A bound on the cross-validation estimate for algorithm assessment

Gianluca Bontempi Mauro Birattari

Iridia - CP 194/6
Université Libre de Bruxelles
email: {gbonte, mbiro}@ulb.ac.be

Abstract

Cross-validation methods are commonly used as an effective way to estimate from a finite data set the generalization properties of a function approximator. It is common belief that cross-validation, at the cost of an increased computational expense, returns an estimate of the real generalization error that is more reliable than the simplest resubstitution estimate. However, few results providing bounds on the accuracy of the cross-validated estimate are reported in literature. This paper suggests that the lack of significant results could be related to a misunderstanding of the quantity which is targeted by a cross-validation estimator. The paper thus distinguishes two different ways of assessing a learning procedure: the *hypothesis-based* approach and the *algorithm-based* approach. We show that while the well-known results of Vapnik's learning theory can be considered as an example of the *hypothesis-based* approach, cross-validation can be more profitably interpreted in the *algorithm-based* framework. Adopting the *algorithm-based* interpretation for cross-validation, we derive a new bound on its accuracy. Unlike previous results, the bound is independent of the Vapnik-Chervonenkis dimension of the hypothesis class, and provides insight into the behavior of cross-validation for large data sets.

1 Introduction

A supervised learning procedure consists in a learning algorithm L which takes as an input a training set D_N made of N input-output pairs, and returns one hypothesis $h(\cdot, \alpha_N)$ chosen in a class Λ. A vast amount of literature focused on the problem of assessing the learning procedure by estimating the generalization ability of the hypothesis $h(\cdot, \alpha_N)$ returned by L. Many estimates of the generalization ability have been proposed and examined in the literature, as the *training error*, the *cross-validation* estimates [8] and the *hold-out* estimate [1]. While a large amount of results have been found in the case of the training error, there are still few results concerning the reliability of the cross-validation estimates. This happens in spite of the common attitude in the learning community which considers cross-validation as more reliable than the training error in measuring the generalization power of a learning machine .

Goal of previous works on cross-validation was to bound the deviation between the cross-validation estimate and the generalization error $R(\alpha_N)$ of the selected hypothesis. To our knowledge, some results are reported in the work of Rogers and Wagner [6], and Devroye and Wagner [2] who proved that for several local algorithms, the leave-one-out estimate can be as close as $O(1/\sqrt{N})$ to $R(\alpha_N)$. A good description of this work can be found also in [1]. These results suggest the leave-one-out as preferable to the training error, yielding an estimate of the true error whose accuracy is independent of any notion of hypothesis complexity.

So far, however, for a large class of learning algorithms, no theoretical proof of the expected supremacy of the cross-validation over the resubstitution estimate has been found. Kearns and Ron [5] proved that, for training error minimization algorithms, the error of the leave-one-out estimate is not much worse than the worst case behavior of the training error estimate (sanity-check bound). Other results on accuracy of the cross-validation estimate were proposed by Holden [4], but, as referred by the author, they did not obtain an improvement over the bound for the resubstitution estimate as desired.

Our paper explores the idea that the lack of definitive results might be a consequence of the particular interpretation of cross-validation adopted by these authors. In fact, these results are based on the idea that cross-validation is an estimate of the true generalization error $R(\alpha_N)$ of the hypothesis function $h(\cdot, \alpha_N)$ chosen by the learning algorithm L. Cross-validation is then viewed essentially as a measure of the performance of a single hypothesis $h(\cdot, \alpha_N)$. This approach requires some form of ***stability*** of the learning algorithm. If the removal of even a single example from the training sample causes the learning algorithm to jump to a different hypothesis $h(\cdot, \alpha_{N-1})$ with much larger error than the full-sample hypothesis $h(\cdot, \alpha_N)$, it seems hard to expect the leave-one-out estimate to be accurate. A similar concern was raised by Holden [4], who stated the difficulty of studying the deviation between the true error of some hypothesis $h(\cdot, \alpha_N)$ and an estimate derived from a different hypothesis $h(\cdot, \alpha_{N-1})$. In this context, it appears problematic to justify why cross-validation estimates are currently deemed more reliable than simple resubstitution estimates.

Here, adopting a different interpretation of the cross-validation procedure, we derive a general upper bound on the cross-validation accuracy. The idea is founded on the definition of two different ways of assessing a learning procedure: the ***hypothesis-based*** approach and the ***algorithm-based*** approach. These two approaches measure the performance of the learning procedure referring to two different indices of performance: the ***hypothesis-based*** approach addresses the generalization error of the hypothesis α_N while the ***algorithm-based*** approach measures the average performance of the algorithm L on training sets of size N. We state that an interpretation of cross-validation in the ***algorithm-based*** framework may be more convenient and more related to its current use among the machine learning practitioners. We show that when cross-validation is intended as an estimator of the algorithm performance, rather than an assessment of the selected hypothesis, a general upper bound can be derived, independently of the complexity of the hypothesis class and of any notion of stability. In fact, while a notion of stability is mandatory if the cross-validation is seen as an estimate of the generalization of the hypothesis $h(cdot, \alpha_N)$, it is no more relevant for measuring the sensitivity of the learning algorithm to different realizations of the training set. However, the aim of the paper is not to demonstrate the superiority of an approach over the other but simply to illustrate how an alternative interpretation of the assessment process of a learning machine can open the way to interesting new developments.

In order to make an analysis of the learning problem, we first need to introduce some terminology and to define a number of mathematical objects.

These are the main actors of the learning problem:

- A data generator of casual input vectors $x \in X \subset \Re^n$ independent and identically distributed according to some unknown (but fixed) input probability distribution $\Pi(x)$.
- A *target* operator, which transforms the vectors x into the output values $y \in Y \subset \Re$ according to some unknown (but fixed) conditional distribution $P_f(y|x)$ (this includes the simplest case where the target implements some function $y = f(x)$). By definition the input distribution $\Pi(x)$ is independent of $P_f(y|x)$.
- A *training set* $D_N = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ made of N pairs $(x_i, y_i) \in Z = X \times Y$ independent and identically distributed according to the joint distribution $P(z) = P((x, y)) = P_f(y|x)\Pi(x)$. Then, $D_N \in Z^N = (X \times Y)^N$
- A learning machine having two components:
  1. A class of *hypothesis* functions $h(x, \alpha)$ with $\alpha \in \Lambda$. We consider only the case where the functions $h(\cdot, \alpha)$ are single valued mappings.
  2. An *algorithm* L which takes as input the training set D_N and returns as output one hypothesis function $h(\cdot, \alpha)$ with $\alpha \in \Lambda$. Throughout the paper, we will consider only the case of *deterministic* and *symmetric* algorithms. This means that they always give the same $h(\cdot, \alpha_N)$ for the same data set D_N and that they are insensitive to the ordering of the examples in D_N, respectively.

     The hypothesis selection is done according to Empirical Risk Minimization (ERM) principle where

     $$\alpha_N = \arg\min_{\alpha \in \Lambda} R_{emp}(\alpha) \tag{1}$$

     is the hypothesis which minimizes the empirical risk

     $$R_{emp}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} C\left(y_i, h(x_i, \alpha)\right) \tag{2}$$

     constructed on the basis of the data set D_N. The empirical risk is often referred to as the *training error* or as the *resubstitution estimate.*

- A *cost* C associated with a particular $f(x)$ and a particular $h(x, \alpha)$, given by a loss function $C(f(x), h(x, \alpha))$ which measures, given an input x, the discrepancy between the output of the supervisor and the output of the selected hypothesis.
- A *functional risk* which averages over the XY-domain the cost C for a given hypothesis $h(\cdot, \alpha_N)$:

$$R(\alpha_N) = E_{x,y}[C|D_N] = \int_{X,Y} C(y, h(x, \alpha_N)) dP_f(y|x) d\Pi(x) \tag{3}$$

- The average of the cost C for a given input x over the ensemble of training sets with N samples:

$$R(x, N) = E_{D_N,y}[C|x] = \int_{Z^N,Y} C(y, h(x, \alpha_N)) dP_f(y|x) dP^N(D_N) \tag{4}$$

  In the case of a quadratic cost function, this quantity is usually referred to as the *mean squared error* (MSE).

2 The assessment of a learning machine

The generalization error of the learning machine can be evaluated at three different levels.

Class of hypotheses: Let $\alpha_0 = \arg\min_{\alpha \in \Lambda} R(\alpha)$ be the vector of parameters of the hypothesis which best approximates the target in the class Λ according to the criterion (3). Here, we assume for simplicity that there exists a minimum value of $R(\alpha)$ achievable by a function in the class Λ. We define with $R(\alpha_0)$ the *generalization error of the class of hypotheses.*

The algorithm: Using Eq. (4) we define the quantity

$$R(N) = \int_X R(x, N) d\Pi(x) \tag{5}$$

that represents the *generalization error of the algorithm L.* In the case of a quadratic cost function this quantity is referred to as the *mean integrated squared error* (MISE).

The single hypothesis. Let α_N be the vector of parameters of the hypothesis generated by the algorithm for a training set D_N according to Eq. (1). By using Eq. (3) we define with $R(\alpha_N)$ the *generalization error of the hypothesis* $h(\cdot, \alpha_N)$.

The three criteria correspond to three different ways to assess the learning machine: the first quantifies the generalization error of the best approximation in the class Λ, the second assesses the average performance of the algorithm over training sets with N samples, the third is a measure to assess the specific hypothesis chosen by the learning machine. All these quantities should be compared to the minimal risk that can be attained by a single valued mapping. To this aim let us define with Λ^* the set of all possible single valued mappings $f : X \to Y$ and consider the quantity $\alpha^* = \arg\min_{\alpha \in \Lambda^*} R(\alpha)$. Thus $R(\alpha^*)$ represents the absolute minimum rate of error obtainable by a single valued approximation for the unknown target.

This allows us to decompose the problem of learning into two subproblems. First, define with $\hat{R}(N) - R(\alpha^*)$ the *learning error* of the algorithm L where $\hat{R}(N)$ is an estimate of (5) and with $\hat{R}(\alpha_N) - R(\alpha^*)$ the *learning error* of the hypothesis $h(\cdot, \alpha_N)$, with $\hat{R}(\alpha_N)$ an estimate of $R(\alpha_N)$.

Consider the two equalities for the algorithm L and the hypothesis $h(\cdot, \alpha_N)$, respectively:

$$\hat{R}(N) - R(\alpha^*) = \big(\hat{R}(N) - R(N)\big) + \big(R(N) - R(\alpha^*)\big) \tag{6}$$

$$\hat{R}(\alpha_N) - R(\alpha^*) = \big(\hat{R}(\alpha_N) - R(\alpha_0)\big) + \big(R(\alpha_0) - R(\alpha^*)\big) \tag{7}$$

It is common practice to define the first right-hand term as *estimation error* and the second term as *approximation error* [1]. The decomposition of the learning errors leads to a decomposition of the learning procedure into two steps:

1. The *error estimation* where the set of available data is used to estimate either the *algorithm-based* criterion or the *hypothesis-based* criterion for a fixed class Λ.

2. The *model selection* where different classes of hypotheses are evaluated and compared in order to select the one which will minimize the learning error.

These procedures are strongly characterized by which criterion is used for assessing the learning machine from a finite set of data. We will distinguish two different ways of assessing a learning procedure: the *hypothesis-based* and the *algorithm-based* approach. The distinction is made in relation to which measure of generalization error is adopted.

Hypothesis-based approaches aim to minimize the learning error (7). As a consequence, the error estimation problem focuses on the difference $\hat{R}(\alpha_N) - R(\alpha_0)$ while the model selection procedure aims to minimize the quantity $R(\alpha_0) - R(\alpha^*)$. The statistical learning theory proposed by Vapnik [9, 10] is a major example of a *hypothesis-based* approach. As far as the estimation error is concerned, Vapnik first defines $\hat{R}(\alpha_N) = R_{emp}(\alpha_N)$, then bounds the accuracy of $R_{emp}(\alpha_N)$ as an estimator of $R(\alpha_0)$ with a term depending on the number of points and a parameter (VC dimension) describing some general properties of the hypothesis class. In terms of model selection, Vapnik proposes the Structural Risk Minimization (SRM) procedure where the bound on the empirical error provides a constructive tool to select the desired model complexity.

Algorithm-based approaches address the learning error (6). In these approaches the error estimation procedure evaluates the quantity $\hat{R}(N) - R(N)$ while model selection targets the quantity $R(N) - R(\alpha^*)$. In this category we will include all the resampling statistics techniques which can be used to assess the average performance of an algorithm L. These techniques include well-known and largely used procedures in data analysis and statistics, like cross-validation and bootstrap methods.

Henceforth, for reasons of space we will focus exclusively on the error estimation problem in an algorithm-based approach.

3 A bound on the cross-validation estimate

Consider the problem of estimating the quantity $R(N)$ from the training set. The empirical risk $R_{emp}(\alpha_N)$ is seen as the most obvious estimate of $R(N)$. However, it is well known that the empirical risk is a biased estimate of $R(N)$ and that tends to be smaller than $R(N)$, because the same data have been used both to construct and to evaluate $h(\cdot, \alpha_N)$. The study of error estimates other than the resubstitution is of significant importance if we wish to obtain results applicable to practical learning scenarios.

Cross-validation [8] is a well-known method in sampling statistics to circumvent the limits of the resubstitution estimate. The basic idea is to build a model from one part of the data and then use that model to predict the rest of the data. The dataset D_N is split l times in a training and a test part, the first containing N_{tr} samples, the second containing $N_{ts} = N - N_{tr}$ samples. Each time N_{tr} examples are used by L to select a hypothesis $h(\cdot, \alpha^i_{N_{tr}})$ $i = 1, \dots, l$ from Λ and the remaining N_{ts} samples are used to estimate the error of $h(\cdot, \alpha^i_{N_{tr}})$

$$R_{ts}(\alpha^i_{N_{tr}}) = \sum_{j=1}^{N_{ts}} C\big(y_j, h(x_j, \alpha^i_{N_{tr}})\big) \tag{8}$$

The resulting average of the l errors is the cross-validation estimate

$$R_{cv}(N) = \frac{1}{l} \sum_{i=1}^{l} R_{ts}(\alpha^i_{N_{tr}}) \tag{9}$$

A common form of cross-validation is the "leave-one-out". In this case l equals the number of training samples and $N_{ts} = 1$. It is well-known in literature [1] that R_{cv} returns an unbiased estimate of $R(N_{tr})$ for any symmetric learning algorithm; thus $R_{cv}(N)$ should be viewed more as an estimator of $R(N_{tr})$ than of $R(N)$. In the following we will provide a bound on the discrepancy between the cross-validation estimate $R_{cv}(N)$ and $R(N_{tr})$.

Previous results in the literature focused on the *hypothesis-based* problem of deriving bounds on the quantity $|R(\alpha_N) - R_{cv}(N)|$. In particular, [5] and [4] studied the probability that the cross-validation estimate differs from the true generalization error by more than a specific constant ε, under quite general conditions.

Here, instead, we study the consistency of the cross-validation as an estimate of the *algorithm-based* cost $R(N_{tr})$. To this aim we intend to bound the quantity:

$$|R(N_{tr}) - \hat{R}(N_{tr})| = |R(N_{tr}) - R_{cv}(N)| \tag{10}$$

Let $C(f, h)$ be a real-valued bounded function where $A \leq C(f, h) \leq B$. By applying to Eq. (8) the Hoeffding inequality for sums of independent identically distributed bounded variables [3] we have for $i = 1, \dots, l$:

$$P\big\{|R(\alpha^i_{N_{tr}}) - R_{ts}(\alpha^i_{N_{tr}})| > \varepsilon\big\} < e^{\frac{-2\varepsilon^2 N_{ts}}{(B-A)^2}} \tag{11}$$

Consider now the quantity (9). The following relation holds:

$$\Big|\frac{1}{l}\sum_{i=1}^{l} R(\alpha^i_{N_{tr}}) - \frac{1}{l}\sum_{i=1}^{l} R_{ts}(\alpha^i_{N_{tr}})\Big| \le \frac{1}{l}\sum_{i=1}^{l} |R(\alpha^i_{N_{tr}}) - R_{ts}(\alpha^i_{N_{tr}})| \qquad (12)$$

Then from (11) we have

$$P\Big\{\Big|\frac{1}{l}\sum_{i=1}^{l} R(\alpha^i_{N_{tr}}) - \frac{1}{l}\sum_{i=1}^{l} R_{ts}(\alpha^i_{N_{tr}})\Big| > \varepsilon\Big\} \le P\Big\{\frac{1}{l}\sum_{i=1}^{l} |R(\alpha^i_{N_{tr}}) - R_{ts}(\alpha^i_{N_{tr}})| > \varepsilon\Big\} \le$$

$$\le P\Big\{\max_{i=1,\dots,l} |R(\alpha^i_{N_{tr}}) - R_{ts}(\alpha^i_{N_{tr}})| > \varepsilon\Big\} < e^{\frac{-2\varepsilon^2 N_{ts}}{(B-A)^2}} \qquad (13)$$

Further, from the Hoeffding inequality for U-statistics [3]

$$P\Big\{\Big|R(N_{tr}) - \frac{1}{l}\sum_{i=1}^{l} R(\alpha^i_{N_{tr}})\Big| > \varepsilon\Big\} =$$

$$= P\Big\{\Big|E_{D_{N_{tr}}}[R(\alpha^i_{N_{tr}})] - \frac{1}{l}\sum_{i=1}^{l} R(\alpha^i_{N_{tr}})\Big| > \varepsilon\Big\} < e^{\frac{-2\varepsilon^2 l_{\text{eff}}}{(B-A)^2}} \qquad (14)$$

where $l_{\text{eff}} = \lceil N/N_{tr} \rceil$ is the largest integer contained in N/N_{tr} and denotes the effective number of independent samples that can be extracted from D_N to estimate $R(N_{tr})$. Notice that the application of the Hoeffding theorem for U-statistics in (14) allows us to avoid any unrealistic assumption of independence among the cross-validated training sets. Since

$$R(N_{tr}) - R_{cv}(N) = R(N_{tr}) - \frac{1}{l}\sum_{i=1}^{l} R(\alpha^i_{N_{tr}}) + \frac{1}{l}\sum_{i=1}^{l} R(\alpha^i_{N_{tr}}) - R_{cv}(N)$$

it follows from (13) and (14) that

$$P\{|R(N_{tr}) - R_{cv}(N)| > \varepsilon\} < e^{\frac{-2\varepsilon^2 l_{\text{eff}}}{(B-A)^2}} + e^{\frac{-2\varepsilon^2 N_{ts}}{(B-A)^2}} \qquad (15)$$

This bound is independent of any definition of complexity of the hypothesis class, and allows the definition of sufficient conditions for consistency. If we assume that the two conditions

$$\lim_{N\to\infty} N_{ts} = \infty \qquad (16)$$

$$\lim_{N\to\infty} \frac{N}{N_{tr}} = \lim_{N\to\infty} \frac{N}{N - N_{ts}} = \infty \qquad (17)$$

are satisfied then the cross-validation returns a consistent estimate of $R(N)$. For instance, the relation $N_{tr} = \sqrt{N}$ satisfies the conditions.

Leave-one-out ($N_{ts} = 1$) and the resubstitution estimate ($N_{ts} = 0$) can be analyzed under this framework. These quantities do not satisfy the sufficient conditions for consistency: the leave-one-out does not satisfy the condition (16), while both the conditions (16) and (17) do not hold in the case of the resubstitution estimate. It is interesting to remark how similar conclusions for linear models were obtained in statistical literature by Shao [7].

4 Concluding remarks

The principles underlying the *hypothesis-based* and the *algorithm-based* approaches to learning are substantially different. The goal of the *hypothesis-based* approach is to estimate the performance of the selected hypothesis. The main assumption is that averaging over all possible training sets would be unnatural given the single realization available. Since the distribution of the data is not known, *hypothesis-based* methods search for distribution-free bounds. As a drawback, the results might be too conservative for a specific learning problem.

In the *algorithm-based* approach, a learned hypothesis is seen as a function of the data D_N. Since D_N is a random variable, the hypothesis is random as well and must be assessed averaging different realizations. It would be desirable to repeat several times the data generation and to run each time the learning algorithm. Unfortunately, the use of repeated realizations is not viable in a real learning problem. As an alternative, resampling methods are employed to simulate the stochastic process underlying the data.

In this paper, the definition of the categories *hypothesis-based* and *algorithm-based* did not aim to demonstrate the superiority of one approach over the other, but it had, in fact, the intention of providing new insight into cross-validation. We believe that this distinction is potentially fruitful and can lead to further developments.

References

[1] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer Verlag, 1996.

[2] L. Devroye and T. J. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE-IT*, 25(2):202–207, 1979.

[3] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association*, 58:13–30, 1963.

[4] S. B. Holden. Cross-validation and the pac learning model. Technical Report RN/96/94, Dept. of CS, Univ. College, London, 1996.

[5] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Tenth Annual Conference on Computational Learning Theory*, 1997.

[6] W. H. Rogers and T. J. Wagner. A fine sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3):506–514, 1978.

[7] J. Shao. Linear model selection by cross-validation. *Journal of American Statistical Association*, 88(422):486–494, 1993.

[8] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36(1):111–147, 1974.

[9] V. N. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4, Denver, CO, 1992.

[10] V. N. Vapnik. *Statistical Learning Theory.* Springer, 1998.