Information-theoretic partition-based models for interpretable machine learning

Yang, L.

Propositions
accompanying the thesis

Information-theoretic Partition-based Models for Interpretable Machine Learning

1. Rule set models must be truly unordered to be interpretable enough for human-guided rule learning. [Chapter 2 & 3]
2. A probabilistic rule set simply summarizes the data without any additional modelling assumptions; thus, it is as reliable as the data itself. [Chapter 2 & 3]
3. Data-driven modelling in critical areas needs to provide interpretable and actionable insights. [Chapter 4]
4. The less number of hyper-parameters, the more unambiguous knowledge discovered. [Chapter 5]
5. The minimum description length (MDL) principle provides a principled way for regularization without a regularization parameter to be tuned.
6. Post-hoc explanations generated by non-transparent algorithms can raise questions about the explainability of the generated explanations themselves.
7. Interactive machine learning requires humans to understand machines, and simultaneously, machines to understand humans. Rule-based models can serve as a common 'language' for humans and machines to communicate.
8. While the goal of designing a machine learning system is often to provide the right answers to the predetermined questions, the goal of designing interactive and human-guided machine learning method involves how to allow humans to ask the right questions.
9. Humans have been disagreeing with each other for a long time, and we will probably also see humans and machines disagreeing with each other in the future.

Lincen Yang
Leiden, 20-09-2024