Electronic Journal of Statistics
Vol. 0 (2020) 1–8
ISSN: 1935-7524

Optimal selection of sample-size dependent common subsets of covariates for multi-task regression prediction

David Azriel

Faculty of Industrial Engineering and Management,
Technion – Israel Institute of Technology, Haifa, Israel
e-mail: davidazr@technion.ac.il

Yosef Rinott

Department of Statistics and Center for the Study of Rationality
The Hebrew University, Jerusalem, Israel
e-mail: yosef.rinott@mail.huji.ac.il

Abstract:

An analyst is given a training set consisting of regression datasets D_j of different sizes, which are distributed according to some G_j, $j = 1, \ldots, \mathcal{J}$, where the distributions G_j are assumed to form a random sample generated by some common source. In particular, the D_j's have a common set of covariates and they are all labeled. The training set is used by the analyst for selection of subsets of covariates denoted by $\mathcal{P}^*(n)$, whose role is described next.

The multi-task problem we consider is as follows: given a number of random labeled datasets (which may be in the training set or not) D_{J_k} of size n_k, $k = 1, \ldots, K$, estimate separately for each dataset the regression coefficients on the subset of covariates $\mathcal{P}^*(n_k)$ and then predict future dependent variables given their covariates.

Naturally, a large sample size n_k of D_{J_k} allows a larger subset of covariates, and the dependence of the size of the selected covariate subsets on n_k is needed in order to achieve good prediction and avoid overfitting. Subset selection is notoriously difficult and computationally demanding, and requires large samples; using all the regression datasets in the training set together amounts to borrowing strength toward better selection under suitable assumptions. Furthermore, using common subsets for all regressions having a given sample size standardizes and simplifies the data collection and avoids having to select and use a different subset for each prediction task. Our approach is efficient when the relevant covariates for prediction are common to the different regressions, while the models' coefficients may vary between different regressions.

Last but not least, we propose a simple and meaningful measure, GENO, that allows comparisons of the predictive value of different subsets of covariates by comparing the sample size they require in order to achieve the same prediction error.

MSC2020 subject classifications: 62J99.
Keywords and phrases: random covariates, model selection, Mallows C_p, equivalent number of observations (ENO), GENO, transfer learning, overfitting.

arXiv: 2012.05949