Reasoning-Based Learning of Interpretable ML Models
Alexey Ignatiev¹, Joao Marques-Silva², Nina Narodytska³, and Peter J. Stuckey¹
¹Monash University, Melbourne, Australia  ²IRIT, CNRS, Toulouse, France  ³VMware Research, CA, USA

Why Status Quo

Interpretability Issue

Perfect and sparse DTs

Decision sets

Decision trees

Interpretable Models

Rule-based models

“transparent” and easy to interpret

come in handy in XAI

Perfect and sparse DLs and DSs

Same Issue with DL Interpretablity

Additional remarks 1

- Comparing to heuristic methods
  - higher accuracy but
  - higher training time
  - evolution of reasoning methods!

- Other interpretable models
  - learning OBDDs
  - SAT-based inference

- Perfect vs. sparse models
  - pros of perfect models:
    - highest possible accuracy
    - pros of sparse models:
      - smaller size
      - easier to compute
      - smaller explanations

Additional remarks 2

- Model expressivity and size
  - DLs are more succinct than DTs
  - how to categorise DSs?
  - empirically, less succinct than DLs!
  - a special case of DSs

- Fairness and other constraints
  - model properties can be enforced
  - in the form of constraints
  - easy to plug in!

- Fairness constraints
  - learning fair DTs and DSs
  - accuracy vs. fairness

- Intepretability
  - empirical considerations:
    - [XP] for perfect DLs < [XP] for perfect DSs
    - [XP] for sparse DLs < [XP] for sparse DSs

- DTs and DLs may be uninterpretable
  - AXPs for DTs — in polytime!
  - not the case for DLs and DSs!

Reasoning-based approaches to DTs

Rule-based approaches to DLs and DSs

DT Interpretablity Issue

Instance v = (1, 0, 1, 1), i.e. 4 literals in the path
actual explanation x₁ = 1 ∧ x₂ = 1, i.e. 2 literals

Instance v = (1, 0, 1, 1), i.e. rule R₅ fires the prediction
actual AXp: x₃ = 1 ∧ x₄ = 1, i.e. 2 literals

Perfect and sparse DTs

sparse DT for Titanic dataset
(training accuracy 79.25%)

Perfect and sparse DLs

smaller perfect DL for Titanic dataset
(training accuracy 79.25%)

Same Issue with DL Interpretablity

Additional remarks 1

- Comparing to heuristic methods
  - higher accuracy but
  - higher training time
  - evolution of reasoning methods!

- Other interpretable models
  - learning OBDDs
  - SAT-based inference

- Perfect vs. sparse models
  - pros of perfect models:
    - highest possible accuracy
    - pros of sparse models:
      - smaller size
      - easier to compute
      - smaller explanations

Additional remarks 2

- Model expressivity and size
  - DLs are more succinct than DTs
  - how to categorise DSs?
  - empirically, less succinct than DLs!
  - a special case of DSs

- Fairness and other constraints
  - model properties can be enforced
  - in the form of constraints
  - easy to plug in!

- Fairness constraints
  - learning fair DTs and DSs
  - accuracy vs. fairness

- Intepretability
  - empirical considerations:
    - [XP] for perfect DLs < [XP] for perfect DSs
    - [XP] for sparse DLs < [XP] for sparse DSs

- DTs and DLs may be uninterpretable
  - AXPs for DTs — in polytime!
  - not the case for DLs and DSs!