r and $\neg f_r$

Classification problems I

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Training data (or **examples**/instances): $\mathcal{E} = \{e_1, \ldots, e_M\}$
- Binary **features**: $\mathcal{F} = \{f_1, \ldots, f_K\}$
  - Literals: f_r and $\neg f_r$
- **Feature space**: $\mathcal{U} \triangleq \prod_{r=1}^{K} \{f_r, \neg f_r\}$

Classification problems I

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Training data (or **examples**/instances): $\mathcal{E} = \{e_1, \ldots, e_M\}$
- Binary **features**: $\mathcal{F} = \{f_1, \ldots, f_K\}$
  - Literals: f_r and $\neg f_r$
- **Feature space**: $\mathcal{U} \triangleq \prod_{r=1}^{K} \{f_r, \neg f_r\}$
- Binary classification: $\mathcal{C} = \{c_0 = 0, c_1 = 1\}$
  - $\mathcal{E}$ partitioned into $\mathcal{E}^-$ and $\mathcal{E}^+$

Classification problems II

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- $e_q \in \mathcal{E}$ represented as a 2-tuple (π_q, ς_q)
  - $\pi_q \in \mathcal{U}$: literals associated with the example
  - $\varsigma_q \in \{0, 1\}$ is the class of example

Classification problems II

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- $e_q \in \mathcal{E}$ represented as a 2-tuple (π_q, ς_q)
  - $\pi_q \in \mathcal{U}$: literals associated with the example
  - $\varsigma_q \in \{0, 1\}$ is the class of example
- A literal l_r on a feature f_r, $l_r \in \{f_r, \neg f_r\}$, **discriminates** an example e_q if $\pi_q[r] = \neg l_r$
  - I.e. feature r takes the value **opposite** to the value in the tuple of literals of the example

Example

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Binary features: $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$
  - $f_1 \triangleq \mathsf{V}$, $f_2 \triangleq \mathsf{C}$, $f_3 \triangleq \mathsf{M}$, and $f_4 \triangleq \mathsf{E}$
- e_1 is represented by the 2-tuple (π_1, ς_1),
  - $\pi_1 = (\neg\mathsf{V}, \neg\mathsf{C}, \mathsf{M}, \neg\mathsf{E})$
  - $\varsigma_1 = 0$
- Literals V, C, ¬M and E **discriminate** e_1
- $\mathcal{U} = \{\mathsf{V}, \neg\mathsf{V}\} \times \{\mathsf{C}, \neg\mathsf{C}\} \times \{\mathsf{M}, \neg\mathsf{M}\} \times \{\mathsf{E}, \neg\mathsf{E}\}$

Goal of explainable classification – our take

Given training data, learn set of DNFs that correctly classify that data, perform suitably well on unseen data, and offer human-understandable explanations for the predictions made

Outline

Itemsets & decision sets

- Given $\mathcal{F}$, an **itemset** π is an element of $\mathcal{I} \triangleq \prod_{r=1}^{K} \{f_r, \neg f_r, \mathfrak{u}\}$
  - $\mathfrak{u}$ represents a don't care value

Itemsets & decision sets

- Given $\mathcal{F}$, an **itemset** π is an element of $\mathcal{I} \triangleq \prod_{r=1}^{K}\{f_r, \neg f_r, \mathfrak{u}\}$
  - $\mathfrak{u}$ represents a don't care value

- A **rule** is a 2-tuple (π, ς), with itemset $\pi \in \mathcal{I}$, and class $\varsigma \in \mathcal{C}$
  Rule (π, ς) interpreted as:

  IF all specified literals in π are true, THEN pick class ς

Itemsets & decision sets

- Given $\mathcal{F}$, an **itemset** π is an element of $\mathcal{I} \triangleq \prod_{r=1}^{K}\{f_r, \neg f_r, \mathfrak{u}\}$
  - $\mathfrak{u}$ represents a don't care value

- A **rule** is a 2-tuple (π, ς), with itemset $\pi \in \mathcal{I}$, and class $\varsigma \in \mathcal{C}$
  Rule (π, ς) interpreted as:

  IF all specified literals in π are true, THEN pick class ς

- A **decision set** $\mathbb{S}$ is a finite set of rules – unordered

Itemsets & decision sets

- Given $\mathcal{F}$, an **itemset** π is an element of $\mathcal{I} \triangleq \prod_{r=1}^{K}\{f_r, \neg f_r, \mathfrak{u}\}$
  - $\mathfrak{u}$ represents a don't care value

- A **rule** is a 2-tuple (π, ς), with itemset $\pi \in \mathcal{I}$, and class $\varsigma \in \mathcal{C}$
  Rule (π, ς) interpreted as:

  IF all specified literals in π are true, THEN pick class ς

- A **decision set** $\mathbb{S}$ is a finite set of rules – unordered

- A rule of the form $\mathfrak{D} \triangleq (\emptyset, \varsigma)$ denotes the **default rule** of a decision set $\mathbb{S}$
  - Default rule is **optional** and used **only** when other rules do not apply on some feature space point

Example

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Rule 1: $((\mathfrak{u}, \mathfrak{u}, \neg \mathrm{M}, \mathfrak{u}), c_1)$
  - Meaning: **if** ¬Meeting **then** Hike
- Rule 2: $((\neg \mathrm{V}, \mathfrak{u}, \mathfrak{u}, \mathfrak{u}), c_0)$
  - Meaning: **if** ¬Vacation **then** ¬Hike

Example

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Rule 1: $((\mathfrak{u}, \mathfrak{u}, \neg\mathsf{M}, \mathfrak{u}), \mathsf{c}_1)$
  - Meaning: **if** ¬Meeting **then** Hike
- Rule 2: $((\neg\mathsf{V}, \mathfrak{u}, \mathfrak{u}, \mathfrak{u}), \mathsf{c}_0)$
  - Meaning: **if** ¬Vacation **then** ¬Hike
- Default rule: $(\emptyset, \mathsf{c}_0)$
  - Meaning: if all other rules do not apply, then pick ¬Hike

Issue with unordered rules

- Itemsets $\pi_1, \pi_2 \in \mathcal{I}$ **clash**, $\pi_1 \cap \pi_2 = \emptyset$, if for some coordinate r:
  - $\pi_1[r] = f_r$ and $\pi_2[r] = \neg f_r$, or $\pi_1[r] = \neg f_r$ and $\pi_2[r] = f_r$

Issue with unordered rules – overlap

- Itemsets $\pi_1, \pi_2 \in \mathcal{I}$ **clash**, $\pi_1 \cap \pi_2 = \emptyset$, if for some coordinate r:
  - $\pi_1[r] = f_r$ and $\pi_2[r] = \neg f_r$, or $\pi_1[r] = \neg f_r$ and $\pi_2[r] = f_r$
- Two rules $r_1 = (\pi_1, \varsigma_1)$ and $r_2 = (\pi_2, \varsigma_2)$ **overlap** if π_1 and π_2 do not clash, i.e.

$$\pi_1 \cap \pi_2 \neq \emptyset$$

  - Can be restricted to some set, e.g. $\mathcal{E}$

Issue with unordered rules – overlap

- Itemsets $\pi_1, \pi_2 \in \mathcal{I}$ **clash**, $\pi_1 \cap \pi_2 = \emptyset$, if for some coordinate r:
  - $\pi_1[r] = f_r$ and $\pi_2[r] = \neg f_r$, or $\pi_1[r] = \neg f_r$ and $\pi_2[r] = f_r$
- Two rules $r_1 = (\pi_1, \varsigma_1)$ and $r_2 = (\pi_2, \varsigma_2)$ **overlap** if π_1 and π_2 do not clash, i.e.

$$\pi_1 \cap \pi_2 \neq \emptyset$$

  - Can be restricted to some set, e.g. $\mathcal{E}$
- Forms of overlap:
  - $\oplus$: overall where rules agree in prediction
  - $\ominus$: overlap where rules **disagree** in prediction

Issue with unordered rules – overlap

- Itemsets $\pi_1, \pi_2 \in \mathcal{I}$ **clash**, $\pi_1 \cap \pi_2 = \emptyset$, if for some coordinate r:
  - $\pi_1[r] = f_r$ and $\pi_2[r] = \neg f_r$, or $\pi_1[r] = \neg f_r$ and $\pi_2[r] = f_r$
- Two rules $r_1 = (\pi_1, \varsigma_1)$ and $r_2 = (\pi_2, \varsigma_2)$ **overlap** if π_1 and π_2 do not clash, i.e.

$$\pi_1 \cap \pi_2 \neq \emptyset$$

  - Can be restricted to some set, e.g. $\mathcal{E}$
- Forms of overlap:
  - $\oplus$: overall where rules agree in prediction
  - $\ominus$: overlap where rules **disagree** in prediction
- Our goal:

> Minimize number of rules in decision set, and provide guarantees in terms of overlap, namely $\ominus$-overlap

Example

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Decision set:

$$\{((\neg V, u, u, u), c_0), ((u, u, \neg M, u), c_1)\}$$

- No $\mathcal{E}^{\ominus}$-overlap

Example

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Decision set:

$$\{((\neg V, u, u, u), c_0), ((u, u, \neg M, u), c_1)\}$$

- No $\mathcal{E}^{\ominus}$-overlap
- But, there exists overlap in feature space
  - $\ominus$-overlap for $(\neg V, \neg C, \neg M, \neg E) \in \mathcal{U} \setminus \mathcal{E}$

Example

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

- Decision set:

$$\{((\neg V, u, u, u), c_0), ((u, u, \neg M, u), c_1)\}$$

- No $\mathcal{E}^{\ominus}$-overlap
- But, there exists overlap in feature space
  - $\ominus$-overlap for $(\neg V, \neg C, \neg M, \neg E) \in \mathcal{U} \setminus \mathcal{E}$
- However, there exists **no** $\mathcal{U}^{\ominus}$-overlap for decision set:

$$\{((V, u, u, u), c_1), ((\neg V, u, u, u), c_0)\}$$

Succinct explanations

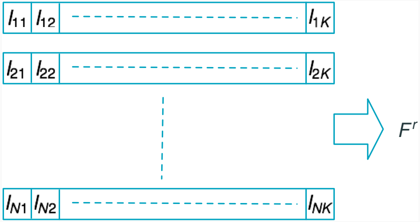
- If a rule fires, the set of literals represents the explanation for the predicted class
  - Explanation is succinct : only the literals in the rule used; independent of example
- For the default class, must pick one falsified literal in every rule that predicts a different class
  - Explanation is not succinct : explanation depends on each example
- Obs: Uninteresting to predict c_1 as negation of c_0 (and vice-versa)
  - Explanations also not succinct

Stating our goals

- Assumptions:
    - Represent $\mathcal{E}^-$ with Boolean function E^0
        - True for each example $\mathcal{E}^-$
    - Represent $\mathcal{E}^+$ with Boolean function E^1
        - True for each example $\mathcal{E}^+$
    - Also, let $E^0 \wedge E^1 \vDash \perp$

Stating our goals

- Assumptions:
  - Represent $\mathcal{E}^-$ with Boolean function E^0
    - True for each example $\mathcal{E}^-$
  - Represent $\mathcal{E}^+$ with Boolean function E^1
    - True for each example $\mathcal{E}^+$
  - Also, let $E^0 \wedge E^1 \vDash \perp$
- DNF functions to compute:
  - F^0 for predicting c_0, while **ensuring** $E^0 \vDash F^0$
  - F^1 for predicting c_1, while **ensuring** $E^1 \vDash F^1$



An ideal model – MinDS_0

- MinDS_0:

  Find the smallest DNF representations of Boolean functions F^0 and F^1, measured in the number of terms, such that:

  1. $E^0 \models F^0$
  2. $E^1 \models F^1$
  3. $F^1 \leftrightarrow F^0 \models \bot$

  - **No** $\mathcal{U}^{\ominus}$-overlap

An ideal model – MinDS_0

- MinDS_0:

  Find the smallest DNF representations of Boolean functions F^0 and F^1, measured in the number of terms, such that:

  1. $E^0 \models F^0$
  2. $E^1 \models F^1$
  3. $F^1 \leftrightarrow F^0 \models \perp$

  - No $\mathcal{U}^{\ominus}$-overlap

- **Obs:** MinDS_0 ensures **succinct** explanations
  - Computes F^0 and F^1 (i.e. **no** negation) and **no** default rule

An ideal model – MinDS_0

- MinDS_0:

  Find the smallest DNF representations of Boolean functions F^0 and F^1, measured in the number of terms, such that:

  1. $E^0 \vDash F^0$
  2. $E^1 \vDash F^1$
  3. $F^1 \leftrightarrow F^0 \vDash \perp$

  - No $\mathcal{U}^{\ominus}$-overlap

- **Obs:** MinDS_0 ensures **succinct** explanations
  - Computes F^0 and F^1 (i.e. **no** negation) and **no** default rule

- Complexity-wise:
  - $\mathsf{MinDS}_0 \in \Sigma_2^{\mathsf{P}}$
  - A conjecture: MinDS_0 hard for Σ_2^{P} (from late 2017)

Curbing our expectations I

- MinDS_4: Minimize F^0, given $F^1 \equiv E^1$ constant, and such that
  1. $E^0 \models F^0$
  2. $F^0 \wedge E^1 \models \bot$
     - No $\ominus$-overlap;
     - No succinct explanations for F^1

Curbing our expectations I

- MinDS$_4$: Minimize F^0, given $F^1 \equiv E^1$ constant, and such that
  1. $E^0 \models F^0$
  2. $F^0 \wedge E^1 \models \bot$
  - No $\ominus$-overlap;
  - No succinct explanations for F^1
- MinDS$_3$: Same as MinDS$_4$, but target F^1 given $F^0 \equiv E^0$ constant
  - Also, no $\ominus$-overlap;
  - No succinct explanations for F^0

Curbing our expectations I

- MinDS$_4$: Minimize F^0, given $F^1 \equiv E^1$ constant, and such that
  1. $E^0 \models F^0$
  2. $F^0 \wedge E^1 \models \bot$
  - No $\ominus$-overlap;
  - No succinct explanations for F^1
- MinDS$_3$: Same as MinDS$_4$, but target F^1 given $F^0 \equiv E^0$ constant
  - Also, no $\ominus$-overlap;
  - No succinct explanations for F^0
- MinDS$_2$: Minimize both F^0 and F^1, such that
  1. $E^0 \models F^0$
  2. $E^1 \models F^1$
  3. $F^0 \wedge E^1 \models \bot$
  4. $F^1 \wedge E^0 \models \bot$
  - Also, no $\mathcal{E}^{\ominus}$-overlap; but $(\mathcal{U} \setminus \mathcal{E})^{\ominus}$-overlap may exist
  - All explanations succinct

Curbing our expectations II

- MinDS$_1$: Minimize both F^0 and F^1, such that
  1. $E^0 \models F^0$
  2. $E^1 \models F^1$
  3. $F^1 \wedge F^0 \models \bot$
  - No $\mathcal{U}^{\ominus}$-overlap
  - Default rule may be required for points in $\mathcal{U} \setminus \mathcal{E}$
  - And, default rule explanations not succinct

Curbing our expectations II

- MinDS$_1$: Minimize both F^0 and F^1, such that
  1. $E^0 \vDash F^0$
  2. $E^1 \vDash F^1$
  3. $F^1 \wedge F^0 \vDash \bot$
  - No $\mathcal{U}^{\ominus}$-overlap
  - Default rule may be required for points in $\mathcal{U} \setminus \mathcal{E}$
  - And, default rule explanations not succinct

- Complexity-wise:
  - Decision formulations of MinDS$_1$, MinDS$_2$, MinDS$_3$, MinDS$_4$ are **complete** for **NP**
    - In principle, could be solved with MaxSAT
    - But no closed MaxSAT models for now

Experimental setup & initial results

- 49 datasets from the PMLB repository
- Assessment of MinDS_1, MinDS_2 and MP92, w/ and w/o SBPs
  - A basic model MP92 developed in the 90s [KKRR92]
  - We devised SBPs for the MinDS and the MP92 models
- Comparison with (state of the art) IDS [LBL16]
  - Heuristic approach, using smooth local search
  - Default settings & additional settings
- All experiments on an Intel Xeon E5-2630 2.60GHz processor with 64GB of memory, running Ubuntu Linux
  - Timeout of 600s and memout of 10GB

Experimental setup & initial results

- 49 datasets from the PMLB repository
- Assessment of MinDS_1, MinDS_2 and MP92, w/ and w/o SBPs
  - A basic model MP92 developed in the 90s [KKRR92]
  - We devised SBPs for the MinDS and the MP92 models
- Comparison with (state of the art) IDS [LBL16]
  - Heuristic approach, using smooth local search
  - Default settings & additional settings
- All experiments on an Intel Xeon E5-2630 2.60GHz processor with 64GB of memory, running Ubuntu Linux
  - Timeout of 600s and memout of 10GB

MP92	MP92+SBP	MinDS_2	MinDS_2+SBP	MinDS_1	MinDS_1+SBP	IDS-supp0.2	IDS-supp0.5
42	45	42	45	6	6	0	2

Outline

Background

Decision Sets

A Word on Decision Trees

Propositional encodings for DTs

- Proposed tight encoding for computing smallest decision tree [NIPM18]
  - Encoding also serves to **pick** the structure of the binary tree

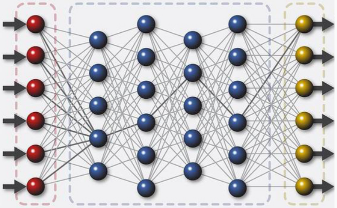
- Encoding much tighter (and more general) than earlier work [BHO09]

SAT	Weather	Mouse	Cancer	Car	Income
DT2*	27K	3.5M	92G	842M	354G
DT1	190K	1.2M	5.2M	4.1M	1.2G

- Several recent alternative proposals [VZ19, HRS19, VNP+19]
  - At least one outperforms our approach [VNP+19]

Part 2

Computing Rigorous Explanations

Component	Representation	Notes
	$\mathcal{C}$	Conjunction of literals, i.e. cube
	$\mathcal{F}$	Model encoding, e.g. SAT/SMT/CP/ILP/FOL
Cat	$\mathcal{E}$	Predicted class, i.e. literal

Relating with abduction

What we know $\mathcal{C} \land \mathcal{F} \models \mathcal{E}$

Relating with abduction

What we know	$\mathcal{C} \wedge \mathcal{F} \models \mathcal{E}$

Propositional Abduction	Hypotheses	$\mathcal{C}$
	Theory	$\mathcal{F}$
	Manifestation	$\mathcal{E}$
Goal	Find $\mathcal{C}_m \subseteq \mathcal{C}$, s.t.	$\mathcal{C}_m \wedge \mathcal{F} \not\models \perp \wedge \mathcal{C}_m \wedge \mathcal{F} \models \mathcal{E}$

Relating with abduction

What we know	$\mathcal{C} \land \mathcal{F} \vDash \mathcal{E}$

Propositional Abduction	Hypotheses	$\mathcal{C}$
	Theory	$\mathcal{F}$
	Manifestation	$\mathcal{E}$
Goal	Find $\mathcal{C}_m \subseteq \mathcal{C}$, s.t.	$\mathcal{C}_m \land \mathcal{F} \nvDash \perp \land \mathcal{C}_m \land \mathcal{F} \vDash \mathcal{E}$

But,	$\mathcal{C}_m \land \mathcal{F} \nvDash \perp$ is tautology
And,	$\mathcal{C}_m \land \mathcal{F} \vDash \mathcal{E}$ iff $\mathcal{C}_m \vDash \mathcal{F} \rightarrow \mathcal{E}$
Thus,	$\mathcal{C}_m$ is **prime implicant** of $\mathcal{F} \rightarrow \mathcal{E}$

Relating with abduction

What we know	$\mathcal{C} \wedge \mathcal{F} \vDash \mathcal{E}$

Propositional Abduction	Hypotheses	$\mathcal{C}$
	Theory	$\mathcal{F}$
	Manifestation	$\mathcal{E}$
Goal	Find $\mathcal{C}_m \subseteq \mathcal{C}$, s.t.	$\mathcal{C}_m \wedge \mathcal{F} \nvDash \perp \wedge \mathcal{C}_m \wedge \mathcal{F} \vDash \mathcal{E}$

But,	$\mathcal{C}_m \wedge \mathcal{F} \nvDash \perp$ is tautology
And,	$\mathcal{C}_m \wedge \mathcal{F} \vDash \mathcal{E}$ iff $\mathcal{C}_m \vDash \mathcal{F} \rightarrow \mathcal{E}$
Thus,	$\mathcal{C}_m$ is **prime implicant** of $\mathcal{F} \rightarrow \mathcal{E}$

We can compute **subset-**/**cardinality-**minimal (prime) implicants

Relating with abduction

What we know	$\mathcal{C} \wedge \mathcal{F} \vDash \mathcal{E}$

Propositional Abduction	Hypotheses	$\mathcal{C}$
	Theory	$\mathcal{F}$
	Manifestation	$\mathcal{E}$
Goal	Find $\mathcal{C}_m \subseteq \mathcal{C}$, s.t.	$\mathcal{C}_m \wedge \mathcal{F} \nvDash \perp \wedge \mathcal{C}_m \wedge \mathcal{F} \vDash \mathcal{E}$

Obs: For any instance consistent with $\mathcal{C}_m$, and given the model $\mathcal{F}$, the prediction is $\mathcal{E}$!

But,	$\mathcal{C}_m \wedge \mathcal{F} \nvDash \perp$ is tautology
And,	$\mathcal{C}_m \wedge \mathcal{F} \vDash \mathcal{E}$ iff $\mathcal{C}_m \vDash \mathcal{F} \rightarrow \mathcal{E}$
Thus,	$\mathcal{C}_m$ is **prime implicant** of $\mathcal{F} \rightarrow \mathcal{E}$

We can compute **subset-/cardinality**-minimal (prime) implicants – i.e. explanations!

Computing one subset-minimal explanation

Input: formula $\mathcal{F}$, input cube $\mathcal{C}$, prediction $\mathcal{E}$
Output: ***Subset-minimal*** explanation $\mathcal{C}_m \subseteq \mathcal{C}$

begin
    for $l \in \mathcal{C}$ **:**
        if $\text{Entails}(\mathcal{C} \setminus \{l\}, \mathcal{F} \rightarrow \mathcal{E})$ **:**
            $\mathcal{C} \leftarrow \mathcal{C} \setminus \{l\}$
    return $\mathcal{C}$
end

Computing one subset-minimal explanation

Input: formula $\mathcal{F}$, input cube $\mathcal{C}$, prediction $\mathcal{E}$
Output: ***Subset-minimal*** explanation $\mathcal{C}_m \subseteq \mathcal{C}$

```
begin
    for l ∈ C :
        if Entails(C \ {l}, F → E) :
            C ← C \ {l}
    return C
end
```

Computes some prime

Computing one cardinality-minimal explanation

Input: formula $\mathcal{F}$, input cube $\mathcal{C}$, prediction $\mathcal{E}$
Output: *Cardinality-minimal* explanation $\mathcal{C}_m \subseteq \mathcal{C}$

```
Γ ← ∅
while true do
    C_m ← MinimumHS(Γ)                    // Implicit hitting set dualization
    if Entails(C_m, F → E) :
        return C_m
    else:
        μ ← GetAssignment()
        C_T ← PickFalseLits(C \ C_m, μ)
        Γ ← Γ ∪ C_T
end
```

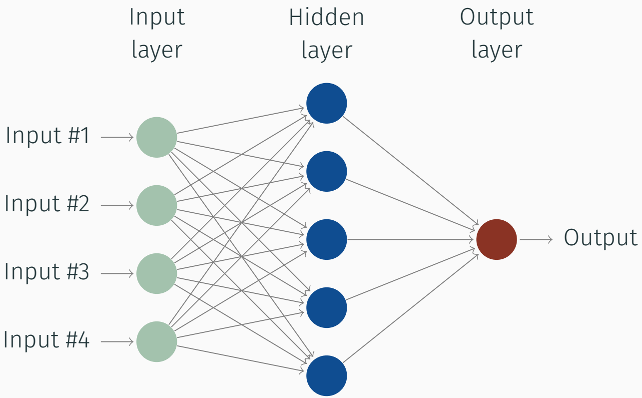
Computing one cardinality-minimal explanation

Input: formula $\mathcal{F}$, input cube $\mathcal{C}$, prediction $\mathcal{E}$
Output: *Cardinality-minimal* explanation $\mathcal{C}_m \subseteq \mathcal{C}$

```
Γ ← ∅
while true do
    C_m ← MinimumHS(Γ)                          // Implicit hitting set dualization
    if Entails(C_m, F → E) :
        return C_m
    else:
        μ ← GetAssignment()
        C_T ← PickFalseLits(C \ C_m, μ)
        Γ ← Γ ∪ C_T
end
```

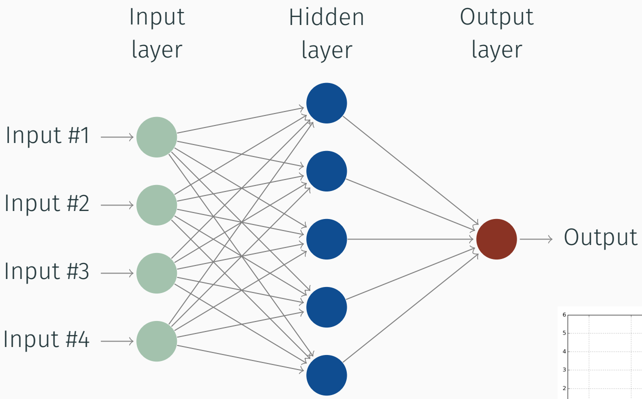
Computes **smallest** prime

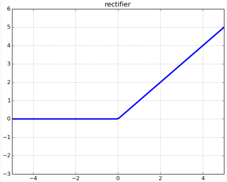
Encodings NNs



- Each layer (except first) viewed as a **block**
  - Compute $\mathbf{x}'$ given input $\mathbf{x}$, weights matrix $\mathbf{A}$, and bias vector $\mathbf{b}$
  - Compute output $\mathbf{y}$ given $\mathbf{x}'$ and activation function

Encodings NNs





- Each layer (except first) viewed as a **block**
  - Compute $\mathbf{x}'$ given input $\mathbf{x}$, weights matrix $\mathbf{A}$, and bias vector $\mathbf{b}$
  - Compute output $\mathbf{y}$ given $\mathbf{x}'$ and activation function
- Each unit uses a ReLU activation function

[NH10]

Encoding NNs using MILP

Computation for a NN ReLU **block**:

$$\mathbf{x}' = \mathbf{A} \cdot \mathbf{x} + \mathbf{b}$$
$$\mathbf{y} = \max(\mathbf{x}', \mathbf{0})$$

Encoding NNs using MILP

Computation for a NN ReLU **block**:

$$\mathbf{x}' = \mathbf{A} \cdot \mathbf{x} + \mathbf{b}$$
$$\mathbf{y} = \max(\mathbf{x}', \mathbf{0})$$

Encoding each **block**: [FJ18]

$$\sum_{j=1}^{n} a_{i,j} x_j + b_i = y_i - s_i$$
$$z_i = 1 \rightarrow y_i \leq 0$$
$$z_i = 0 \rightarrow s_i \leq 0$$
$$y_i \geq 0, s_i \geq 0, z_i \in \{0, 1\}$$

Simpler encodings exist, but **not** as effective [KBD+17]

Sample of experimental results

Dataset			Minimal explanation			Minimum explanation		
			size	SMT (s)	MILP (s)	size	SMT (s)	MILP (s)
australian	(14)	m	1	0.03	0.05	—	—	—
		a	8.79	1.38	0.33	—	—	—
		M	14	17.00	1.43	—	—	—
backache	(32)	m	13	0.13	0.14	—	—	—
		a	19.28	5.08	0.85	—	—	—
		M	26	22.21	2.75	—	—	—
breast-cancer	(9)	m	3	0.02	0.04	3	0.02	0.03
		a	5.15	0.65	0.20	4.86	2.18	0.41
		M	9	6.11	0.41	9	24.80	1.81
cleve	(13)	m	4	0.05	0.07	4	—	0.07
		a	8.62	3.32	0.32	7.89	—	5.14
		M	13	60.74	0.60	13	—	39.06
hepatitis	(19)	m	6	0.02	0.04	4	0.01	0.04
		a	11.42	0.07	0.06	9.39	4.07	2.89
		M	19	0.26	0.20	19	27.05	22.23
voting	(16)	m	3	0.01	0.02	3	0.01	0.02
		a	4.56	0.04	0.13	3.46	0.3	0.25
		M	11	0.10	0.37	11	1.25	1.77
spect	(22)	m	3	0.02	0.02	3	0.02	0.04
		a	7.31	0.13	0.07	6.44	1.61	0.67
		M	20	0.88	0.29	20	8.97	10.73

Part 3

Assessing Heuristic Explanations

Computing heuristic explanations

- Many (highly visible) heuristic explanation approaches:
  - LIME [RSG16]
  - SHAP [LL17]
  - Anchor [RSG18]
  - ...

Computing heuristic explanations

- Many (highly visible) heuristic explanation approaches:
  - LIME [RSG16]
  - SHAP [LL17]
  - Anchor [RSG18]
  - ...
- **Q:** How to assess the quality of heuristic explanations? [NSM+19, INM19b]

What is the global quality of heuristic explanations in light of computed local explanations?

Approach

- Learn ML model
  - Focused on boosted trees obtained with XGBoost

Approach

- Learn ML model
  - Focused on boosted trees obtained with XGBoost
- Compute local explanation for some instance

Approach

- Learn ML model
  - Focused on boosted trees obtained with XGBoost
- Compute local explanation for some instance
- Use our abduction-based approach to assess whether local explanation holds globally, and

Approach

- Learn ML model
  - Focused on boosted trees obtained with XGBoost
- Compute local explanation for some instance
- Use our abduction-based approach to assess whether local explanation holds globally, and
  1. If it does **not** (i.e. it's optimistic), then **fix** it

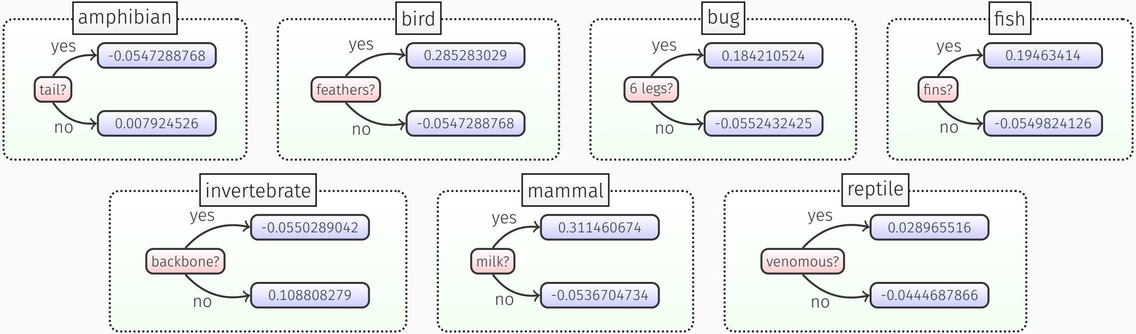
Approach

- Learn ML model
  - Focused on boosted trees obtained with XGBoost
- Compute local explanation for some instance
- Use our abduction-based approach to assess whether local explanation holds globally, and
  1. If it does **not** (i.e. it's optimistic), then **fix** it
  2. It it holds but has redundant literals (i.e. it's pessimistic), then refine it

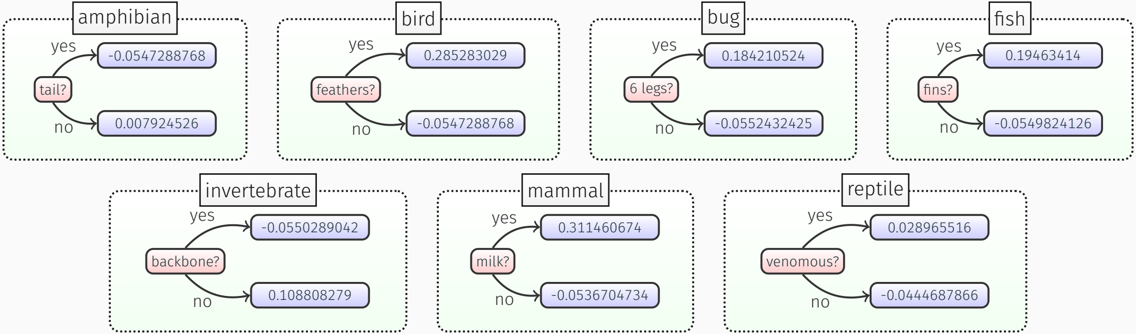
Approach

- Learn ML model
  - Focused on boosted trees obtained with XGBoost
- Compute local explanation for some instance
- Use our abduction-based approach to assess whether local explanation holds globally, and
  1. If it does **not** (i.e. it's optimistic), then **fix** it
  2. It it holds but has redundant literals (i.e. it's pessimistic), then refine it
  3. Otherwise, report the local explanation as a global explanation

An example – zoo dataset



An example – zoo dataset

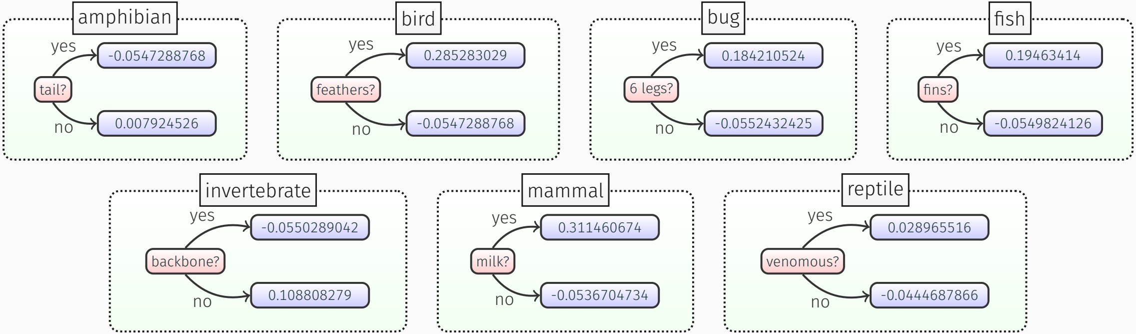


- Example instance:

IF $(\text{animal\_name} = \text{pitviper}) \wedge \neg\text{hair} \wedge \neg\text{feathers} \wedge \text{eggs} \wedge \neg\text{milk} \wedge \neg\text{airborne} \wedge \neg\text{aquatic} \wedge \text{predator} \wedge \neg\text{toothed} \wedge \text{backbone} \wedge \text{breathes} \wedge \text{venomous} \wedge \neg\text{fins} \wedge (\text{legs} = 0) \wedge \text{tail} \wedge \neg\text{domestic} \wedge \neg\text{catsize}$

THEN $(\text{class} = \text{reptile})$

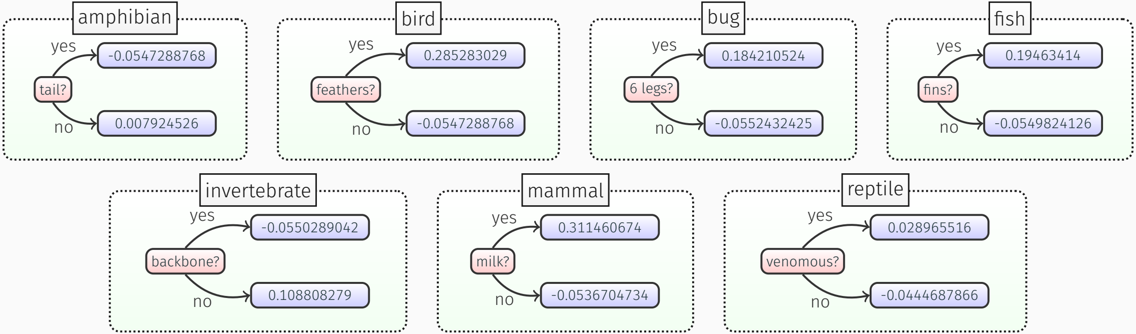
An example – zoo dataset



- Explanation obtained with Anchor [RSG18]

IF	$\neg\text{hair} \wedge \neg\text{milk} \wedge \neg\text{toothed} \wedge \neg\text{fins}$
THEN	$(\text{class} = \text{reptile})$

An example – zoo dataset



- But, explanation **incorrectly** holds on another instance (from training data)

IF (animal_name = toad) ∧ ¬***hair*** ∧ ¬feathers ∧ eggs ∧ ¬***milk*** ∧ ¬airborne ∧ ¬aquatic ∧ ¬predator ∧ ¬***toothed*** ∧ backbone ∧ breathes ∧ ¬venomous ∧ ¬***fins*** ∧ (legs = 4) ∧ ¬tail ∧ ¬domestic ∧ ¬catsize

THEN (class = amphibian)

Some results

Dataset	(# unique)	Explanations					
		optimistic		pessimistic		realistic	
		LIME	Anchor	LIME	Anchor	LIME	Anchor
adult	(5579)	61.3%	80.5%	7.9%	1.6%	30.8%	17.9%
lending	(4414)	24.0%	3.0%	0.4%	0.0%	75.6%	97.0%
recidivism	(3696)	94.1%	99.4%	4.6%	0.4%	1.3%	0.2%
compas	(778)	71.9%	84.4%	20.6%	1.7%	7.5%	13.9%
german	(1000)	85.3%	99.7%	14.6%	0.2%	0.1 %	0.1 %

Some results

Dataset	(# unique)	Explanations					
		optimistic		pessimistic		realistic	
		LIME	Anchor	LIME	Anchor	LIME	Anchor
adult	(5579)	61.3%	80.5%	7.9%	1.6%	30.8%	17.9%
lending	(4414)	24.0%	3.0%	0.4%	0.0%	75.6%	97.0%
recidivism	(3696)	94.1%	99.4%	4.6%	0.4%	1.3%	0.2%
compas	(778)	71.9%	84.4%	20.6%	1.7%	7.5%	13.9%
german	(1000)	85.3%	99.7%	14.6%	0.2%	0.1%	0.1%

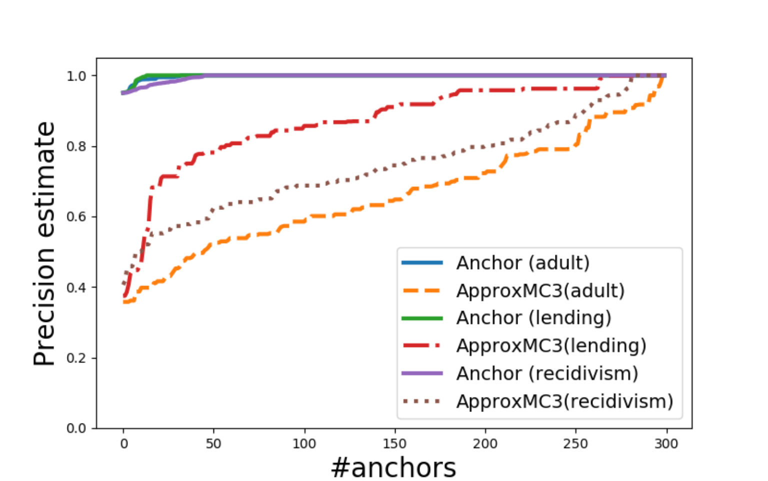
& Google XAI service
most likely similar...

How often are local explanations consistent with prediction?

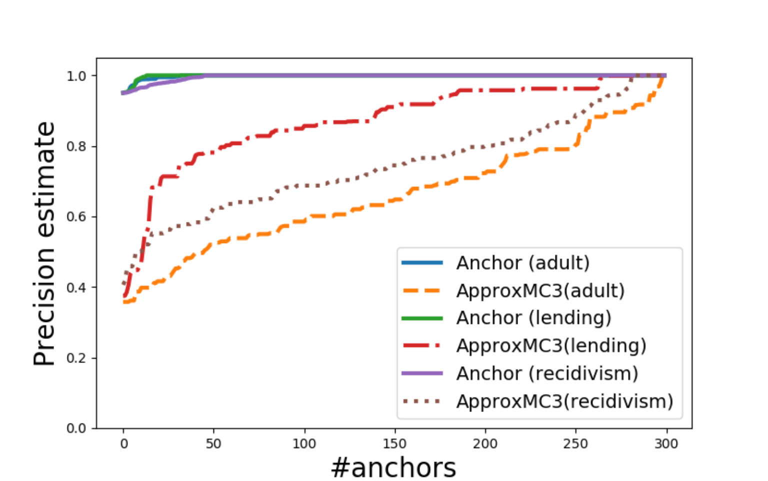
Approach

- Exploit ML model with SAT-based encoding
    - In our case: used binarized neural networks (BNNs)
- Compute local explanations with Anchor (similar results with LIME or SHAP)
- Use (approximate) model counter to assess how often explanation is consistent with prediction

Preliminary results



- Anchor often claims $\approx 99\%$ precision



- Anchor often claims $\approx 99\%$ precision; out results **demonstrate** otherwise

Part 4

Explanations vs. Adversarial Examples

- Vast body of work on computing explanations (XPs)
  - Mostly heuristic approaches, with recent rigorous solutions

- Vast body of work on computing explanations (XPs)
  - Mostly heuristic approaches, with recent rigorous solutions

- Vast body of work on coping with adversarial examples (AEs)
  - Both heuristic and rigorous approaches

Overview

- Vast body of work on computing explanations (XPs)
  - Mostly heuristic approaches, with recent rigorous solutions
- Vast body of work on coping with adversarial examples (AEs)
  - Both heuristic and rigorous approaches
- Can XPs and AEs be somehow related?

Overview

- Vast body of work on computing explanations (XPs)
  - Mostly heuristic approaches, with recent rigorous solutions
- Vast body of work on coping with adversarial examples (AEs)
  - Both heuristic and rigorous approaches
- Can XPs and AEs be somehow related?
  - Recent work observed that some connection existed, but formal connection has been elusive

- Vast body of work on computing explanations (XPs)
  - Mostly heuristic approaches, with recent rigorous solutions
- Vast body of work on coping with adversarial examples (AEs)
  - Both heuristic and rigorous approaches
- Can XPs and AEs be somehow related?
  - Recent work observed that some connection existed, but formal connection has been elusive
- We recently proposed a (first) link between XPs and AEs [INMS19]

- Vast body of work on computing explanations (XPs)
  - Mostly heuristic approaches, with recent rigorous solutions
- Vast body of work on coping with adversarial examples (AEs)
  - Both heuristic and rigorous approaches
- Can XPs and AEs be somehow related?
  - Recent work observed that some connection existed, but formal connection has been elusive
- We recently proposed a (first) link between XPs and AEs [INMS19]
  - The work exploits hitting set duality, first studied in model-based diagnosis [Rei87]

A well-known example

[RN10]

Example	Input Attributes										Goal
	Alt	Bar	Fri	Hun	Pat	Price	Rain	Res	Type	Est	WillWait
x_1	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	$y_1 =$ Yes
x_2	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	$y_2 =$ No
x_3	No	Yes	No	No	Some	$	No	No	Burger	0–10	$y_3 =$ Yes
x_4	Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10–30	$y_4 =$ Yes
x_5	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	$y_5 =$ No
x_6	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	$y_6 =$ Yes
x_7	No	Yes	No	No	None	$	Yes	No	Burger	0–10	$y_7 =$ No
x_8	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	$y_8 =$ Yes
x_9	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	$y_9 =$ No
x_{10}	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	$y_{10} =$ No
x_{11}	No	No	No	No	None	$	No	No	Thai	0–10	$y_{11} =$ No
x_{12}	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	$y_{12} =$ Yes

A well-known example (Cont.)

- 10 features:

$$\{\text{A(lternate)}, \text{B(ar)}, \text{W(eekend)}, \text{H(ungry)}, \text{Pa(trons)}, \text{Pr(ice)}, \text{Ra(in)}, \text{Re(serv.)}, \text{T(ype)}, \text{E(stim.)}\}$$

- Example instance (x_1, with outcome $y_1 = \text{Yes}$):

$$\{\text{A}, \neg\text{B}, \neg\text{W}, \text{H}, (\text{Pa} = \text{Some}), (\text{Pr} = \$\$\$), \neg\text{Ra}, \text{Re}, (\text{T} = \text{French}), (\text{E} = 0\text{–}10)\}$$

- A possible decision set (obtained with some off-the-shelf tool, & function*):

$$
\begin{array}{llllr}
\textbf{IF} & (\text{Pa} = \text{Some}) \land \neg(\text{E} = {>}60) & \textbf{THEN} & (\text{Wait} = \text{Yes}) & (\text{R1}) \\
\textbf{IF} & \text{W} \land \neg(\text{Pr} = \$\$\$) \land \neg(\text{E} = {>}60) & \textbf{THEN} & (\text{Wait} = \text{Yes}) & (\text{R2}) \\
\textbf{IF} & \neg\text{W} \land \neg(\text{Pa} = \text{Some}) & \textbf{THEN} & (\text{Wait} = \text{No}) & (\text{R3}) \\
\textbf{IF} & (\text{E} = {>}60) & \textbf{THEN} & (\text{Wait} = \text{No}) & (\text{R4}) \\
\textbf{IF} & \neg(\text{Pa} = \text{Some}) \land (\text{Pr} = \$\$\$) & \textbf{THEN} & (\text{Wait} = \text{No}) & (\text{R5})
\end{array}
$$

Counterexamples & breaks

Counterexamples & breaks

- Counterexamples:

  A subset-minimal set $\mathcal{C}$ of literals is a counterexample (CEx) to a prediction π, if $\mathcal{C} \vDash (\mathcal{M} \rightarrow \rho)$, with $\rho \in \mathbb{K} \wedge \rho \neq \pi$

Counterexamples & breaks

- Counterexamples:

  A subset-minimal set $\mathcal{C}$ of literals is a counterexample (CEx) to a prediction π, if $\mathcal{C} \vDash (\mathcal{M} \rightarrow \rho)$, with $\rho \in \mathbb{K} \wedge \rho \neq \pi$

- Breaks:

  A literal τ_i breaks a set of literals $\mathcal{S}$ (each denoting a different feature) if $\mathcal{S}$ contains a literal inconsistent with τ_i

Counterexamples & breaks

- Counterexamples:

  A subset-minimal set $\mathcal{C}$ of literals is a counterexample (CEx) to a prediction π, if $\mathcal{C} \vDash (\mathcal{M} \rightarrow \rho)$, with $\rho \in \mathbb{K} \wedge \rho \neq \pi$

- Breaks:

  A literal τ_i breaks a set of literals $\mathcal{S}$ (each denoting a different feature) if $\mathcal{S}$ contains a literal inconsistent with τ_i

- Back to the example, consider prediction $(\text{Wait} = \text{Yes})$:

Counterexamples & breaks

- Counterexamples:

  A subset-minimal set $\mathcal{C}$ of literals is a counterexample (CEx) to a prediction π, if $\mathcal{C} \vDash (\mathcal{M} \rightarrow \rho)$, with $\rho \in \mathbb{K} \land \rho \neq \pi$

- Breaks:

  A literal τ_i breaks a set of literals $\mathcal{S}$ (each denoting a different feature) if $\mathcal{S}$ contains a literal inconsistent with τ_i

- Back to the example, consider prediction $(\mathsf{Wait} = \mathsf{Yes})$:
  - Using (R1) (and assuming a consistent instance), an explanation is:

$$(\mathsf{Pa} = \mathsf{Some}) \land \neg(\mathsf{E} = {>}60)$$

Counterexamples & breaks

- Counterexamples:

  A subset-minimal set $\mathcal{C}$ of literals is a counterexample (CEx) to a prediction π, if $\mathcal{C} \vDash (\mathcal{M} \to \rho)$, with $\rho \in \mathbb{K} \land \rho \neq \pi$

- Breaks:

  A literal τ_i breaks a set of literals $\mathcal{S}$ (each denoting a different feature) if $\mathcal{S}$ contains a literal inconsistent with τ_i

- Back to the example, consider prediction $(\text{Wait} = \text{Yes})$:
  - Using (R1) (and assuming a consistent instance), an explanation is:

$$(\text{Pa} = \text{Some}) \land \neg(\text{E} = {>}60)$$

  - Due to (R5), a counterexample is:

$$\neg(\text{Pa} = \text{Some}) \land (\text{Pr} = \$\$\$)$$

Counterexamples & breaks

- Counterexamples:

  A subset-minimal set $\mathcal{C}$ of literals is a counterexample (CEx) to a prediction π, if $\mathcal{C} \vDash (\mathcal{M} \rightarrow \rho)$, with $\rho \in \mathbb{K} \wedge \rho \neq \pi$

- Breaks:

  A literal τ_i breaks a set of literals $\mathcal{S}$ (each denoting a different feature) if $\mathcal{S}$ contains a literal inconsistent with τ_i

- Back to the example, consider prediction $(\text{Wait} = \text{Yes})$:
  - Using (R1) (and assuming a consistent instance), an explanation is:

    $$(\text{Pa} = \text{Some}) \wedge \neg(\text{E} = {>}60)$$

  - Due to (R5), a counterexample is:

    $$\neg(\text{Pa} = \text{Some}) \wedge (\text{Pr} = \$\$\$)$$

  - XP $\mathcal{S}_1 = \{(\text{Pa} = \text{Some}), \neg(\text{E} = {>}60)\}$ breaks CEx $\mathcal{S}_2 = \{\neg(\text{Pa} = \text{Some}), (\text{Pr} = \$\$\$)\}$ and vice-versa

Some preliminary results

1. Relationship between XPs with CEx's:

Some preliminary results

1. Relationship between XPs with CEx’s:
    - Each XP breaks every CEx

Some preliminary results

1. Relationship between XPs with CEx's:
    - Each XP breaks every CEx
    - Each CEx breaks every XP

Some preliminary results

1. Relationship between XPs with CEx's:
    - Each XP breaks every CEx
    - Each CEx breaks every XP

$\therefore$ XPs can be computed from all CEx's (by HSD) and vice-versa

Some preliminary results

1. Relationship between XPs with CEx's:
    - Each XP breaks every CEx
    - Each CEx breaks every XP

    $\therefore$ XPs can be computed from all CEx's (by HSD) and vice-versa

2. Given instance $\mathcal{I}$, an AE can be computed from closest CEx

Revisiting the example

- Restaurant dataset
- ML model is decision set (shown earlier)
- Prediction is $(\text{Wait} = \text{Yes})$

- Global explanations:
  1. $(\text{Pa} = \text{Some}) \land \neg(\text{E} = {>}60)$
  2. $\text{W} \land \neg(\text{Pr} = \$\$\$) \land \neg(\text{E} = {>}60)$

- Counterexamples:
  1. $\neg\text{W} \land \neg(\text{Pa} = \text{Some})$
  2. $(\text{E} = {>}60)$
  3. $\neg(\text{Pa} = \text{Some}) \land (\text{Pr} = \$\$\$)$

- The XP's break the CEx's and vice-versa

Conclusions & roadmap

- Glimpse of work on learning interpretable ML models (using SAT)
  - Smallest decision trees & decision sets
- New approach for finding explanations of black-box models by computing prime implicants (using ILP&SMT)
  - Results for NNs and for BTs
- Hitting set duality between explanations and counterexamples
  - Can compute CEx's from XP's and AEs from CEx's

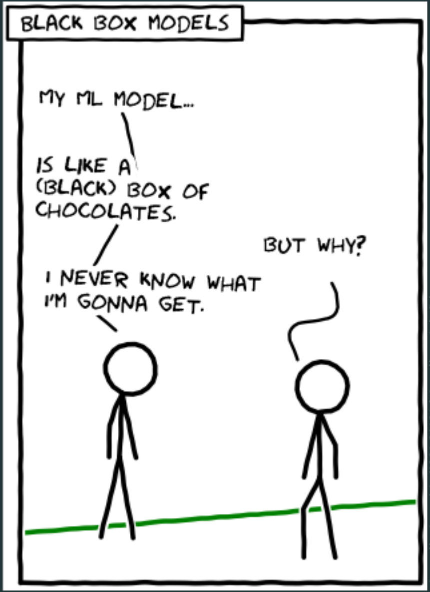
Conclusions & roadmap

- Glimpse of work on learning interpretable ML models (using SAT)
  - Smallest decision trees & decision sets
- New approach for finding explanations of black-box models by computing prime implicants (using ILP&SMT)
  - Results for **NNs** and for **BTs**
- Hitting set duality between explanations and counterexamples
  - Can compute CEx's from XP's and AEs from CEx's
- Our remit **@ ANITI**:

  To explain, to verify & to learn ML models

  with guarantees of **rigor**, by using AR tools & techniques

Questions?



https://arxiv.org/abs/1901.01686 & http://cmx.io/edit/

[BHO09] Christian Bessiere, Emmanuel Hebrard, and Barry O'Sullivan.
Minimising decision tree size as combinatorial optimisation.
In *CP*, pages 173–187, 2009.

[FJ18] Matteo Fischetti and Jason Jo.
Deep neural networks and mixed integer linear optimization.
Constraints, 23(3):296–309, 2018.

[HRS19] Xiyang Hu, Cynthia Rudin, and Margo I. Seltzer.
Optimal sparse decision trees.
CoRR, abs/1904.12847, 2019.

[INM19a] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva.
Abduction-based explanations for machine learning models.
In *AAAI*, pages 1511–1519, 2019.

[INM19b] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva.
On validating, repairing and refining heuristic ML explanations.
CoRR, abs/1907.02509, 2019.

[INMS19] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva.
On relating explanations and adversarial examples.
In *NeuIPS*, 2019.

References ii

[IPNM18] Alexey Ignatiev, Filipe Pereira, Nina Narodytska, and João Marques-Silva.
A SAT-based approach to learn explainable decision sets.
In *IJCAR*, pages 627–645, 2018.

[KBD+17] Guy Katz, Clark W. Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer.
Reluplex: An efficient SMT solver for verifying deep neural networks.
In *CAV*, pages 97–117, 2017.

[KKRR92] Anil P. Kamath, Narendra Karmarkar, K. G. Ramakrishnan, and Mauricio G. C. Resende.
A continuous approach to inductive inference.
Math. Program., 57:215–238, 1992.

[LBL16] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec.
Interpretable decision sets: A joint framework for description and prediction.
In *KDD*, pages 1675–1684, 2016.

[LL17] Scott M. Lundberg and Su-In Lee.
A unified approach to interpreting model predictions.
In *NIPS*, pages 4765–4774, 2017.

[NH10] Vinod Nair and Geoffrey E. Hinton.
Rectified linear units improve restricted boltzmann machines.
In *ICML*, pages 807–814, 2010.

[NIPM18] Nina Narodytska, Alexey Ignatiev, Filipe Pereira, and Joao Marques-Silva.
Learning optimal decision trees with SAT.
In *IJCAI*, pages 1362–1368, 2018.

[NSM+19] Nina Narodytska, Aditya A. Shrotri, Kuldeep S. Meel, Alexey Ignatiev, and Joao Marques-Silva.
Assessing heuristic machine learning explanations with model counting.
In *SAT*, pages 267–278, 2019.

[Rei87] Raymond Reiter.
A theory of diagnosis from first principles.
Artif. Intell., 32(1):57–95, 1987.

[RN10] Stuart J. Russell and Peter Norvig.
***Artificial Intelligence - A Modern Approach*.**
Pearson Education, 2010.

[RSG16] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin.
”why should I trust you?”: Explaining the predictions of any classifier.
In *KDD*, pages 1135–1144, 2016.

[RSG18] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin.
Anchors: High-precision model-agnostic explanations.
In *AAAI*, pages 1527–1535. AAAI Press, 2018.

[VNP+19] Helene Verhaeghe, Siegfried Nijssen, Gilles Pesant, Claude-Guy Quimper, and Pierre Schaus.
Learning optimal decision trees using constraint programming.
In *CP*, 2019.

[VZ19] Sicco Verwer and Yingqian Zhang.
Learning optimal classification trees using a binary linear program formulation.
In *AAAI*, pages 1625–1632, 2019.