For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics

PHD STUDENT SEMINAR

Rule-Prioritised Entropy Logic Explanations: Integrating Domain Knowledge into Interpretable Classification

By

Miss Ho Yi Alexis HO

Abstract

In many real-world applications, practitioners possess not only data-driven concept representations but also domain-specific rules of thumb, such as clinical heuristics or regulatory criteria. The Entropy Logic Explanation (ELE) framework provides a principled way to train classifiers on human-interpretable concepts and distil them into Boolean classification rules. However, naïvely combining domain rules with base concepts into a single input space leads to redundant explanations and weakened sparsity. We propose a rule-prioritised variant of ELE that integrates domain knowledge through a two-phase sequential training procedure. Phase 1 trains an ELE exclusively on domain rule indicators in a compact, low-dimensional space. If this already achieves sufficient accuracy, the process terminates with a highly concise model. Otherwise, Phase 2 freezes the learned rule contributions and trains a second ELE on base concepts, with the frozen rule output acting as a fixed offset. Because the two sets of features never compete within the same attention mechanism, each phase induces sparsity independently and effectively. The resulting model produces hierarchical Boolean explanations that can be read at both the rule level and the expanded concept level. We further show that the method degrades gracefully: when domain rules are uninformative, the framework naturally recovers a standard concept-based ELE with negligible overhead.

Date : 28 April 2026, Tuesday

Time : 10:00am (Hong Kong time)

Zoom Meeting : https://hkust.zoom.us/j/96728799934
(Passcode: 806784)

All are Welcome!