1. Related Work. We will discuss related literature in several subfields.

Interpretable Models: There is a growing interest in interpretable (transparent, comprehensible) models because of practical societal importance(see Rüping, 2006; Bratko, 1997; Dawes, 1979; Vellido et al., 2012; Giraud-Carrier, 1998; Holte, 1993; Shmueli, 2010; Huysmans et al., 2011; Freitas, 2014). There are now regulations on algorithmic decision-making in the European Union on the "right to an explanation" (Goodman and Flaxman, 2016) that would legally require interpretability in predictions.

Optimal Decision Tree Modeling: The body of work closest to ours is possibly that of optimal decision tree modeling. There is work starting in the late 1990's on building optimal decision trees using optimization techniques (e.g., Bennett and Blue, 1996; Dobkin et al., 1996), continuing until the present (e.g., Farhangfar et al., 2008). A particularly interesting paper along these lines is that of Nijssen and Fromont (2010), who created a "bottom-up" way to form optimal decision trees. Their method performs an expensive search step, mining all possible leaves (rather than all possible rules), and using those leaves to form trees. Their method can lead to memory problems, but it is possible that these memory issues can be mitigated using the theorems in this paper. [1] Another work close to ours is that of Garofalakis et al. (2000), who introduce an algorithm to generate more interpretable decision trees by allowing constraints to be placed on the size of the decision tree. During tree construction, they bound the possible Minimum Description Length (MDL) cost of every different split at a given node. If every split at that node is more expensive than the actual cost of the current subtree, then that node can be pruned. In this way, they were able to prune the tree while constructing it instead of just constructing the tree and then pruning at the end. They do not aim for optimal trees; they build trees that obey constraints, and find optimal subtrees within the trees that were built during the building phase.

Greedy splitting and pruning: On the other end of the spectrum from optimal decision tree methods are methods like CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993). They do not perform exploration of the search space beyond greedy splitting. There are a huge number of algorithms in this class.

Bayesian tree and rule list methods: Some of these methods that aim to explore the space of trees (Dension et al., 1998; Chipman et al., 2002, 2010) using Monte Carlo exploration methods, however, the space of trees of a given depth is much larger than the space of rule lists of that same level of depth, and the trees within these algorithms are grown in a top-down greedy way. Because of this, the authors noted that the MCMC chain tends to reach only locally optimal solutions. This is why Bayesian rule-based methods (Letham et al., 2015; Yang et al., 2016) have tended to be more successful in escaping local minima. This work builds specifically on that of Yang et al. (2016). In particular, we use their fast bit-vector manipulations, and build on their bounds. Note that the 1995 algorithm RIPPER Cohen (1995) is similar to the Bayesian tree methods in that it grows, prunes, and then locally optimizes.

Rule learning methods: Most rule learning methods are not designed for optimality or interpretability, but mainly for computational speed and/or accuracy. In *associative classification* (Vanhoof and Depaire, 2010; Liu et al., 1998; Li et al., 2001; Yin and Han, 2003), classifiers are often formed greedily from the top down as rule lists, or they are formed by taking the simple union of pre-mined rules, whereby any observation that fits into any of the rules is classified as positive. In *inductive logic programming* (Muggleton and De Raedt, 1994), algorithms form disjunctive normal form patterns via a set of operations (rather than using optimization). These approaches are not appropriate for obtaining a guarantee of optimality. Methods for decision list learning construct

[1] There is no public version of their code for distribution as of this writing.

rule lists iteratively in a greedy way (Rivest, 1987; Sokolova et al., 2003; Marchand and Sokolova, 2005; Rudin et al., 2013; Goessling and Kang, 2015), which again have no guarantee on optimality, and tend not to produce optimal rule lists in general. Some methods allow for interpretations of single rules, without constructing rule lists (McCormick et al., 2012).

There is a tremendous amount of related work in other subfields that are too numerous to discuss at length here. We have not discussed *rule mining* algorithms since they are part of an interchangeable preprocessing step for our algorithm, and are deterministically fast (that is, they will not generally slow our algorithm down). We also did not discuss methods that create disjunctive normal form models, e.g. logical analysis of data, and many associative classification methods).

Related problems concerning interpretable lists of rules: Beyond trees that are optimized for accuracy and sparsity, rule lists have been developed that have exotic types of constraints, and for various applications. This includes Falling Rule Lists (Wang and Rudin, 2015), which are constrained to have decreasing probabilities down the list, as well as rule lists for dynamic treatment regimes (Zhang et al., 2015) and cost-sensitive dynamic treatment regimes (Lakkaraju and Rudin, 2017). Both Wang and Rudin (2015) and Lakkaraju and Rudin (2017) use Monte Carlo searches through the space of rule lists. The method proposed in this work could potentially be adapted to handle these kinds of interesting problems. We are currently working on bounds for Falling Rule Lists (Chen and Rudin, 2017) similar to those presented here.

References.

K. P. Bennett and J. A. Blue. Optimal decision trees. Technical report, R.P.I. Math Report No. 214, Rensselaer Polytechnic Institute, 1996.

I. Bratko. Machine learning: between accuracy and interpretability. In G. Della Riccia, H.-J. Lenz, and R. Kruse, editors, *Learning, Networks and Statistics*, volume 382 of *International Centre for Mechanical Sciences*, pages 163–177. Springer Vienna, 1997. ISBN 978-3-211-82910-3.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, 1984.

C. Chen and C. Rudin. Optimized falling rule lists and softly falling rule lists. work in progress, 2017.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian treed models. *Machine Learning*, 48(1/3):299–320, 2002.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

W. W. Cohen. Fast effective rule induction. In *Twelfth International Conference on Machine Learning (ICML)*, pages 115–123, 1995.

R. M. Dawes. The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7):571–582, 1979.

D. Dension, B. Mallick, and A. Smith. A Bayesian CART algorithm. *Biometrika*, 85(2):363–377, 1998.

D. Dobkin, T. Fulton, D. Gunopulos, S. Kasif, and S. Salzberg. Induction of shallow decision trees, 1996.

A. Farhangfar, R. Greiner, and M. Zinkevich. A fast way to produce optimal fixed-depth decision trees. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM 2008), Fort Lauderdale, Florida, USA, January 2-4, 2008*, 2008.

A. A. Freitas. Comprehensible classification models. *SIGKDD Explorations*, 2014.

M. Garofalakis, D. Hyun, R. Rastogi, and K. Shim. Efficient algorithms for constructing decision trees with constraints. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 335–339, New York, NY, USA, 2000. ACM.

C. Giraud-Carrier. Beyond predictive accuracy: what? In *Proceedings of the ECML-98 Workshop on Upgrading Learning to Meta-Level: Model Selection and Data Transformation*, pages 78–85, 1998.

M. Goessling and S. Kang. Directional decision lists. ArXiv e-prints 1508.07643, Aug 2015.

B. Goodman and S. Flaxman. EU regulations on algorithmic decision-making and a “right to explanation”. *arXiv preprint arXiv:1606.08813*, 2016.

R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11 (1):63–91, 1993.

J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

H. Lakkaraju and C. Rudin. Cost-sensitive and interpretable dynamic treatment regimes based on rule lists. In *Proceedings of the Artificial Intelligence and Statistics (AISTATS)*, 2017.

B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.

W. Li, J. Han, and J. Pei. CMAR: accurate and efficient classification based on multiple class-association rules. *IEEE International Conference on Data Mining*, pages 369–376, 2001.

B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, KDD '98, pages 80–96, 1998.

M. Marchand and M. Sokolova. Learning with decision lists of data-dependent features. *Journal of Machine Learning Research*, 6:427–451, 2005.

T. H. McCormick, C. Rudin, and D. Madigan. Bayesian hierarchical rule modeling for predicting medical conditions. *The Annals of Applied Statistics*, 6:652–668, 2012.

S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679, 1994.

S. Nijssen and E. Fromont. Optimal constraint-based decision tree induction from itemset lattices. *Data Mining and Knowledge Discovery*, 21(1):9–51, 2010. ISSN 1384-5810.

J. Quinlan. C4.5: Programs for machine learning. 1993.

R. L. Rivest. Learning decision lists. *Machine Learnning*, 2(3):229–246, Nov. 1987.

C. Rudin, B. Letham, and D. Madigan. Learning theory analysis for association rules and sequential event prediction. *Journal of Machine Learning Research*, 14:3384–3436, 2013.

S. Rüping. *Learning interpretable models*. PhD thesis, Universität Dortmund, 2006.

G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, August 2010. ISSN 0883-4237.

M. Sokolova, M. Marchand, N. Japkowicz, and J. Shawe-Taylor. The decision list machine. In *Advances in Neural Information Processing Systems*, volume 15 of *NIPS '03*, pages 921–928, 2003.

K. Vanhoof and B. Depaire. Structure of association rule classifiers: a review. In *Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering*, ISKE '10, pages 9–12, 2010.

A. Vellido, J. D. Martín-Guerrero, and P. J. Lisboa. Making machine learning models interpretable. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2012.

F. Wang and C. Rudin. Falling rule lists. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 2015.

H. Yang, C. Rudin, and M. Seltzer. Scalable Bayesian rule lists. *Preprint at arXiv:1602.08610*, 2016.

X. Yin and J. Han. Cpar: classification based on predictive association rules. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, ICDM '03, pages 331–335, 2003.

Y. Zhang, E. B. Laber, A. Tsiatis, and M. Davidian. Using Decision Lists to Construct Interpretable and Parsimonious Treatment Regimes. *ArXiv e-prints*, Apr. 2015.