

LARS: the how and why

Including variable constraints

Johan Van Kerckhoven

April 1st, 2011

In regression modelling tasks, one of the problems is selecting a subset of variables which adequately explain the response variable while at the same time being as sparse as possible. Variable selection criteria, such as AIC or SIC, are commonly used to compare candidate models and select the “best”. However, choosing candidate models is itself already a troublesome task for the researcher, as full subset search is unfeasible for even smallish datasets (with 30 or more variables). Hence, a procedure to choose candidate models is needed. One procedure to do this is the Least Angle Regression (LARS), developed by Efron, Hastie, Johnstone and Tibshirani (2004).

We first give a detailed explanation of the algorithm, and elaborate the reasoning behind each step. We also touch on the advantages and disadvantages of the procedure, and show how the algorithm can be modified to compute a full Lasso solution path. Finally, we examine how the LARS deals with variable “constraints”, meaning situations in which one variable cannot be included in the model unless another variable is included. Examples of this are interaction terms or higher order terms, where the researcher prefers to have the original variables included as well. We illustrate a possible disadvantage of the procedure outlined in Efron et al. and present a modification to the LARS algorithm that allows us to deal with this situation in a more suitable fashion.