Information-theoretic partition-based models for interpretable machine learning

Yang, L.

Information-theoretic Partition-based Models for Interpretable Machine Learning

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op vrijdag 20 september 2024
klokke 11:30 uur

door

Yang, Lincen 阳林岑

geboren te Zigong 自贡, China

in 1993

Promotores:

Dr. M. van Leeuwen
Prof.dr. A. Plaat

Promotiecommissie:

Prof.dr. E. Fromont	Université de Rennes
Prof.dr. T. De Bie	University of Ghent
Dr. S. Yu	Vrije Universiteit Amsterdam
Prof.dr. M. Bonsangue	
Dr. A. Knobbe	

This PhD project was conducted in the Explanatory Data Analysis Group, Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands.

SIKS Dissertation Series No. 2024-27. The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Research presented in this thesis was funded by the Dutch Research Council (NWO), under the the research programme 'Human-Guided Data Science by Interactive Model Selection' with Project Number 612.001.804.

ISBN: 978-94-6506-377-5
Printed by: Ridderprint

To my grandfather.

Abstract

Data-driven modeling in applied research often requires both predictive modeling and understanding data. Predictive models in supervised learning are indispensable for decision-making, early warning systems, and forming robust associations; meanwhile, algorithms in data mining and unsupervised learning search for patterns that are crucial for understanding data, getting insights into the physical process behind it, and taking action in the application domain.

In this dissertation, we study partition-based models that can be used both for interpretable predictive modeling and for understanding data via interpretable patterns. Specifically, we study probabilistic rule-based models for multi-class classification and histogram models for discretization, explanatory data analysis, and conditional mutual information estimation.

For rule-based models, we address the long-unresolved problem that interpretability of rule-based models is harmed by the need for conflict-resolving schemes for "overlaps" among rules (i.e., instances covered by multiple rules). Based on the intuitions that 1) overlaps can be used for characterizing the uncertainty of a model, and 2) only rules with similar class probability estimates are "allowed" to overlap, we formally introduce a new probabilistic model based on probabilistic rules, which we name *Truly Unordered Rule Set (TURS)*. In a series of three research papers (Chapters 2–4) we showcase that our proposed method learns "independent" rules that are not "entangled" with each other, which significantly improves the comprehensibility of the induced rule sets, as (explicit and implicit) orders within rules are eliminated. Building upon our proposed TURS model, we conduct a pilot study to demonstrate how it facilitates *interactive rule learning*: rules can be updated after receiving feedback from domain experts regarding disliked variables.

Next, in the realm of histogram-based methods, we first consider two dimensional datasets (Chapter 5), motivated by the ubiquitous spatial datasets collected by GPS devices. While histograms are widely used to discretize and summarize data, how to incorporate dependency among variables in *multivariate unsupervised discretization* is understudied. Further, the lack of a principled way for parameter setting leads to ambiguity to data analysts, as different parameter settings for histograms lead to significantly different results.

We address these issues by introducing a two dimensional histogram method based on the minimum description length (MDL) principle, with enhanced interpretability and transparency in the following aspects. First, we formally define the optimal histogram under the MDL principle, and hence eliminate the need for setting bin sizes, which increases the transparency of how histogram-based data discretization/summarization is obtained. Second, we propose the problem of learning two dimensional histograms in an expressive and flexible model class, in which the data space can be partitioned into subsets consisting of unions of *disjoint* rectangles. Based on this, we increase the interpretability of the model by learning histograms in which neighboring "bins" must have density estimates that are "dissimilar enough" under the MDL framework.

Following this line, we lastly study multi-dimensional adaptive histograms for conditional mutual information (CMI) estimation (Chapter 6), which is a fundamental task in understanding (conditional) independence and dependence relationships among variables. Thus, CMI estimation has applications in feature selection, independence testing, probabilistic graphic models, and causal inference. We specifically consider discrete-continuous mixture data, which is common in application areas where data can be truncated or is collected in a way that numeric values (instead of discrete levels) are only recorded in specific (e.g., anomaly) situations. We introduce histograms that can handle such mixture data, and support it with theoretical underpinnings, including measure-theoretic entropy definitions and consistency proofs. Notably, histogram bins with large differences between the (empirical) joint entropy and the sum of marginal entropies can be regarded as interpretable patterns for explaining dependency.

In conclusion, this dissertation explores MDL-based partition-based models and advances the field of interpretable machine learning by introducing innovative methods for a variety of tasks.

Contents