LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Mathias Fuchs, Xiaoyu Jiang, Anne-Laure Boulesteix

The computationally optimal test set size in simulation studies on supervised learning

Technical Report Number 189, 2016
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

The computationally optimal test set size in simulation studies on supervised learning

M.Fuchs[a*] X.Jiang[b] A.-L.Boulesteix[a]

[a]*Institut für Medizinische Informationsverarbeitung Biometrie und Epidemiologie, Ludwig-Maximilians-Universität, München, Germany;* [b]*Novartis Institutes for Biomedical Research, 250 Massachusetts Ave, Cambridge, MA 02139, United States*

(*v4.0 released January 2015*)

We consider simulation studies on supervised learning which measure the performance of a classification- or regression method based on *i.i.d.* samples randomly drawn from a prespecified distribution. In a typical setting, a large number of data sets are generated and split into training and test sets used to train and evaluate models, respectively. Here, we consider the problem of the choice of an adequate number of test observations.
In this setting, the expectation of the method's performance is independent of this choice, but the variance and hence the convergence speed may depend substantially on the trade-off between the number of test observations and the number of simulation iterations. Therefore, it is an important matter of computational convenience to choose it carefully.
Here, we show that this problem can be formulated in terms of a well-defined optimization problem that possesses a solution in terms of a simple closed-form expression. We give examples to show that the relative contributions of each term can vary considerably between data sets and settings. We discuss the statistical estimation of the solution, giving a confidence interval for the optimal number of test observations.

Keywords: simulation study; supervised learning;

AMS Subject Classification: 65C60; 68U20;

1. Introduction

Suppose we are interested in the performance of a considered supervised learning method in terms of the prediction error. An option is to use real data sets to estimate the performance. In this approach, each data set is split into non-overlapping training and test data sets. The considered learning method is used to construct a prediction rule from the training data set, and this prediction rule is subsequently applied to classify the test data. By comparing the true and predicted responses for this test data set, one estimates the prediction error. This procedure can be repeated for several splits into training and test sets. The well-known cross-validation procedure can be viewed as a variant of this. However, each data set's response and predictor variables follow an (unkown) distribution which typically leads to a high variance of the observed errors [1], making reliable comparisons of learning methods on real data difficult. In contrast, in a simulation study the underlying distribution is chosen by the experimenter who can, therefore, draw a large number of observations in order for the observed error to converge.

*Corresponding author. Email: fuchs@ibe.med.uni-muenchen.de

In the following, we will use the term "test chunk", defined as an arbitrary but small fixed number of test observations used to measure a model's performance. We treat any scenario where a performance measure on a metric scale is associated to a learned model and a test chunk in such a way that the interesting quantity is its expectation and can, consequently, be estimated by an average across several test chunks. We simultaneously treat regression where the performance is usually measured by the mean-squared error, and classification where the performance may be measured by an arbitrary loss function or sensitivity/specificity etc. Sometimes, one is interested in performance measures such as the area under the ROC-curve (AUC) that require several observations to compute, in contrast to measures such as the misclassification loss which are already defined on a single test observation, i.e., test chunks of size one. Thus, typically, the test chunk size can be set to one, but for AUC a larger number is required. Note that the expected value of the AUC does not depend on the test chunk size, nor does any other performance measure introduced above. In any case, we will always refer to the true expectation of the performance measure as to the "error" because the case of binary classification is the most intuitive example case, and to the estimated error computed as an average across several test chunks, as to the "observed error".

In some cases, it is advisable to investigate the error of the considered prediction method as a function of particular parameters of the distribution from which the data are drawn. For example, one might be interested in the relationship between the method's performance and distribution characteristics such as the correlation between predictors, the number of predictors having an effect on the response, or their effect sizes. Determining the prediction error of the prediction method of interest on simulated data may then be a valuable approach, which has often been taken in the literature [1].

The goal of this paper is to determine the optimal number of test chunks in this situation, providing guidance to design such a simulation in such a way that it achieves the most reliable approximation of the error within a given computation time. More precisely, we are going to determine the test chunk size that minimizes the total variance of the observed error. Before describing further the contribution of the paper, let us introduce a few important notions.

First, let us give a simple example of a joint distribution P to make things clearer. Consider a case with two continuous predictors X_1 and X_2 and a binary response Y. A joint distribution P for X_1, X_2 and Y is defined by letting the random vector $X = (X_1, X_2)^T$ follow a multivariate normal distribution with mean $(0, 0)^\top$ and identity covariance matrix, and then setting the conditional distribution of Y given X_1 and X_2 to $log(P(Y = 1)/P(Y = 0)) = \beta_0 + \sum \beta_i X_i$ for some coefficients β_1, β_2. The joint distribution P is thus fully specified. Logistic regression is a simple example of a method, which is known to perform well for this distribution P for a sufficiently large size g of the training data set since the relationship between response Y and predictors X_1 and X_2 follows exactly a logistic regression model.

One now defines the unconditional error as the true (expected) error for test data drawn from P of a prediction rule constructed with the considered method from a fixed data set consisting of g learning observations drawn from P, i.e., from a data set drawn from P^g. The unconditional error is to be contrasted with the error of a specific prediction rule, which corresponds to the error conditioned on the observation of a training data set. A simulation allows to approximate the unconditional error of the considered method for a given joint distribution P of response class and predictors, for instance the simple one defined above, and a given size of the training set. Note that in this definition the training data set of size g is considered as a random variable, the constructed regression/classification rule and its observed error are thus also random variables, and the

unconditional error is just the expectation of the latter. It is called unconditional error because it does not refer to a specific training data set. It is a parameter which depends only on the distribution P, the size g of the training data set and the considered regression/classification method.
the learning method sometimes involves a cross-validation resampling scheme, often with the goal to optimize an inner parameter. In such cases, some authors [2] speak of three different kinds of data: the training data, the validation data, and the test data, where the cross-validation happens on the first two.
Here, we will not treat the case involving cross-validation separately. We will always simply consider a training data set of size g, no matter whether the learning procedure internally splits these data or not, subsuming the validation data under what we call training data. Thus, Hastie's training and validation data together constitute what will be called training data in this note, whereas for the test set, consisting of all test chunks for a given learning set, our terminology coincides with Hastie's in all cases.
To approximate the unconditional error of a given learning method for a given P and a given g through a simulation, one has to repeat the following procedure a large number of times N:

(1) A training data set is randomly drawn from P^g.
(2) A prediction rule is constructed from this training data using the considered learning method.
(3) The error of this prediction rule is approximated by comparing the true and predicted response by means of a loss such as the mean squared error in the case of regression, or by comparing true and predicted classes for test data from a test data set in the case of classification. The test data set randomly drawn from P^n_{test}, where n_{test} denotes the size of the test data set; $P^{b \cdot n_{\text{test}}}$ where b is the test chunk size, in case $b \neq 1$, and n_{test} denotes the number of test chunks.

when this procedure has been performed a large number of times, the errors obtained in step three are averaged, yielding an approximation of the unconditional error. At his stage, it is important to note that:

(1) the more times this procedure is repeated (i.e. the larger N is), the better the approximation of the expectation is,
(2) the bigger the test data set used in step 3 to approximate the error of the prediction rule constructed in step 2 (i.e. the larger n_{test} is), the better the approximation of the error is.

Briefly, this simulation procedure involves two parameters N and n_{test} which should both be set to large values if a good approximation of the unconditional error is desired. In practice, computation time is limited, and one cannot increase N and n_{test} arbitrarily.
When test observations are not re-used, the procedure is guaranteed to converge to the true unconditional error Θ with probability one, as the number N of iterations converges to infinity, *no matter how many test observations are used in each iteration.* This is a consequence of the strong law of large numbers. (The convergence is almost sure, and therefore also in probability and in distribution).
Since the computation time available for a simulation study is in practice limited, one has to set n_{test} and N to reasonable values and hereby compromise between the precision of the approximation of the error of each prediction rule —this precision increases with n_{test}— and the precision of the expectation, which increases with N. To date, there is to our knowledge no literature on how to set N and n_{test} in practice to achieve the most

precise approximation of the unconditional error within a fixed computation time. Even worse, some researchers are not aware that n_{test} can and should be set to a large value: they simply set it to a (often small) value "typical for real data sets", thereby giving up one of the advantages of performing a simulation study.
In this note, we derive mathematically the optimal number of test chunks to achieve the best possible approximation of the unconditional error within a fixed computation time. Its practical usefulness is demonstrated through applications to several realistic scenarios (i.e. different distributions P, different training set sizes n and different learning methods) yielding very different optimal sizes n_{test} of the test data set.
The paper is structured as follows: In Section 2, we will carefully define the setup, in Section 3, we present the solution, in Section 4 we define an estimator for the optimal number of test chunks. Finally, Section 5 presents a calculation which shows that even in a very simple example the optimal number of test chunks can vary considerably.

2. Definitions and notations

In a simulation study, data are drawn from a pre-specified distribution P; in each iteration, a learning set (always of the same size) is drawn and its conditional error is assessed by means of test observations. The estimation target, the error Θ, is estimated by an average taken across several such iterations.
One may either re-use the same test observations across the iterations, or draw new ones for each iteration. The latter method leads to independence between the iterations, whence it yields a valid confidence interval for the error Θ–which is, obviously, a great advantage. The confidence interval is simply that for the mean, taken across the conditional errors, as it is implemented in any standard statistical software.
Suppose each learning set takes the same time C to draw and fit a model on. For practical reasons, it lends itself to draw test sets in chunks of, say, 100 observations, rather than one at a time. Also, as noted above, the usage of test chunks instead of test observations allows to treat the AUC as well. Suppose, furthermore, each such chunk of test observation on one learning set takes the same time B to draw and evaluate.
As usual, we will denote the predictors by X and the response variable by Y, and pairs (X, Y) in such a way that a learning sample is $((X_1, Y_1), \ldots, (X_g, Y_g)) = (Z_1, \ldots, Z_g)$ which we will abbreviate by Z_ℓ where the ℓ stands for "learning". We will denote the i-th learning set by Z_ℓ^i and the error of the decision rule learnt on the i-the learning set and evaluated on the j-the test chunk by K_{ij}, where $1 \leq i \leq N$ and $1 \leq j \leq n_{\text{test}}$. Thus, the data comes in the form of a $N \times n_{\text{test}}$-matrix K. Since the marginal expectation of any K_{ij} is the same, namely the true unconditional error which we will call Θ, the values K_{ij} can be considered as "elementary" estimators of the error. Also, we will consider an additional "generic" learning set $i = 0$ and test chunk $j = 0$ on which the error K_{00} is only considered as random variable.
Let us also denote the true conditional error of learning set $i = 1 \ldots N$ by $m_i(Z_\ell^i) := \mathbb{E}(K_{i0}|Z_\ell^i)$.
In case the researcher wants to re-use the same test observations for all learning sets–thereby sacrificing the independence between the learing iterations–the following remains approximately valid when B is taken to be the time for evaluation of a pre-existing test observation; in reality, the dependence between the learning iterations then leads to a slighty optimistic expression for the variance which then becomes very hard to estimate.

3. The optimal n_{test}

Let us denote the number of learning sets by N and the number of test observations (for each learning set) by n_{test}. Then, the total processor time required for the numerical study is approximately

$$t = N(C + Bn_{\text{test}}) \tag{1}$$

The conditional errors m_i are i.i.d. random variables. Each $m_i(Z_\ell^i)$, or m_i for short, is estimated by an average taken across n_{test} test observations:

$$\widehat{m_i} := n_{\text{test}}^{-1} \sum_{j=1}^{n_{\text{test}}} K_{ij}. \tag{2}$$

The obivous estimator of the unconditional error is the average of the m_i, taken across all i:

$$\widehat{\Theta} := N^{-1} n_{\text{test}}^{-1} \sum_{i,j} K_{ij} = N^{-1} \sum_{i=1}^{N} \widehat{m_i}$$

As introduced above, we ill, for sake of the argument, consider another random independent learning set with conditional error m_0 and test evaluations $K_{01}, K_{02}, \ldots$, and another random independent test chunk $j = 0$, with associated elementary error estimator K_{00} and conditional error $m_0 = m_0(Z_\ell^0) = \mathbb{E}(K_{00}|Z_\ell^0)$. The true variance of $\widehat{\Theta}$ is

$$\mathbb{V}(\widehat{\Theta}) = N^{-1}\mathbb{V}(\widehat{m_0}) \tag{3}$$

due to the independence between the learning iterations. The true variance can, of course, be expressed only in terms of the additional learning set $i = 0$ and test chunk $j = 0$ due to the i.i.d. setup. We are faced with the optimization problem

$$\mathbb{V}(\widehat{\Theta}) \to \text{ min}$$

subject to the constraint

$$t = \text{ const.} \tag{4}$$

The variance appearing in (3) can be understood by conditioning on the random variable m_1: By the law of total variance, we have

$$\mathbb{V}(\widehat{m_0}) = \mathbb{E}(\mathbb{V}(\widehat{m_0}|Z_\ell^0)) + \mathbb{V}(\mathbb{E}(\widehat{m_0}|Z_\ell^0)). \tag{5}$$

Both summands can be made more explicit. For the first, we calculate

$$\mathbb{V}(\widehat{m_0}|Z_\ell^0) = n_{\text{test}}^{-1}\mathbb{V}(K_{00}|Z_\ell^0) \tag{6}$$

where we resort, again, to the independent test chunk $j = 0$ using the i.i.d. assumption. The second summand of (5) is $\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))$ because $\widehat{m_0}$ is, of course, an unbiased

estimator of m_0, i.e., $\mathbb{E}(\widehat{m_0}|Z_\ell^0) = m_0$ by (2). Plugging this and (6) into (5), we obtain

$$\mathbb{V}(\widehat{m_0}) = n_{\text{test}}^{-1}\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)) + \mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0)) \tag{7}$$

The first summand describes the variances "within"; the second the variance "between" the learning sets. Thus, by (3)

$$\mathbb{V}(\widehat{\Theta}) = N^{-1}\left[n_{\text{test}}^{-1}\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)) + \mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))\right] \tag{8}$$

Abbreviating the constant of the right-hand side of (4) by T–the scheduled simulation running time– and using (1), we have a fixed relation between N and n_{test}:

$$N = (C + Bn_{\text{test}})^{-1}T$$

Plugging this into (8), we arrive at the optimization problem

$$(C + Bn_{\text{test}})T^{-1}\left[n_{\text{test}}^{-1}\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)) + \mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))\right] \to \min \tag{9}$$

which is by expanding the terms equivalent to

$$T^{-1}\left(Cn_{\text{test}}^{-1}\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)) + C\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0)) + B\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)) + Bn_{\text{test}}\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))\right) \to \min$$

for pre-specified T. Since only n_{test} can be chosen by the user, this is equivalent to

$$Cn_{\text{test}}^{-1}\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)) + Bn_{\text{test}}\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0)) \to \min$$

For the problem at hand we can view n_{test} as a real-valued variable, so we can derive the left-hand side by n_{test} and equal the result to zero

$$-Cn_{\text{test}}^{-2}\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)) + B\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0)) = 0$$

Solving this for n_{test}, we obtain for the optimal n_{test} the expression

$$n_{\text{test}} = \sqrt{\frac{C\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))}{B\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))}} \tag{10}$$

and one easily checks, using elementary analysis, that this expression indeed minimizes the left-hand side of (9). The formula (10) makes sense: The longer the learning procedure takes, compared to the duration of an evaluation, the more testing should be done in an iteration. Also, the higher the ratio of the conditional variance of K_{00} is, compared to the variance of the conditional errors, the more testing should be done in an iteration. For instance, suppose hypothetically, the variance $\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))$ across the learning sets is very small compared to that conditional variance. Then, there is no point in drawing many learning sets, and all variance is due to the conditional variance between the test observations. On the other hand, suppose the conditional variance between the test observations is very small, but the condtional errors vary a lot. Then, there is less need to assess each single conditional error to a very high precision.

4. Confidence intervals

Let us first construct an estimator and a confidence interval for the ratio $\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))/\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))$ on N' training sets and n'_{test} test chunks each (Since this estimation procedure is to be done before the simulation, we use N' and n'_{test} instead of N and n_{test}.) Thus, the data has the form of an $N' \times n'_{\text{test}}$-matrix.
We consider the estimator

$$v_i := (n'_{\text{test}}(n'_{\text{test}}-1))^{-1}\sum_{l\neq m}(K'_{il}-K'_{im})^2/2$$

as the obvious variance estimator of the i-th row as well as its average

$$v := (N'n_{\text{test}}(n_{\text{test}}-1))^{-1}\sum_{i=1}^{N'}\sum_{l\neq m}(K'_{il}-K'_{im})^2/2$$

taken over all rows. Their expectations are $\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))$, the mean row variance.
The variance estimator of the i-th row is (under normality) marginally distributed as

$$v \sim \frac{\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))}{n'_{\text{test}}-1}\chi^2_{n'_{\text{test}}-1}$$

which has variance

$$2(\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)))^2(n'_{\text{test}}-1)^{-1}$$

Thus, the variance of v is $2(\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)))^2(N'(n'_{\text{test}}-1))^{-1}$. Since a linear combination of chi-squares with different coefficients is difficult to treat analytically, we approximate the distribution of v by a single chi-square by matching the expected value $\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))$ and the variance $2(\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)))^2(N'(n'_{\text{test}}-1))^{-1}$ with that of

$$(\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)))(N'(n'_{\text{test}}-1))^{-1}\chi^2_{N'(n'_{\text{test}}-1)}$$

which is a very good approximation to the distribution of v.
In contrast, the term $\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))$ can be estimated by

$$w := \sum_{i\neq j}\big(n_{\text{test}}^{-1}\sum_l (K'_{il}-K'_{jl})\big)^2/2$$

which has expectation $\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))$ and is approximately distributed as

$$\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))(N'-1)^{-1}\chi^2_{N'-1}$$

Thus, the ratio v/w can be approximated by

$$\frac{(\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)))(N'(n'_{\text{test}}-1))^{-1}}{\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))(N'-1)^{-1}}\frac{N'(n'_{\text{test}}-1)}{N'-1}F_{N'(n'_{\text{test}}-1),N'-1} = \frac{\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))}{\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))}F_{N'(n'_{\text{test}}-1),N'-1}$$

and we can "invert" this information to a confidence interval for $\frac{\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))}{\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))}$ as follows: With probability $1-\alpha$, one has

$$\frac{\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))}{\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))}F_{N'(n'_{\text{test}}-1),N'-1}(\alpha/2) \leq \frac{v}{w} \leq \frac{\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))}{\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))}F_{N'(n'_{\text{test}}-1),N'-1}(1-\alpha/2)$$

which happens if and only if

$$\frac{v}{w}F_{N'-1,N'(n'_{\text{test}}-1)}(\alpha/2) \leq \frac{\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))}{\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))} \leq \frac{v}{w}F_{N'-1,N'(n'_{\text{test}}-1)}(1-\alpha/2)$$

The approximation gets better when new data are used for w so that v and w become independent.
Thus, a confidence interval at level $1-\alpha$ for $\sqrt{C\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))/(B\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0)))}$ is

$$\left[\sqrt{\frac{Cv}{Bw}F_{N'-1,N'(n'_{\text{test}}-1)}(\alpha/2)}, \sqrt{\frac{Cv}{Bw}F_{N'-1,N'(n'_{\text{test}}-1)}(1-\alpha/2)}\right]$$

5. An analytical example

In order to illustrate the within- and between-iterations variances, we are going to consider a simple example where these terms are analytically accessible. The random variable X is univariate with arbitrary distribution, and the joint distribution of (Y, X) is

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with β_0 and σ^2 unknown, and we suppose that β_1 is already known. Being given the generic learning sample $Z_\ell^0 = (X_1^0, Y_1^0, \ldots, X_g^0, Y_g^0)$ of size g, we introduce the natural abbreviation satisfying

$$Y_i^0 = \beta_0 + \beta_1 X_i^0 + \epsilon_i^0,$$

where $(\epsilon_1^0, \ldots, \epsilon_g^0)^T =: \epsilon_\ell^0$ is the i.i.d. vector from $\mathcal{N}(0, \sigma^2)$ of errors occuring in the generic learning sample. (In this analytical example, we only speak of the generic learning data and a generic test observations, so that no confusion with the index i introduced in Section 2 should arise.) the intercept estimator is the average residual

$$\widehat{\beta_0} = g^{-1}\sum_{i=1}^{g}(Y_i^0 - \beta_1 X_i^0) = g^{-1}\sum_{i=1}^{g}(\beta_0 + \beta_1 X_i^0 + \epsilon_i^0 - \beta_1 X_i^0) = \beta_0 + \overline{\epsilon_\ell^0}.$$

Then, the generic test observation is (X_0^0, Y_0^0) where $Y_0^0 = \beta_0 + \beta_1 X_0^0 + \epsilon_0^0$ where $\epsilon_0^0 \sim \mathcal{N}(0, \sigma^2)$, and the "generic" simulated squared loss can be written in terms of the errors

$\epsilon_\ell^0, \epsilon_0^0$ as

$$\begin{aligned} K_{00} &= \big(\widehat{\beta}_0 + \beta_1 X_0^0 - Y_0^0\big) \\ &= \big(\beta_0 + \overline{\epsilon_\ell^0} + \beta_1 X_0^0 - (\beta_0 + \beta_1 X_0^0 + \epsilon_0^0)\big) \\ &= (\overline{\epsilon_\ell^0} - \epsilon_0^0)^2 \end{aligned}$$

which follows the $(1+g^{-1})\sigma^2$-fold of a chi-squared distribution on one degree of freedom because $\overline{\epsilon_\ell^0}$ is normal with mean zero and variance σ^2/g and $\overline{\epsilon_\ell^0} - \epsilon_0^0$ is normal with mean zero and variance $\sigma^2(1+g^{-1})$. Thus, the true unconditional error is

$$\mathbb{E}K_{00} = \mathbb{E}((1+g^{-1})\sigma^2\chi_1^2) = (1+g^{-1})\sigma^2$$

and the total variance of K_{00} is

$$\mathbb{V}K_{00} = \mathbb{V}((1+g^{-1})\sigma^2\chi_1^2) = 2\sigma^4(1+2g^{-1}+g^{-2}). \tag{11}$$

The conditional error is

$$\begin{aligned} \mathbb{E}(K_{00}|\epsilon_\ell^0) &= \mathbb{E}((\overline{\epsilon_\ell^0} - \epsilon_0^0)^2|\epsilon_\ell^0) \\ &= \mathbb{E}((\overline{\epsilon_\ell^0})^2 - 2(\overline{\epsilon_\ell^0})\epsilon_0^0 + (\epsilon_0^0)^2|\epsilon_\ell^0) \\ &= (\overline{\epsilon_\ell^0})^2 + 0 + \sigma^2 \end{aligned}$$

which is distributed as σ^2 plus the $g^{-1}\sigma^2$-fold of a chi-square on one degree of freedom. This implies that the between-iterations variance is

$$\mathbb{V}(\mathbb{E}(K_{00}|\epsilon_\ell^0)) = \mathbb{V}(g^{-1}\sigma^2\chi_2^2) = 2\sigma^4 g^{-2}. \tag{12}$$

The conditional variance of K_{00}, given a learning sample Z_ℓ^0 with errors $\epsilon_\ell^0 = \epsilon_1^0, \ldots, \epsilon_g^0$, can be seen to be

$$\mathbb{V}(K_{00}|\epsilon_\ell^0) = 4(\overline{\epsilon_\ell^0})^2\sigma^2 + 2\sigma^4$$

after a short calculation. Thus, the average within-iterations variance is

$$\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0)) = \mathbb{E}\big(4(\overline{\epsilon_\ell^0})^2\sigma^2 + 2\sigma^4\big) = 2\sigma^4(1+2g^{-1}). \tag{13}$$

One checks that the sum of (12) and (13) is (11), as the law of total variance implies. Thus, the ratio of within- to between-iterations variances is

$$\frac{\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))}{\mathbb{V}(\mathbb{E}(K_{00}|Z_\ell^0))} = \frac{2\sigma^4(1+2g^{-1})}{2\sigma^4 g^{-2}} = g^{-2} + 2g^{-3}$$

which decreases, as expected, in g. Thus, the larger g is, the fewer testing observations need to be drawn. One might conjecture that this relationship holds in more generality. As a side result, we have seen that error estimates after a test observations vary for a single learning set on the same order of magnitude as the conditional errors vary between learning sets. This might be interesting in itself.

Table 1. Estimation of n_{test} for a logistic learner on data from a logistic model. We compared three scenarios, each of which is represented in a row of the table. We report the resulting times A and B in nanoseconds, the point estimator for the variance ratio $\mathbb{E}(\mathbb{V}(K_{00}|Z_\ell^0))/vbl$, and the confidence interval for n_{test}. Clearly, the resulting n_{test} can vary considerably.

g	es	A	B	varRatioPoint	ntestLower	ntestPoint	ntestUpper
10	1000	6721594	107476.0	21.15	28.66	36.37	44.13
30	10	6909372	126840.4	20.95	26.78	33.78	40.83
1000	1	176303859	732111.9	33.58	71.49	89.93	108.49

Also, it seems that the ratio of between- to across-iteration variance merits independent attention in further research.

We drew of $p = 4$ features and a response with the logistic model using the linear predictor $X\beta$ where $\beta = (beta_0, 0, 0, 0)^T$ and fitted coefficients on $g = 10, 30, 1000$ learning observations with $\beta_0 = 1000, 10, 1$, respectively, by a usual support vector machine. The results are shown in Table 1, illustrating that the optimal number of test chunks can vary considerably.

Funding

This work was supported by a Novartis funding.

Supplemental material

The file simTestSize _ 1.0.tar.gz contains an R-package that contains the proposed methodology. It also contains a function `reproduce` that allows to reproduce the results of this paper.

References

[1] Dougherty ER, Zollanvari A, Braga-Neto UM. The illusion of distribution-free small-sample classification in genomics. Current genomics. 2011;12(5):333.

[2] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. Springer Series in Statistics; Springer, New York; 2009; data mining, inference, and prediction; Available from: http://dx.doi.org/10.1007/978-0-387-84858-7.