Available online at www.sciencedirect.com
SCIENCE DIRECT®

Statistical Methodology 2 (2005) 131–154

Statistical Methodology
Official Journal of the International Indian Statistical Association

www.elsevier.com/locate/stamet

Asymptotics of cross-validated risk estimation in estimator selection and performance assessment

Sandrine Dudoit*, Mark J. van der Laan

Division of Biostatistics, University of California, Berkeley, 140 Earl Warren Hall, #7360, Berkeley, CA 94720-7360, United States

Received 27 December 2003

Abstract

Risk estimation is an important statistical question for the purposes of selecting a good estimator (i.e., model selection) and assessing its performance (i.e., estimating generalization error). This article introduces a general framework for cross-validation and derives distributional properties of cross-validated risk estimators in the context of estimator selection and performance assessment. Arbitrary classes of estimators are considered, including density estimators and predictors for both continuous and polychotomous outcomes. Results are provided for general full data loss functions (e.g., absolute and squared error, indicator, negative log density). A broad definition of cross-validation is used in order to cover leave-one-out cross-validation, V-fold cross-validation, Monte Carlo cross-validation, and bootstrap procedures. For estimator selection, finite sample risk bounds are derived and applied to establish the asymptotic optimality of cross-validation, in the sense that a selector based on a cross-validated risk estimator performs asymptotically as well as an optimal oracle selector based on the risk under the true, unknown data generating distribution. The asymptotic results are derived under the assumption that the size of the validation sets converges to infinity and hence do not cover leave-one-out cross-validation. For performance assessment, cross-validated risk estimators are shown to be consistent and asymptotically linear for the risk under the true data generating distribution and confidence intervals are derived for this unknown risk. Unlike previously published results, the theorems derived in this and our related articles apply to general data generating distributions, loss functions (i.e., parameters), estimators, and cross-validation procedures.

* Corresponding author. Tel.: +1 510 643 1108; fax: +1 510 643 5163.
E-mail address: sandrine@stat.berkeley.edu (S. Dudoit).

1572-3127/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2005.02.003

Keywords: Asymptotic linearity; Asymptotic optimality; Classification; Confidence interval; Cross-validation; Density estimation; Estimator selection; Generalization error; Indicator loss function; Loss function; L_2 loss function; Model selection; Multifold cross-validation; Performance assessment; Prediction; Quadratic loss function; Regression; Resubstitution estimator; Risk

1. Introduction

1.1. Motivation

Risk estimation is an important statistical question for at least two purposes. Risk estimation is used for: (i) *estimator selection*, or *model selection*, where the "best" estimator is chosen to minimize risk over a given class of estimators; (ii) *estimator performance assessment*, i.e., the estimation of *generalization error*. These two fundamental problems have been referred to variously in the statistical literature as "submodel selection and evaluation" [4] and "choice and assessment of statistical predictions" [32]. For example, regression problems often involve the data-driven selection of a predictor for an outcome Y given covariates W (e.g., linear model with k explanatory variables, regression tree with k terminal nodes), with the intention of predicting the outcome of interest for future observations. A common measure of performance in this context is the mean squared error between predicted and true responses, i.e., the risk for the quadratic loss function. An immediate difficulty is that the risk of a given estimator is the expected value of a loss function under the typically *unknown* data generating distribution. This means that the available data (i.e., the learning set or empirical distribution) have to be used for both tasks (i) and (ii), that is, to select a good estimator (specifically, estimate the risk criterion used to select the estimator) and to assess the performance of this selected estimator.

A number of approaches have been proposed for selection and performance assessment. As discussed by Breiman [4] in the context of dimensionality selection in regression, criteria such as Mallow's C_p, Akaike's information criterion (AIC), and the Bayesian information criterion (BIC), do not account for the data-driven selection of the sequence of estimators (i.e., submodels) and thus provide biased assessment of generalization error. Instead, risk estimation methods based on sample reuse have been favored in the recent literature. The main procedures include: leave-one-out cross-validation, V-fold cross-validation (i.e., random partition of the learning set into V mutually exclusive and exhaustive sets), Monte Carlo cross-validation (i.e., repeated random splits of the learning set into a training and a validation set), the jackknife, and the bootstrap [5,6,7 (Chapter 3), 8,10,14,15 (Chapter 17),16,17,19 (Chapters 7 and 8),21 (Chapter 7),23,24,27,29 (Chapter 3),30,32,33,39]. Another important class of approaches for model selection, described in [2], uses sieve theory to define penalized empirical loss criteria. Connections with cross-validation methods are discussed in [3].

A variety of *cross-validation* (CV) procedures are available for estimating the risk of a given estimator and for performing estimator selection. A natural question then concerns the distributional properties of the resulting risk estimators, i.e., their performances in terms of identifying a good estimator (model selection) and as estimators of generalization error,