Optimising over Machine Learning Tree-Based Models

Miten Mistry[1], Dimitrios Letsios[1], Gerhard Krennrich[2], Robert M. Lee[2], and Ruth Misener[1]

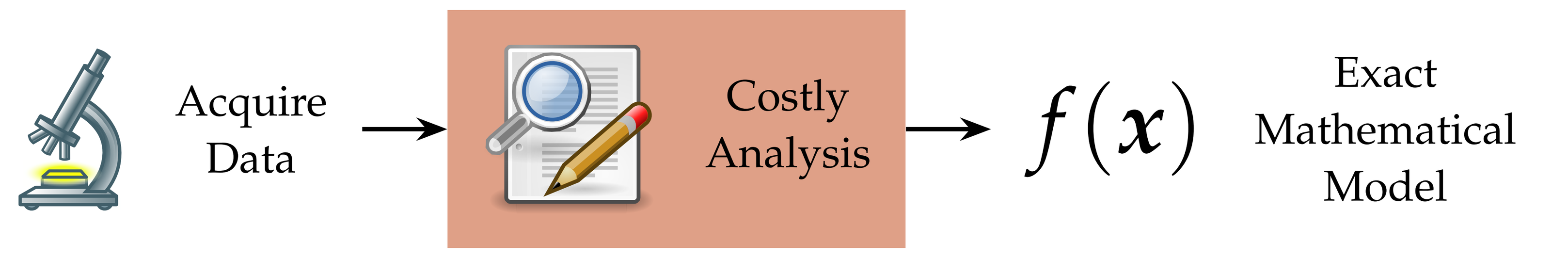
[1]Imperial College London, UK [2]BASF SE, Germany

miten.mistry11@imperial.ac.uk

STEM for Britain 2019

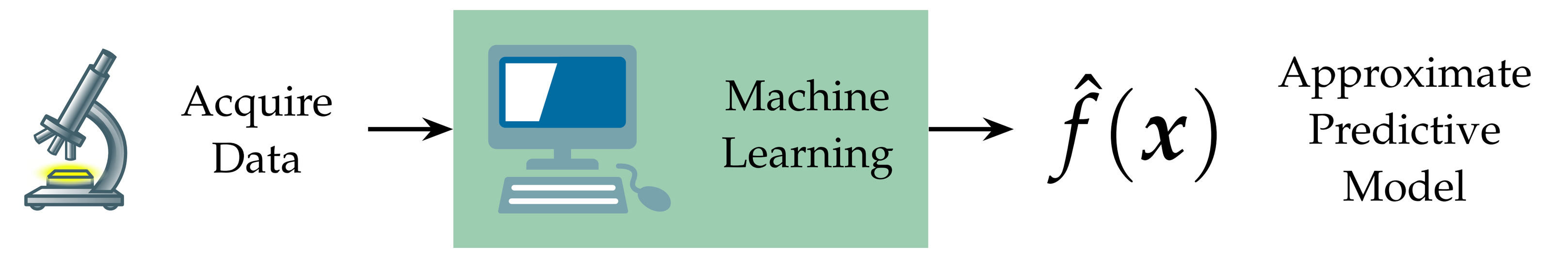
Mathematical Modelling: Exact Data Analysis

Data can involve many variables, contain measurement errors, or may be very large. Deriving an exact mathematical model for the data is a costly task requiring expertise and time. These mathematical models are derived from our understanding of the physical world. Exact mathematical models are readily incorporated into decision-making problems.



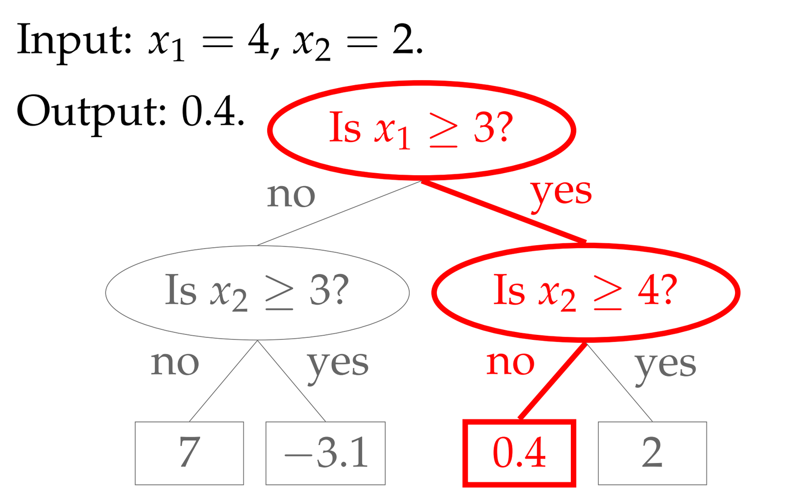
Machine Learning: Approximate Data Analysis

Machine learning approximates how data behaves without needing contextual knowledge. Machine learning is cheaper than exact analysis. While machine learning models are effective, they tend to be complex. We question how to integrate and manage complex machine learning models into larger decision-making problems?



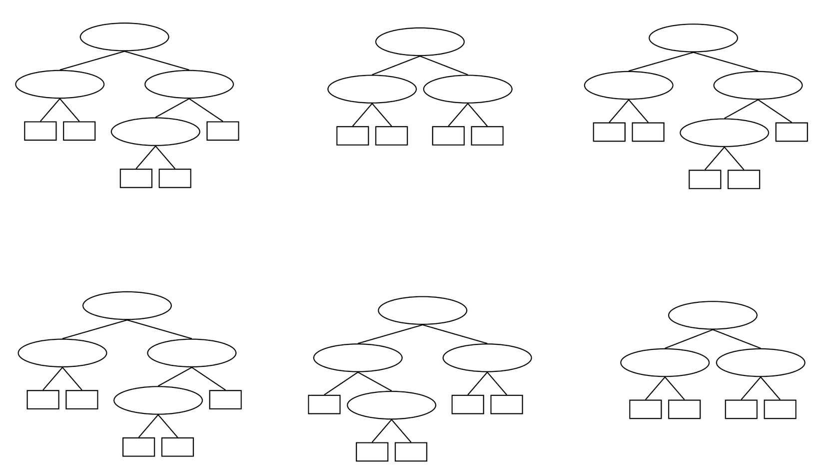
Decision Trees: Explainable Machine Learning

Decision Tree



Query a sequence of yes/no questions to find prediction. Sequence of yes/no responses explain a decision tree's prediction.

Tree-Based Model



Contains several decision trees that collaboratively outperform a single decision tree. Prediction sums all decision tree responses.

Applications

Catalysis

Catalysts speed up chemical reactions and are essential for energy efficiency at BASF. BASF finds tree-based models effective for modelling catalyst behaviour. Developing the best-performing catalysts requires optimising over the BASF tree-based models.

Concrete Mixture Design

Concrete is a fundamental building material that obtains different properties dependent on ingredient proportions. Tree-based models can predict concrete properties and offer explanations for the prediction. We can repurpose strength predicting tree-based models for an optimisation context.

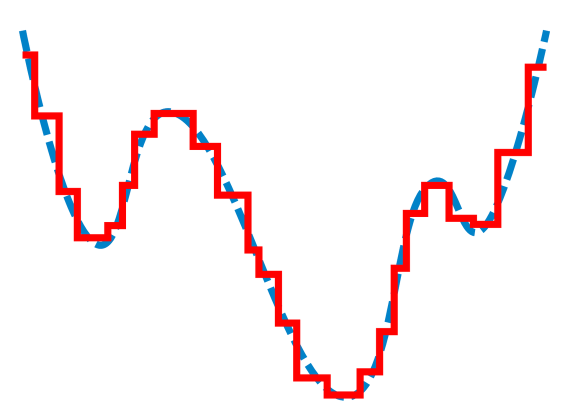
From Prediction to Optimal Decision-Making

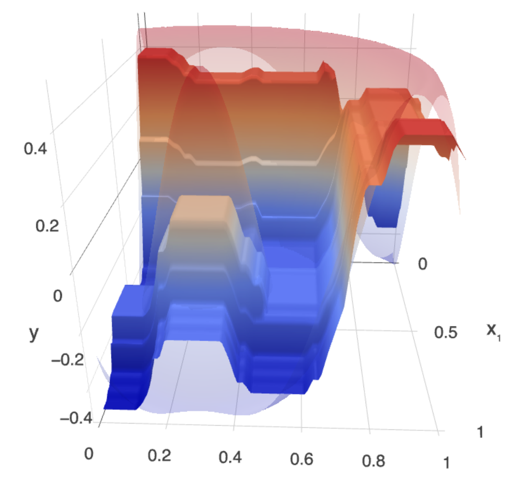
Goal: Integrate and manage tree-based models in larger decision-making problems.

Acquire data → Learn tree-based model → Optimise over tree-based model

Difficulty in Optimising Tree-Based Models

Optimising over tree-based models is difficult because they lack smoothness. Our research considers how to leverage the tree-based structure inherent to these machine learnt functions.

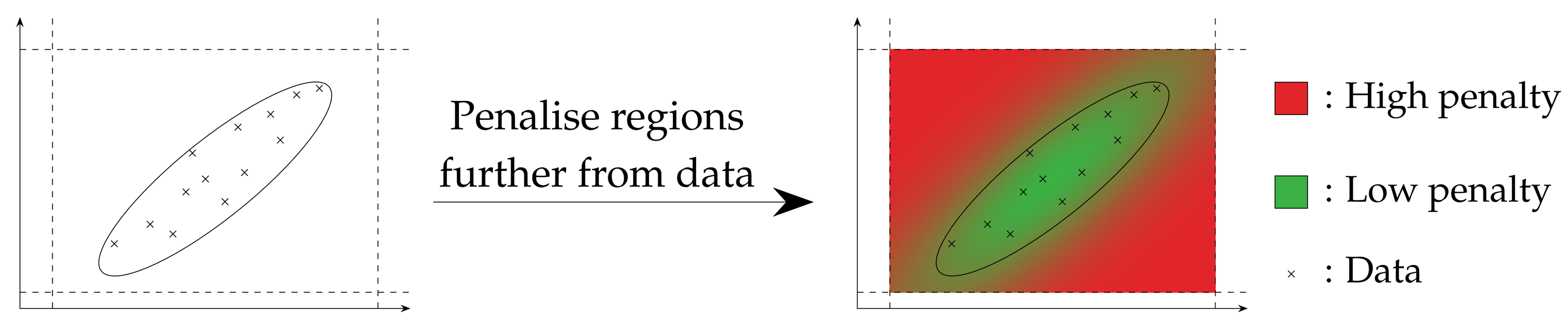




Tree-based models move in steps. This makes optimisation challenging.

Where a Tree-Based Model Makes Sense

A tree-based model is trustworthy in regions close to data and acquiring more data may be expensive. We represent trustworthiness with a penalty that makes regions further from data less optimal.



Decision-Making Optimisation Problem:

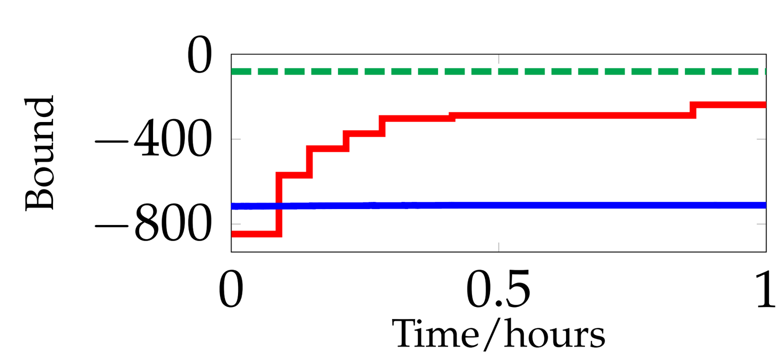
OPTIMISE [Tree-based model] + [Penalty]

Guaranteeing Decision Quality

We develop an algorithm that guarantees the best solution of the decision-making optimisation problem. The algorithm automatically analyses tree-based model structure to dynamically split the problem into easier-to-solve subproblems.

Key algorithm elements:

- Leverage efficient mathematical solvers
- Specialised approximation for tree-based models
- Removing non-optimal regions by dynamically dividing domain.



Closeness to the green dashed line yields a better guarantee. Our approach (red line) outperforms off-the-shelf approaches (blue line).

Getting Good Decisions Quickly

Particle swarm optimisation:
Generate particles near low penalty regions and focus search in a collaborative manner.

Decomposing tree-based models:
Approximate tree-based model with a decomposition and use a mathematical solver. Iteratively generate multiple candidate solutions.

Future Work

Design methods for handling additional machine learnt models in a decision-making context.

Develop freely available decision-making software that integrates and manages machine learnt functions.

Acknowledgements

The support of BASF SE, Lugwigshafen am Rhein, the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems to M.M. (HiPEDS, EP/L016796/1) and an EPSRC Research Fellowship to R.M. (EP/P016871/1) is gratefully acknowledged.

References

[1] M. Mistry, D. Letsios, R. Misener, G. Krennrich, and R. M. Lee. Mixed-Integer Convex Nonlinear Optimization with Gradient-Boosted Trees Embedded. *ArXiv*, 2018. arXiv:1803.00952.