Piecewise linear regularized solution paths

Saharon Rosset
IBM T.J. Watson Research Center

Ji Zhu
Department of Statistics
University of Michigan

Submission for a workshop presentation at NIPS 2003

Abstract

Regularization is critical for successful statistical modeling of *modern* data, which is high-dimensional, sometimes noisy and often contains a lot of irrelevant predictors. It exists *implicitly or explicitly* at the heart of all successful methods.

In this talk we consider the generic explicit regularized optimization problem:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} L(y, X\beta) + \lambda J(\beta),$$

where L is a loss function, J is a model complexity penalty function, and λ is the *tuning* regularization parameter.

Many methods for regression and classification, both traditional and modern, can be cast in this form: ridge regression, the Lasso, support vector machines and more.

This formulation gives rise to interesting statistical questions: what are good (loss L, penalty J) pairs? How should we determine the value of the regularization parameter λ? It also presents a computational challenge, as we would like to solve this problem for many values of λ to determine what a good solution is.

We tackle both of these aspects by characterizing (loss L, penalty J) pairs for which the regularized optimization problem can be solved efficiently for *all* values of λ, since the path of optimal solutions $\hat{\beta}(\lambda)$ follows a *piecewise linear* curve as a function of λ. This family turns out to include many interesting members, including 1-norm and 2-norm support vector machines, the Lasso and its extensions. We then use statistical motivations of robustness and sparsity to select interesting (loss L, penalty J) pairs. We give explicit algorithms for generating the path of regularized solutions to these problems. The resulting methods are adaptable (because we can choose an *optimal* regularization parameter), efficient (because we can generate the whole regularized path efficiently) and robust (because we choose to use robust loss functions).

This is joint work with Trevor Hastie (Stanford University) and Rob Tibshirani (Stanford University).