zbMATH Open
THE FIRST RESOURCE FOR MATHEMATICS

Rauschenberger, Armin; **Ciocănea-Teodorescu, Iuliana**; **Jonker, Marianne A.**; **Menezes, Renée X.**; **van de Wiel, Mark A.**

Sparse classification with paired covariates. (English) Zbl 1459.62007

Adv. Data Anal. Classif., ADAC 14, No. 3, 571-588 (2020).

Summary: This paper introduces the paired lasso: a generalisation of the lasso for paired covariate settings. Our aim is to predict a single response from two high-dimensional covariate sets. We assume a one-to-one correspondence between the covariate sets, with each covariate in one set forming a pair with a covariate in the other set. Paired covariates arise, for example, when two transformations of the same data are available. It is often unknown which of the two covariate sets leads to better predictions, or whether the two covariate sets complement each other. The paired lasso addresses this problem by weighting the covariates to improve the selection from the covariate sets and the covariate pairs. It thereby combines information from both covariate sets and accounts for the paired structure. We tested the paired lasso on more than 2000 classification problems with experimental genomics data, and found that for estimating sparse but predictive models, the paired lasso outperforms the standard and the adaptive lasso. The R package `palasso` is available from CRAN.

MSC:

62-04	Software, source code, etc. for problems pertaining to statistics
62J07	Ridge regression; shrinkage estimators (Lasso)
62J12	Generalized linear models (logistic models)
62H30	Classification and discrimination; cluster analysis (statistical aspects)
62P10	Applications of statistics to biology and medical sciences; meta analysis

Cited in **1** Document

Keywords:

prediction; sparsity; Lasso regression; paired data

Software:

CRAN; glmnet; SuperLearner; TCGAbiolinks; palasso; ipflasso; R; CorShrink; TANDEM

Full Text: DOI

References:

[1] Aben, N.; Vis, DJ; Michaut, M.; Wessels, LF, TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types, Bioinformatics, 32, 17, i413-i420 (2016)· doi:10.1093/bioinformatics/btw449

[2] Bergersen, LC; Glad, IK; Lyng, H., Weighted lasso with data integration, Stat Appl Genet Mol Biol, 10, 1, 39 (2011) · Zbl 1296.92017 · doi:10.2202/1544-6115.1703

[3] Boulesteix, AL; De Bin, R.; Jiang, X.; Fuchs, M., IPF-LASSO: Integrative \(L_1\)-penalized regression with penalty factors for prediction based on multi-omics data, Comput Math Methods Med, 2017, 7691937 (2017) · Zbl 1370.92016 · doi:10.1155/2017/7691937

[4] Bühlmann, P.; van de Geer, S., Statistics for high-dimensional data: methods, theory and applications (2011), Berlin: Springer, Berlin · Zbl 1273.62015

[5] Campbell, F.; Allen, GI, Within group variable selection through the exclusive lasso, Electron J Stat, 11, 2, 4220-4257 (2017) · Zbl 1408.62132 · doi:10.1214/17-EJS1317

[6] Colaprico, A.; Silva, TC; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, TS; Malta, TM; Pagnotta, SM; Castiglioni, I., TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data, Nucleic Acids Res, 44, 8, e71 (2016)· doi:10.1093/nar/gkv1507

[7] Cortes, C.; Mohri, M.; Thrun, S.; Saul, LK; Schölkopf, B., AUC optimization vs. error rate minimization, Advances in neural information processing systems 16, 313-320 (2004), Cambridge: MIT Press, Cambridge

[8] Dey KK, Stephens M (2018) CorShrink: empirical Bayes shrinkage estimation of correlations, with applications. bioRxiv 10.1101/368316

[9] Fan, J.; Lv, J., Sure independence screening for ultrahigh dimensional feature space, J R Stat Soc Ser B (Stat Methodol), 70, 5, 849-911 (2008) · Zbl 1411.62187 · doi:10.1111/j.1467-9868.2008.00674.x

© 2023 FIZ Karlsruhe GmbH

[10] Friedman, J.; Hastie, T.; Tibshirani, R., Regularization paths for generalized linear models via coordinate descent, J Stat Softw (2010)· doi:10.18637/jss.v033.i01

[11] Gade, S.; Porzelius, C.; Fälth, M.; Brase, JC; Wuttig, D.; Kuner, R.; Binder, H.; Sültmann, H.; Beißbarth, T., Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction in prostate cancer, BMC Bioinform, 12, 1, 488 (2011)· doi:10.1186/1471-2105-12-488

[12] Huang, J.; Ma, S.; Zhang, CH, Adaptive lasso for sparse high-dimensional regression models, Stat Sin, 18, 4, 1603-1618 (2008) · Zbl 1255.62198

[13] Huang, X.; Stern, DF; Zhao, H., Transcriptional profiles from paired normal samples offer complementary information on cancer patient survival-evidence from TCGA pan-cancer data, Sci Rep, 6, 20567 (2016)· doi:10.1038/srep20567

[14] Reid, S.; Tibshirani, R., Sparse regression and marginal testing using cluster prototypes, Biostatistics, 17, 2, 364-376 (2016)· doi:10.1093/biostatistics/kxv049

[15] Robinson, MD; Oshlack, A., A scaling normalization method for differential expression analysis of RNA-seq data, Genome Biol, 11, 3, R25 (2010)· doi:10.1186/gb-2010-11-3-r25

[16] Rodríguez-Girondo, M.; Kakourou, A.; Salo, P.; Perola, M.; Mesker, WE; Tollenaar, RA; Houwing-Duistermaat, J.; Mertens, BJ; Datta, S.; Mertens, BJ, On the combination of omics data for prediction of binary outcomes, Statistical analysis of proteomics, metabolomics, and lipidomics data using mass spectrometry, 259-275 (2017), Cham: Springer, Cham

[17] Shmulevich, I.; Zhang, W., Binary analysis and optimization-based normalization of gene expression data, Bioinformatics, 18, 4, 555-565 (2002)· doi:10.1093/bioinformatics/18.4.555

[18] Telonis, AG; Magee, R.; Loher, P.; Chervoneva, I.; Londin, E.; Rigoutsos, I., Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types, Nucleic Acids Res, 45, 6, 2973-2985 (2017)· doi:10.1093/nar/gkx082

[19] Ternès, N.; Rotolo, F.; Heinze, G.; Michiels, S., Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces, Biom J, 59, 4, 685-701 (2017) · Zbl 1369.62306 · doi:10.1002/bimj.201500234

[20] Tibshirani, R., Regression shrinkage and selection via the lasso, J R Stat Soc Ser B (Methodol), 58, 1, 267-288 (1996) · Zbl 0850.62538

[21] Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; Knight, K., Sparsity and smoothness via the fused lasso, J R Stat Soc Ser B (Stat Methodol), 67, 1, 91-108 (2005) · Zbl 1060.62049 · doi:10.1111/j.1467-9868.2005.00490.x

[22] van de Wiel, MA; Lien, TG; Verlaat, W.; van Wieringen, WN; Wilting, SM, Better prediction by use of co-data: adaptive group-regularized ridge regression, Stat Med, 35, 3, 368-381 (2016)· doi:10.1002/sim.6732

[23] van der Laan, MJ; Polley, EC; Hubbard, AE, Super learner, Stat Appl Genet Mol Biol, 6, 1, 25 (2007) · Zbl 1166.62387 · doi:10.2202/1544-6115.1309

[24] van Wieringen, WN; Kun, D.; Hampel, R.; Boulesteix, AL, Survival prediction using gene expression data: a review and comparison, Comput Stat Data Anal, 53, 5, 1590-1603 (2009) · Zbl 1453.62225 · doi:10.1016/j.csda.2008.05.021

[25] Westfall, PH; Armitage, P.; Colton, T., Combining \(P\) values, Encyclopedia of biostatistics (2005), Hoboken: Wiley, Hoboken

[26] Yuan, M.; Lin, Y., Model selection and estimation in regression with grouped variables, J R Stat Soc Ser B (Stat Methodol), 68, 1, 49-67 (2006) · Zbl 1141.62030 · doi:10.1111/j.1467-9868.2005.00532.x

[27] Zou, H., The adaptive lasso and its oracle properties, J Am Stat Assoc, 101, 476, 1418-1429 (2006) · Zbl 1171.62326 · doi:10.1198/016214506000000735

[28] Zou, H.; Hastie, T., Regularization and variable selection via the elastic net, J R Stat Soc Ser B (Stat Methodol), 67, 2, 301-320 (2005) · Zbl 1069.62054 · doi:10.1111/j.1467-9868.2005.00503.x

[29] Zwiener, I.; Frisch, B.; Binder, H., Transforming RNA-Seq data to improve the performance of prognostic gene signatures, PLoS ONE, 9, 1, e85150 (2014)· doi:10.1371/journal.pone.0085150

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.