



Universiteit
Leiden
The Netherlands

Information-theoretic partition-based models for interpretable machine learning

Yang, L.

Citation

Yang, L. (2024, September 20). *Information-theoretic partition-based models for interpretable machine learning*. SIKS Dissertation Series. Retrieved from <https://hdl.handle.net/1887/4092882>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4092882>

Note: To cite this publication please use the final published version (if applicable).

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
Prof.dr. T. De Bie
Dr. S. Yu
Prof.dr. M. Bonsangue
Dr. A. Knobbe

Université de Rennes
University of Ghent
Vrije Universiteit Amsterdam



**Universiteit
Leiden**
The Netherlands



Copyright © 2024 *Lincen Yang*. All rights reserved.

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

Abstract	v
1 Introduction	1
1.1 Data-driven Models	2
1.1.1 Interpretable patterns for knowledge discovery	2
1.1.2 Interpretable predictive models	3
1.2 What is Interpretability?	4
1.3 Partition-based Models	6
1.3.1 Probabilistic rule sets	6
1.3.2 Multi-dimensional adaptive histograms	7
1.4 A Gentle Introduction to the MDL Principle	8
1.5 Research Questions	9
1.5.1 Towards rule sets for interactive rule learning	10
1.5.2 Adaptive histograms for discretization	11
1.5.3 Histograms for learning dependency structure	12
1.6 Contributions	12
2 Rule Sets with Overlaps that Represent Uncertainty	15
2.1 Introduction	17
2.2 Related Work	19
2.3 Rule Sets as Probabilistic Models	20
2.3.1 Probabilistic Rules	20
2.3.2 Truly Unordered Rule Sets as Probabilistic Models	21
2.4 Rule Set Learning as Probabilistic Model Selection	24
2.4.1 Normalized Maximum Likelihood Distributions for Rule Sets	24

Contents

2.4.2	Approximating the NML Distribution	25
2.5	Learning Truly Unordered Rule Sets from Data	27
2.5.1	Evaluating Incomplete Rule Sets with a Surrogate Score	27
2.5.2	Two-phase Rule Growth	28
2.5.3	Beam Search for Two-phase Rule Growth	30
2.5.4	Iterative search for the rule set	32
2.6	Experiments	33
2.6.1	Results	33
2.7	Conclusion	36
2.8	Appendix I: Reproducibility for Experiments	37
2.9	Appendix II: Proof of Proposition 1	37
2.10	Appendix III: Proof of Proposition 2	38
3	Probabilistic Truly Unordered Rule Sets	39
3.1	Introduction	41
3.2	Related Work	46
3.3	Truly Unordered Rule Sets	49
3.3.1	Probabilistic rules	49
3.3.2	The TURS model	50
3.3.3	Predicting for a new instance	52
3.4	Rule Set Learning as Probabilistic Model Selection	53
3.4.1	Normalized Maximum Likelihood Distributions for Rule Sets	53
3.4.2	Approximating the NML Distribution	54
3.4.3	Code length of model	56
3.4.4	MDL-based model selection	58
3.5	Learning Truly Unordered Rules from Data	59
3.5.1	Learning a rule set	59
3.5.2	Learning a single rule	61
3.6	Experiments	68
3.6.1	Setup	69
3.6.2	Classification performance	70
3.6.3	Prediction with ‘random picking’ for overlaps	71
3.6.4	Generalizability of local probabilistic estimates	72
3.6.5	Model complexity	74

3.6.6	Ablation study 1: diverse patience beam search	75
3.6.7	Ablation study 2: MDL-based local testing	76
3.6.8	Runtime	78
3.7	Conclusion	79
3.8	Appendix: Comparison to the Previous Work	81
4	Case Study: Towards Interactive Rule Learning for ICU Readmission Analysis	89
4.1	Introduction	91
4.2	Updating Rule Sets with Human Feedback	92
4.2.1	Human feedback format	93
4.2.2	Updating a rule set	93
4.3	An Empirical Pilot Study	94
4.3.1	Experiment setup	95
4.3.2	Rule set for the ICU dataset	96
4.3.3	Rule-based competitor methods	96
4.3.4	Human-AI collaboration	97
4.4	Conclusion and Discussion	99
4.4.1	Discussion for future work	100
5	Summarizing Two-dimensional Data with MDL-based Discretization by Histograms	103
5.1	Introduction	105
5.2	Related work	109
5.3	Problem Statement	112
5.3.1	Notation and definitions of data, model, and model class . .	112
5.3.2	Histogram model selection by the MDL principle	113
5.4	Calculating the code length	115
5.4.1	Code length of the data	116
5.4.2	Code length of the model	118
5.5	Revisiting MDL histograms for one-dimensional data	120
5.6	The PALM Algorithm for Partitioning and Merging	124
5.7	Experiments	126
5.7.1	Measuring the difference between two-dimensional histograms	127
5.7.2	Revealing ground truth two-dimensional histograms	128

Contents

5.7.3	Approximating histogram models outside model class \mathbb{M} . . .	130
5.7.4	Gaussian random variables	131
5.7.5	Comparison with IPD	132
5.7.6	Empirical runtime	134
5.8	Case study	135
5.8.1	Datasets	136
5.8.2	Case study tasks	136
5.8.3	Case study results	137
5.8.4	Empirical runtime	138
5.8.5	Algorithm settings	139
5.9	Conclusions	140
5.10	Appendix A: Proof of Proposition 3	140
5.11	Appendix B: Proof of Proposition 4	142
5.12	Appendix C: IPD visualizations on case study datasets	144
6	Interpretable Conditional Mutual Information Estimation with Adaptive Histograms	149
6.1	Introduction	151
6.2	Entropy for Mixed Random Variables	152
6.2.1	A Generalized Definition of Entropy	153
6.3	Adaptive Histogram Models	155
6.3.1	One-Dimensional Histogram Models	155
6.3.2	Multi-Dimensional Histograms	156
6.3.3	Maximum Likelihood Estimator	156
6.3.4	Conditional Mutual Information Estimator	157
6.4	Learning Adaptive Histograms from Data	158
6.4.1	MDL and Stochastic Complexity	158
6.4.2	Code Length of the Model	159
6.5	Implementation	160
6.5.1	Algorithm	160
6.5.2	Complexity	161
6.6	Related Work	161
6.7	Experiments	163
6.7.1	Mutual Information Estimation	163

6.7.2 Independence Testing	166
6.8 Conclusion	168
6.9 Supplementary Material	169
6.9.1 Proofs	169
6.9.2 Implementation Details	174
6.9.3 Data Generation and Additional Experiments	174
7 Conclusions	179
7.1 Summary	180
7.2 Answers to Research Questions	181
7.3 Future Work	184
Acknowledgements	199
Summary	201
Samenvatting	203
Titles in the SIKS dissertation series since 2016	205
Curriculum Vitae	223