Stat Med. 2015 Sep 13. doi: 10.1002/sim.6732. [Epub ahead of print]

Better prediction by use of co-data: adaptive group-regularized ridge regression.

van de Wiel MA[1,2], Lien TG[3], Verlaat W[4], van Wieringen WN[1,2], Wilting SM[4].

Author information

[1]Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands.
[2]Department of Mathematics, VU University, Amsterdam, The Netherlands.
[3]Department of Mathematics, University of Oslo, Oslo, Norway.
[4]Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands.

Abstract

For many high-dimensional studies, additional information on the variables, like (genomic) annotation or external p-values, is available. In the context of binary and continuous prediction, we develop a method for adaptive group-regularized (logistic) ridge regression, which makes structural use of such 'co-data'. Here, 'groups' refer to a partition of the variables according to the co-data. We derive empirical Bayes estimates of group-specific penalties, which possess several nice properties: (i) They are analytical. (ii) They adapt to the informativeness of the co-data for the data at hand. (iii) Only one global penalty parameter requires tuning by cross-validation. In addition, the method allows use of multiple types of co-data at little extra computational effort. We show that the group-specific penalties may lead to a larger distinction between 'near-zero' and relatively large regression parameters, which facilitates post hoc variable selection. The method, termed GRridge, is implemented in an easy-to-use R-package. It is demonstrated on two cancer genomics studies, which both concern the discrimination of precancerous cervical lesions from normal cervix tissues using methylation microarray data. For both examples, GRridge clearly improves the predictive performances of ordinary logistic ridge regression and the group lasso. In addition, we show that for the second study, the relatively good predictive performance is maintained when selecting only 42 variables. Copyright © 2015 John Wiley & Sons, Ltd.