Cross-Validation Werks

Jean Feng*
Department of Biostatistics, University of Washington
and
Noah Simon
Department of Biostatistics, University of Washington

August 2, 2016

Abstract

In the setting of penalized regression, cross-validation is a widely used technique for tuning penalty parameters. With an oracle set of parameters, one can guarantee a particular rate of convergence of the prediction error, but it is unknown if cross-validation is able to recover the same rate. We prove that the model chosen from cross-validation will converge to the true model at the optimal rate since it converges the oracle at a near-parametric rate $(J(c + \kappa \log n)/n)^{1/2}$ where n is the number of samples and J is the number of penalty parameters. The results are counter to the common belief that increasing the number of penalty parameters drastically increase the model complexity. In fact, for nonparametric models, our error bounds allow the number of penalty parameters to increase with the number of samples while retaining the optimal rate. The proof allows cross-validation over an infinite set of penalty parameters and the lower limit of the range can decrease at any polynomial rate. For smooth regression problems, the proof only requires convexity of the loss and penalty functions; additional assumptions are required if the penalty functions are non-smooth. The proof uses techniques from entropy and an implicit differentiation trick. The simplicity of the proof may extend itself to other problems in cross-validation. Our simulation studies show that increasing the penalty parameters can substantially decrease model bias if one uses optimization algorithms that effectively minimize the validation loss.

Keywords: ...?

*Jean Feng was supported by NIH grants DP5OD019820 and T32CA206089. Noah Simon was supported by NIH grant DP5OD019820. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1 Introduction

In the usual regression setting, we have data (x, y) where x are the covariates and y is the response. The goal of model estimation is to minimize the prediction error of y as specified by some loss function L. If the problem is ill-posed or high-dimensional ($p >> n$), one needs to balance the "bias" and "variance" of the model in order to obtain good generalization error. A popular technique is to use regularization or penalization. There one introduces penalty functions into the model criterion to control model complexity and induce desired structure. The most common examples of regularization methods include the ridge penalty, lasso, and etc etc etc. Every penalty function is accompanied by a weight parameter that indicates how strongly the structure corresponding to that penalty should be enforced. Since the penalty parameters determine the fitted model, it is important to select these parameters properly.

A popular strategy to tune the penalty parameters is by cross-validation (CV), which is based on the simpler algorithm called holdout. In holdout, one randomly splits the data into two sets. Models are trained on one partition of the data and its generalization error is estimated on the other half. The algorithm then chooses the model with the minimum estimated risk. CV splits the data into multiple partitions, performs holdout over the partitions, and then selects the model with the minimum average loss. CV's popularity can be explained by two main traits: one, CV can be applied to almost any algorithm in almost any framework since it only assume that the data are independent, and two, it has been shown to be highly effective in practice. Hence, many authors have recommended using CV to tune parameters in penalized regression problems (e.g. lasso, elastic net, etc).

However, there is little theoretical foundation for CV though a lot of attempts have been made. Van Der Laan attempts to build a general theory for CV, but (I can't read his papers). Mitchell approaches the problem using entropy methods, but requires strong assumptions on the design matrix when applied to the simple problem of the Lasso. Others have taken a more specific approach. In regression, Gyorfi (2002) proved asymptotic results for a truncated least squares estimate and Wegkamp (2003) proved an oracle inequality for a penalized least squares holdout procedure. Even more specifically, Golub, Heath and Wahba proved CV for penalized ridge regression (read) and Chaterjee and Chetverikov address the lasso. (Are there

drawbacks to their methods?) In classification, Kearns (1997) proved an oracle inequality when using piecewise constant classifiers. In density estimation, van der laan shows the asymptotic optimality of CV. For a more complete review of CV methods, refer to Arlot.

Our paper addresses CV in the penalized regression setting and provides a finite sample upper bound for the prediction error of a model selected by CV. We find that this CV-selected model converges towards the oracle model at a near-parametric rate. Hence for most nonparametric problems, the convergence rate of the CV model to the true model is dominated by convergence rate of oracle. Unlike many of the previous methods, we allow CV to be performed over an infinite (possibly uncountable) number of penalty parameters and (?) allow the lower bound on the penalty parameters to decrease at any polynomial rate with respect to the number of samples. (Does anyone even do multiple penalty parameters??) The proof is based on entropy methods (sara's book) and an implicit differentiation trick (bengio, Foo, Feng). We assume that the loss and penalty functions are convex and smooth. If the penalty is non-smooth but smooth almost everywhere, such as in the Lasso, further assumptions are needed on the behavior of the the directional derivatives and the local optimality space. The assumptions here are minimal.

Important extension: What if we don't find the minimizer of the validation error or the training error. Does this proof fail? Does this method fail? Is this a minimal assumption.

This inequality bound also provides insight into how to improve model estimation. The error bounds suggest that one way to reduce generalization error is to increase the number of penalty parameters with the sample size, contrary to popular belief. However, it also gives an upper bound for the number of penalty parameters to tune. We support this idea through simulation studies.

Section 1 provides the theorem. Section 2 applies our technique to various problems. Section 3 provides simulation studies. Section 4 is a discussion and bids you farewell. Section 5 gives the hairy proof details.

2 Discussion