Electronic Journal of Statistics > Vol. 2 (2008) open journal systems

Structured variable selection in support vector machines

Seongho Wu, *University of Minnesota*
Hui Zou, *University of Minnesota*
Ming Yuan, *Georgia Institute of Technology*

Abstract

When applying the support vector machine (SVM) to high-dimensional classification problems, we often impose a sparse structure in the SVM to eliminate the influences of the irrelevant predictors. The lasso and other variable selection techniques have been successfully used in the SVM to perform automatic variable selection. In some problems, there is a natural hierarchical structure among the variables. Thus, in order to have aninterpretable SVM classifier, it is important to respect the heredity principle when enforcing the sparsity in the SVM. Many variable selection methods, however, do not respect the heredity principle. In this paper we enforce both sparsity and the heredity principle in the SVM by using the so-called structured variable selection (SVS) framework originally proposed in Yuan, Joseph and Zou (2007). We minimize the empirical hinge loss under a set of linear inequality constraints and a lasso-type penalty. The solution always obeys the desired heredity principle and enjoys sparsity. The new SVM classifier can be efficiently fitted, because the optimization problem is a linear program. Another contribution of this work is to present a nonparametric extension of the SVS framework, and we propose nonparametric heredity SVMs. Simulated and read data are used to illustrate the merits of the proposed method.

AMS 2000 subject classifications: Primary 68T10; secondary 62G05.

Keywords: Classification, Heredity, Nonparametric estimation, Support vector machine, Variable selection.

Full Text: PDF

Wu, Seongho, Zou, Hui, Yuan, Ming, Structured variable selection in support vector machines, Electronic Journal of Statistics, 2, (2008), 103-117 (electronic). DOI: 10.1214/07-EJS125.

References

[1] Bradley, P. and Mangasarian, O. (1998). Feature selection via concave minimization and support vector machines. In J. Shavlik (eds), ICML'98. Morgan Kaufmann.

[2] Breiman, L. (1995). Better subset regression using the nonnegative garrote. Technometrics 37, 4, 373–384. MR1365720

[3] Chipman, H. (1996). Bayesian variable selection with related predictors. Canad. J. Statist. 24, 1, 17–36. MR1394738

[4] Chipman, H., Hamada, M. and Wu, C. F. J. (1997). A Bayesian variable selection approach for analyzing designed experiments with complex aliasing. Technometrics 39, 372–381.

[5] Choi, N. and Zhu, J. (2006). Variable selection with strong heredity / marginality constraints. Technical Report, Department of Statistics, University of Michigan, Ann Arbor.

[6] de Boor, C. (1978). A practical guide to splines. Applied Mathematical Sciences,

Vol. 27. Springer-Verlag, New York. MR0507062

[7] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. Ann. Statist. 32, 2, 407–499. With discussion, and a rejoinder by the authors. MR2060166

[8] Green, P. J. and Silverman, B. W. (1994). Nonparametric regression and generalized linear models. Monographs on Statistics and Applied Probability, Vol. 58. Chapman & Hall, London. A roughness penalty approach. MR1270012

[9] Hamada, M. and Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing, Journal of Quality Technology 24, 130–137.

[10] Hastie, T. and Tibshirani, R. (2004). Efficient Quadratic Regularization for Expression Arrays, Biostatistics 5, 329–340.

[11] Hastie, T., Tibshirani, R., and Friedman, J. (2001). The elements of statistical learning. Springer Series in Statistics. Springer-Verlag, New York. Data mining, inference, and prediction. MR1851606

[12] Hastie, T. J. and Tibshirani, R. J. (1990). Generalized additive models. Monographs on Statistics and Applied Probability, Vol. 43. Chapman and Hall Ltd., London. MR1082147

[13] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58, 1, 267–288. MR1379242

[14] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. Ann. Statist. 32, 2, 407–499. With discussion, and a rejoinder by the authors. MR2060166

[15] Vapnik, V. (1996). The Nature of Statistical Learning. Springer Verlag, New York.

[16] Venables, W. N. and Ripley, B. D. (1994). Modern applied statistics with S-Plus. Statistics and Computing. Springer-Verlag, New York. With 1 IBM-PC floppy disk (3.5 inch; HD). MR1337030

[17] Wahba, G. (1990). Spline models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR1045442

[18] Wu, C. F. J. and Hamada, M. (2000). Experiments. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons Inc., New York. Planning, analysis, and parameter design optimization, A Wiley-Interscience Publication. MR1780411

[19] Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. J. Amer. Statist. Assoc. 100, 472, 1215–1225. MR2236436

[20] Yuan, M., Joseph, R. and Zou, H. (2007). Structured Variable Selection and Estimation, Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology.

[21] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68, 1, 49–67. MR2212574

[22] Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. J. R. Stat. Soc. Ser. B Stat. Methodol. 69, 2, 143–161. MR2325269

[23] Zhao, P., Rocha, G. and Yu, B. (2006). Grouped and hierarchical model selection through composite absolute penalties. Technical Report, Department of Statistics, University of California, Berkeley.

[24] Zhu, J., Rosset, S., Hastie, T. & Tibshirani, R. (2004). 1-norm support vector machines. Advances in Neural nformation Processing Systems 16.

[25] Zou, H. and Yuan, M. (2005). The F_∞-norm Support Vector Machine. Statisica Sinica. To appear.

Electronic Journal of Statistics. ISSN: 1935-7524