Ustun, Berk; Rudin, Cynthia
Learning optimized risk scores. (English) Zbl 1440.68242
J. Mach. Learn. Res. 20, Paper No. 150, 75 p. (2019).

Summary: Risk scores are simple classification models that let users make quick risk predictions by adding and subtracting a few small numbers. These models are widely used in medicine and criminal justice, but are difficult to learn from data because they need to be calibrated, sparse, use small integer coefficients, and obey application-specific constraints. In this paper, we introduce a machine learning method to learn risk scores. We formulate the risk score problem as a mixed integer nonlinear program, and present a cutting plane algorithm to recover its optimal solution. We improve our algorithm with specialized techniques that generate feasible solutions, narrow the optimality gap, and reduce data-related computation. Our algorithm can train risk scores in a way that scales linearly in the number of samples in a dataset, and that allows practitioners to address application-specific constraints without parameter tuning or post-processing. We benchmark the performance of different methods to learn risk scores on publicly available datasets, comparing risk scores produced by our method to risk scores built using methods that are used in practice. We also discuss the practical benefits of our method through a real-world application where we build a customized risk score for ICU seizure prediction in collaboration with the Massachusetts General Hospital.

MSC:
68T05 Learning and adaptive systems in artificial intelligence
62H30 Classification and discrimination; cluster analysis (statistical aspects)
90C30 Nonlinear programming

Keywords:
score systems; classification; calibration; customization; interpretability; cutting-planes methods; discrete optimization; mixed integer nonlinear programming

Software:
ORL; KNITRO; CPLEX; UCI-ml

Full Text: arXiv Link

References:
[1] Alba, Ana Carolina, Thomas Agoritsas, Michael Walsh, Steven Hanna, Alfonso Iorio, PJ Devereaux, Thomas McGinn, and Gordon Guyatt. Discrimination and calibration of clinical prediction models: Users’ guides to the medical literature. Journal of the American Medical Association, 318 (14):1377-1384, 2017.
[2] American Clinical Neurophysiology Society. Standardized Critical Care EEG Terminology Training Module, 2012.
[3] Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Man’e. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565, 2016.
[4] Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. Journal of Machine Learning Research, 18(234): 1-78, 2018.
[5] Antman, Elliott M, Marc Cohen, Peter ILM Bernink, Carolyn H McCabe, Thomas Horacek, Gary Papuchis, Branco Mautner, Ramon Corbalan, David Radley, and Eugene Braunwald. The TIMI risk score for unstable angina/non-ST elevation MI. Journal of the American Medical Association, 284(7):835-842, 2000.
[6] Austin, James, Roger Ocker, and Avi Bhati. Kentucky Pretrial Risk Assessment Instrument Validation. Bureau of Justice Statistics, 2010.
[7] Bache, K. and M. Lichman. UCI Machine Learning Repository, 2013.
[8] Bai, Lihui and Paul A Rubin. Combinatorial Benders Cuts for the Minimum Tollbooth Problem. Operations Research, 57(6):1510-1522, 2009.
[9] Bardenet, Rémi and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. Bernoulli, 21(3):1361-1385, 2015.
[10] Beneish, Messod D, Charles MC Lee, and D Craig Nichols. Earnings manipulation and expected returns. Financial Analysts
Bertsimas, Dimitris, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. The Annals of Statistics, 44(2):813-852, 2016.

Billiet, Lieven, Sabine Van Huffel, and Vanya Van Belle. Interval coded scoring extensions for larger problems. In Proceedings of the IEEE Symposium on Computers and Communications, pages 198-203. IEEE, 2017.

Billiet, Lieven, Sabine Van Huffel, and Vanya Van Belle. Interval Coded Scoring: A toolbox for interpretable scoring systems. PeerJ Computer Science, 4:e150, 04 2018.

Bolko, Philip, Philip L Roth, and Maury A Buster. The usefulness of unit weights in creating composite scores. A literature review, application to content validity, and meta-analysis. Organizational Research Methods, 10(4):689-709, 2007.

Bonami, Pierre, Mustafa Kili ın, c, and Jeff Linderoth. Algorithms and software for convex mixed integer nonlinear programs. In Mixed Integer Nonlinear Programming, pages 1-39. Springer, 2012.

Böckenholt, Ulrich. The use of information weighting schemes in discriminant analysis. Biometrika, 75(2):339-343, 1988.

Boyd, Stephen P and Lieven Vandenberghe. Convex Optimization. Cambridge University Press, 2004.

Burgess, Ernest W. Factors determining success or failure on parole. The workings of the indeterminate sentence law and the parole system in Illinois, pages 221-234, 1928.

Byrd, Richard H, Jorge Nocedal, and Richard A Waltz. KNITRO: An Integrated Package for Nonlinear Optimization. In Large-scale Nonlinear Optimization, pages 35-59. Springer, 2006.

Calmon, Flavio, Dennis Wei, Bhanukrishan Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems, pages 3995-4004, 2017.

Carrizosa, Emilio, Amaya Nogales-Gómez, and Dolores Romero Morales. Strongly agree or strongly disagree?: Rating features in support vector machines. Information Sciences, 329:256-273, 2016.

Caruana, Rich and Alexandre Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 69-78. ACM, 2004.

Caruana, Rich and Alexandre Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine Learning, pages 161-168. ACM, 2006.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligent models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1721-1730. ACM, 2015.

Cawley, Gavin C and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11(Jul):2079-2107, 2010.

Chang, Allison, Cynthia Rudin, Michael Cavaretta, Robert Thomas, and Gloria Chou. How to reverse-engineer quality rankings. Machine Learning, 88:369-398, September 2012.

Chen, Chaofan and Cynthia Rudin. An optimization approach to learning falling rule lists. In Proceedings of the 21st International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 604-612. PMLR, 09-11 Apr 2018.

Chen, Chaofan, Kauchen Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. An interpretable model with globally consistent explanations for credit risk. In Proceedings of NeurIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy, 2018.

Chevaleyre, Yann, Frederic Koriche, and Jean-Daniel Zucker. Rounding methods for discrete linear classification. In Proceedings of the 30th International Conference on Machine Learning, pages 651-659, 2013.

Cranor, Lorrie Faith and Brian A LaMacchia. SpanCommunications of the ACM, 41(8):74-83, 1998.

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

DeGroot, Morris H and Stephen E Fienberg. The comparison and evaluation of forecasters. The Statistician, pages 12-22, 1983.

Duwe, Grant and KiDeuk Kim. Sacrificing accuracy for transparency in recidivism risk assessment: The impact of classification method on predictive performance. pages 1-22, 2016.

Einhorn, Hillel J and Robin M Hogarth. Unit weighting schemes for decision making. Organizational Behavior and Human Performance, 13(2):171-192, 1976.

Elber, M, R Schultz-Wendtland, and T Wittenberg. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. Medical Physics, 34:4164, 2007.

Ertekin, S, cyda and Cynthia Rudin. On equivalence relationships between classification and ranking algorithms. Journal of Machine Learning Research, 12:2905-2929, 2011.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 259-268. ACM, 2015.

FICO. Introduction to Scorecard for FICO ModelBuilder. http://www.fico.com/en/node/8140?file=7900, 2011.

Finlay, Steven. Credit scoring, response modeling, and insurance rating: a practical guide to forecasting consumer behavior. Palgrave Macmillan, 2012.

Franc, Vojtˇech and Soeren Sonnenburg. Optimized cutting plane algorithm for support vector machines. In Proceedings of
the 25th International Conference on Machine Learning, pages 320-327. ACM, 2008.

[40] Franc, Vojtˇech and S´oren Sonnenburg. Optimized cutting plane algorithm for large-scale risk minimization. Journal of Machine Learning Research, 10:2157-2192, 2009.

[41] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1-22, 2010.

[42] Gage, Brian F, Amy D Waterman, William Shannon, Michael Boechler, Michael W Rich, and Martha J Radford. Validation of clinical classification schemes for predicting stroke. Journal of the American Medical Association, 285(22):2864-2870, 2001.

[43] Goel, Sharad, Justin M Rao, and Ravi Shroff. Precinct or Prejudice? Understanding Racial Disparities in New York City’s Stop-and-Frisk Policy. Annals of Applied Statistics, 10(1):365-394, 2016.

[44] Geh, Gabriel, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In Advances in Neural Information Processing Systems, pages 2415-2423, 2016.

[45] Geh, Siong Thye and Cynthia Rudin. Box drawings for learning with imbalanced data. In Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 333-342. ACM, 2014.

[46] Goldberg, Noam and Jonathan Eckstein. Sparse weighted voting classifier selection and its linear programming relaxations. Information Processing Letters, 112:481-486, 2012.

[47] Gottfredson, Don M and Howard N Snyder. The mathematics of risk classification: Changing data into valid instruments for juvenile courts. NCJ 209158: Office of Juvenile Justice and Delinquency Prevention Washington, D.C., 2005.

[48] Guan, Wei, Alex Gray, and Sven Leyffer. Mixed-integer support vector machine. In NIPS Workshop on Optimization for Machine Learning, 2009.

[49] Gupta, Maya, Andrew Cotter, Jan Pfeifer, Konstantin Voevodski, Kevin Canini, Alexander Mangylov, Wojciech Moczydlowski, and Alexander Van Esbroeck. Monotonic calibrated interpolated look-up tables. Journal of Machine Learning Research, 17(1):3795-3836, 2016.

[50] Hirsch, L.I., SM LaRuche, N Gaspard, E Gerard, A Svoronos, ST Herman, R Mani, H Arif, N Jette, Y Minazad, et al. American clinical neurophysiology society’s standardized critical care EEG terminology: 2012 version. Journal of Clinical Neurophysiology, 30(1):1-27, 2013.

[51] Holte, Robert C. Very simple classification rules perform well on most commonly used datasets. Machine Learning, 11(1):63-90, 1993.

[52] Holte, Robert C. Elaboration on Two Points Raised in “Classifier Technology and the Illusion of Progress”. Statistical Science, 21(1):24-26, February 2006.

[53] Hu, Xiyang, Cynthia Rudin, and Margo Seltzer. Optimal sparse decision trees. In Proc. Neural Information Processing Systems, 2019.

[54] ILOG, IBM. CPLEX Optimizer 12.6. https://www-01.ibm.com/software/commerce/optimization/cplexoptimizer/, 2017.

[55] Joachims, Thorsten. Training linear SVMs in linear time. In Proceedings of the 12th ACM SIGKDD International conference on Knowledge Discovery and Data Mining, pages 217-226. ACM, 2006.

[56] Joachims, Thorsten, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural SVMs. Machine Learning, 77(1):27-59, 2009.

[57] Kamishima, Toshihiro, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In 2011 IEEE 11th International Conference on Data Mining Workshops, pages 643-650. IEEE, 2011.

[58] Kelley, James E., Jr. The cutting-plane method for solving convex programs. Journal of the Society for Industrial and Applied Mathematics, 8(4):703-712, 1960.

[59] Kessler, Ronald C, Lenard Adler, Minnie Ames, Olga Demler, Steve Faraone, Eva Hiripi, Mary J Howes, Robert Jin, Kristina Secnik, Thomas Spencer, et al. The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population. Psychological Medicine, 35(02):245-256, 2005.

[60] Kohavi, Ron. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pages 202-207. AAAI Press, 1996.

[61] Kotlowski, Wojciech, Krzysztof J Dembczynski, and Eyke Huellermeier. Bipartite ranking through minimization of univariate loss. In Proceedings of the 28th International Conference on Machine Learning, pages 1113-1120. 2011.

[62] Kronqvist, Jan, David E Bernal, Andreas Lundell, and Ignacio E Grossmann. A review and comparison of solvers for convex MINLP. Optimization and Engineering, 20(2):397-455, 2019.

[63] Lakkaraju, Himabindu, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1675-1684. ACM, 2016.

[64] Latessa, Edward, Paula Smith, Richard Lemke, Matthew Makarios, and Christopher Lowenkamp. Creation and validation of the Ohio risk assessment system: Final report, 2009.

[65] Le Gall, Jean-Roger, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. Journal of the American Medical Association, 270(24):2957-2963, 1993.

[66] Letham, Benjamin, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics, 9(3):1350-1371, 2015.

[67] Li, Oscar, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In 32nd AAAI Conference on Artificial Intelligence, 2018.
[68] Liu, Yufeng and Yichao Wu. Variable selection via a combination of the L0 and L1 penalties. Journal of Computational and Graphical Statistics, 16(4), 2007.

[69] Lubin, Miles, Emre Yamangil, Russell Bent, and Juan Pablo Vielma. Polyhedral approximation in mixed-integer convex optimization. Mathematical Programming, 172(1-2):139-168, 2018.

[70] Malinovtov, Dmitry and Kush Varshney. Exact rule learning via boolean compressed sensing. In Proceedings of the 30th International Conference on Machine Learning, volume 28 of Proceedings of Machine Learning Research, pages 765-773. PMLR, 17-19 Jun 2013.

[71] Mangasarian, Olvi L, W Nick Street, and William H Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4):570-577, 1995.

[72] McGinley, Ann and Rupert M Pearse. A National Early Warning Score for Acutely Ill Patients, 2012.

[73] Menon, Aditya Krishna, Xiaoqian J Jiang, Shankar Venbu, Charles Elkan, and Lucila Ohno-Machado. Predicting accurate probabilities with a ranking loss. In Proceedings of the International Conference on Machine Learning, volume 2012, page 703, 2012.

[74] Moreno, Rui P, Philipp GH Metzutz, Eduardo Almeida, Barbara Jordan, Peter Bauer, Ricardo Abizanda Campos, Gaetano Iapichino, David Edbrooke, Maurizia Capuzzo, and Jean-Roger Le Gall. SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Medicine, 31(10):1345-1355, 2005.

[75] Moro, S’erigo, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62:22-31, 2014.

[76] Naeini, Mahdi Pakdaman, Gregory F Cooper, and Miles Hauskrecht. Binary classifier calibration: A bayesian non-parametric approach. InProc. SIAM Int Conf Data Mining (SDM), pages 208-216, 2015.

[77] Naoum-Sawaya, Joe and Samir Elhedhli. An interior-point Benders based branch-and-cut algorithm for mixed integer programs. Annals of Operations Research, 210(1):33-55, November 2010.

[78] Nguyen, Hai Thanh and Katrin Franke. A general Lp-norm support vector machine via mixed 0-1 programming. In Machine Learning and Data Mining in Pattern Recognition, pages 40-49. Springer, 2012.

[79] Park, Jaehyun and Stephen Boyd. A semidefinite programming method for integer convex quadratic minimization. Optimization Letters, 12(3):499-518, 2018.

[80] Pennsylvania Bulletin. Sentence Risk Assessment Instrument, April 2017.

[81] Pennsylvania Commission on Sentencing. Interim Report 4: Development of Risk Assessment Scale, June 2012.

[82] Pietroski, Joseph D. Value investing: The use of historical financial statement information to separate winners from losers. Journal of Accounting Research, pages 1-41, 2000.

[83] Platt, John C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers, 10(3):61-74, 1999.

[84] Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In Advances in Neural Information Processing Systems, pages 5684-5693, 2017.

[85] Reid, Mark D and Robert C Williamson. Composite binary losses. Journal of Machine Learning Research, 11:2387-2422, 2010.

[86] Reilly, Brendan M and Arthur T Evans. Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. Annals of Internal Medicine, 144(3):201-209, 2006.

[87] Rudin, Cynthia. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1:206-215, May 2019.

[88] Rudin, Cynthia and S. eyda Ertekin. Learning customized and optimized lists of rules with mathematical programming. Mathematical Programming C (Computation), 10:659-702, 2018.

[89] Rudin, Cynthia and Berk Ustun. Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice. INFORMS Journal on Applied Analytics, 48:399-486, 2018. Special Issue: 2017 Daniel H. Wagner Prize for Excellence in Operations Research Practice.

[90] Rudin, Cynthia and Yining Wang. Direct Learning to Rank And Rerank. In Proceedings of the 21st International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 775-783. PMLR, 09-11 Apr 2018.

[91] Rudin, Cynthia, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. Harvard Data Science Review, 2019. Forthcoming.

[92] Sato, Toshiki, Yuichi Takano, Ryuhei Miyashiro, and Akiko Yoshise. Feature subset selection for logistic regression via mixed integer optimization. Computational Optimization and Applications, 64(3):865-880, July 2016.

[93] Sato, Toshiki, Yuichi Takano, and Ryuhei Miyashiro. Piecewise-linear approximation for feature subset selection in a sequential logit model. Journal of the Operations Research Society of Japan, 60(1):1-14, March 2017.

[94] Schlimmer, Jeffrey C. Concept acquisition through representational adjustment. PhD thesis, University of California, Irvine, 1987. AAI8723227.

[95] Shah, Nilay D, Ewout W Steyerberg, and David M Kent. Big Data and Predictive Analytics: Recalibrating Expectations. Journal of the American Medical Association, 2018.

[96] Siddiqi, Naeem. Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards. John Wiley & Sons, second edition, January 2017. ISBN 978-1-119-27915-0.
[119]

[118]

[116]

[115]

[114]

[113]

[112]

[111]

[110]

[109]

[108]

[107]

[106]

[105]

[104]

[103]

[102]

[101]

[100]

[99]

[98]

[97]

[96]

[95]

[94]

[93]

[92]

[91]

[90]

[89]

[88]

[87]

[86]

[85]

[84]

[83]

[82]

[81]

[80]

[79]

[78]

[77]

[76]

[75]

[74]

[73]

[72]

[71]

[70]

[69]

[68]

[67]

[66]

[65]

[64]

[63]

[62]

[61]

[60]

[59]

[58]

[57]

[56]

[55]

[54]

[53]

[52]

[51]

[50]

[49]

[48]

[47]

[46]

[45]

[44]

[43]

[42]

[41]

[40]

[39]

[38]

[37]

[36]

[35]

[34]

[33]

[32]

[31]

[30]

[29]

[28]

[27]

[26]

[25]

[24]

[23]

[22]

[21]

[20]

[19]

[18]

[17]

[16]

[15]

[14]

[13]

[12]

[11]

[10]

[9]

[8]

[7]

[6]

[5]

[4]

[3]

[2]

[1]