Hypothesis Testing for Cross-Validation

Article

2 authors, including:

Y. Bengio
Université de Montréal
921 PUBLICATIONS **533,678** CITATIONS

SEE PROFILE

Hypothesis Testing for Cross-Validation

Yves Grandvalet
CNRS, Heudiasyc UMR 6599
UTC, Génie Informatique
B.P. 20529, 60205 Compiègne, France
Yves.Grandvalet@utc.fr

Yoshua Bengio
Dept. IRO, Université de Montréal
P.O. Box 6128, Downtown Branch, Montreal, H3C 3J7, QC, Canada
bengioy@iro.umontreal.ca

Technical Report 1285
Département d'Informatique et Recherche Opérationnelle

August 29th, 2006

Abstract

K-fold cross-validation produces variable estimates, whose variance cannot be estimated unbiasedly. However, in practice, one would like to provide a figure related to the variability of this estimate. The first part of this paper lists a series of restrictive assumptions (on the distribution of cross-validation residuals) that allow to derive unbiased estimates. We exhibit three such estimates, corresponding to differing assumptions. Their similar expressions however entail almost identical empirical behaviors. Then, we look for a conservative test for detecting significant differences in performances between two algorithms. Our proposal is based on the derivation of the form of a t-statistic parametrized by the correlation of residuals between each validation set. Its calibration is compared to the usual t-test. While the latter is overconfident in concluding that differences are indeed significant, our test is bound to be more skeptical, with smaller type-I error.

1 Introduction

The empirical assessment of performance of a new algorithm is one of the main steps in its introduction. The most convincing approach consists in benchmarking the new algorithm on a series of real data sets, comparing it with one or more baseline algorithms. For real data sets, the underlying distribution is

unknown. Hence, the generalization error cannot be measured. It has to be estimated on hold-out test sets or by means of resampling schemes such as K-fold cross-validation. Drawing statistically significant conclusions thus requires to evaluate the uncertainty of these estimates.

While K-fold cross-validation is recognized as the mehod of choice for estimating the generalization error for small sample sizes [1, 2], it suffers from one major drawback: variance. Bengio and Grandvalet show that there exists no universal (valid under all distributions) unbiased estimator of the variance of K-fold cross-validation [3]. Here, we derive a set of variance estimates, which are unbiased under restrictive assumptions on the distribution of cross-validation residuals. Then, we derive a t-statistic parametrized by the between-block correlation and based on it, we propose a conservative estimate of the uncertainty in the cross-validation estimation of performance difference between algorithms.

2 Framework and background

2.1 Algorithm evaluation

We dispose of a data set $D = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, with $\mathbf{z}_i \in \mathcal{Z}$, assumed independently sampled from an unknown distribution P. We have a learning algorithm $A : \mathcal{Z}^* \to \mathcal{F}$, which maps data sets of (almost) arbitrary size to predictors. [1] Let $f = A(D)$ be the function returned by algorithm A on the training set D. The discrepancy between a prediction and an observation $\mathbf{z}$ is measured by a loss functional $L : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$. For example one may take $L(f, (x, y)) = (f(x) - y)^2$ in regression , and $L(f, (x, y)) = 1_{f(x) \neq y}$ in classification.

In application-based evaluations, the goal of learning is usually stated as the minimization of the expected loss of $f = A(D)$ on future test examples:

$$\mathrm{PE}(D) = E[L(f, \mathbf{z})] \ , \tag{1}$$

where the expectation is taken with respect to $\mathbf{z} \sim P$.

In algorithm-based evaluations, we are not really interested in the performance on a specific training set. We want to compare algorithms on a more general basis [2], such as the expected performance of learning algorithm A over different training sets:

$$\mathrm{EPE}(n) = E[L(A(D), \mathbf{z})] \ , \tag{2}$$

where the expectation is taken with respect to $D \times \mathbf{z}$ independently sampled from $P^n \times \ P$. When P is unknown, EPE is estimated, and the most widely accepted estimator is cross-validation.

2.2 K-fold cross-validation estimation

In K-fold cross-validation, the data set D is first chunked into K disjoint subsets (or *blocks*) of the same size (to simplify the analysis below we assume that n is a multiple of K), with $m = n/K$. Following [3], we write T_k for the k-th such block, and D_k the training set obtained by removing the elements in T_k from D. The cross-validation estimator is

$$\mathrm{CV} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{m} \sum_{\mathbf{z}_i \in T_k} L(A(D_k), \mathbf{z}_i) \ , \tag{3}$$

[1] Here we consider symmetric algorithms, which are insensitive to the ordering of examples in the data set.

which is an unbiased estimate of EPE$(n-m)$. A version of cross-validation is specifically dedicated to comparing algorithms, using matched pairs

$$\Delta\text{CV} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{\mathbf{z}_i\in T_k} L(A_1(D_k),\mathbf{z}_i) - L(A_2(D_k),\mathbf{z}_i) \ . \tag{4}$$

In what follows, CV and ΔCV will generically be denoted by $\hat{\mu}$:

$$\hat{\mu} \ = \ \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mu_k} \ , \qquad \widehat{\mu_k} = \frac{1}{m}\sum_{i\in T_k} e_i \ ,$$

where, slightly abusing notation, $i \in T_k$ means $\mathbf{z}_i \in T_k$ and

$$\forall i \in T_k, \ e_i = \begin{cases} L(A(D_k),\mathbf{z}_i) & \text{for } \hat{\mu} = \text{CV} \ , \\ L(A_1(D_k),\mathbf{z}_i) - L(A_2(D_k),\mathbf{z}_i) & \text{for } \hat{\mu} = \Delta\text{CV} \ . \end{cases}$$

2.3 Variance of K-fold cross-validation

It is crucial to assess the uncertainty of $\hat{\mu}$ to derive trustworthy conclusions. We first recall three theoretical results of [3] that will be useful in the following Sections. (Lemma 1 and 2, and Theorem 1).

Lemma 1 *The variance of the cross-validation estimator is a linear combination of three moments:*

$$\theta = \frac{1}{n^2}\sum_{i,j}\text{Cov}(e_i,e_j) = \frac{1}{n}\sigma^2 + \frac{m-1}{n}\omega + \frac{n-m}{n}\gamma, \tag{5}$$

where $\text{Cov}(e_i,e_j) = E[e_ie_j] - E[e_i]E[e_j]$ is the covariance between variables e_i and e_j; σ^2 is the variance of $e_i, \ i = 1,\ldots,n$; ω is the covariance for residuals from the same test block ($(i,j) \in T_k^2, \ j \neq i$); γ is the covariance for residuals from different test blocks ($(i,j) \in T_k \times T_\ell, \ \ell \neq k$).

Lemma 2 *Let $\hat{\theta}$ be any quadratic estimate of* $\text{Var}[\hat{\mu}]$*, its expectation is a linear combination of three terms*

$$E[\hat{\theta}] = a(\sigma^2+\mu^2) + b(\omega+\mu^2) + c(\gamma+\mu^2) \ , \tag{6}$$

and a representer of estimators with this expected value is

$$\hat{\theta} = as_1 + bs_2 + cs_3 \ , \tag{7}$$

where (s_1,s_2,s_3) are defined from $\mathbf{e} = (e_1,\ldots,e_n)^T$ as follows:

$$\begin{cases} s_1 & \triangleq & \frac{1}{n}\sum_{i=1}^{n} e_i^2 \ , \\ s_2 & \triangleq & \frac{1}{n(m-1)}\sum_{k=1}^{K}\sum_{i\in T_k}\sum_{j\in T_k : j\neq i} e_ie_j \ , \\ s_3 & \triangleq & \frac{1}{n(n-m)}\sum_{k=1}^{K}\sum_{\ell\neq k}\sum_{i\in T_k}\sum_{j\in T_\ell} e_ie_j \ . \end{cases} \tag{8}$$

Theorem 1 *There exists no universally unbiased estimator of* $\text{Var}[\hat{\mu}]$.

The proof of the above Theorem shows that the bias of any quadratic estimator is a linear combination of μ^2, σ^2, ω and γ. However, the following new proposition shows that one may ignore the covariances ω and γ for large sample sizes. It motivates, in the asymptotic limit, some of the variance estimators presented in the forthcoming Section.

Proposition 1 *The covariances ω and γ go to zero as n goes to infinity provided the learning algorithm A is consistent and the loss function L is bounded.*

Proof

$\widehat{\mu_k}$ *is the mean loss on a test sample of size $m = n/K$, which is independant from the training sample of size $n(1-K)$. As both training and test sample sizes go to infinity as n goes to infinity, the consistency of A implies that, for every $\varepsilon > 0$*

$$\lim_{n\to\infty} P(|\widehat{\mu_k} - \mathrm{EPE}(n-m)| > \varepsilon) = 0 \ .$$

This convergence, together with the fact that L is bounded, implies $\lim_{n\to\infty} \mathrm{Var}[\widehat{\mu_k}] = 0$. By the Cauchy-Schwarz inequality, we have $\mathrm{Var}[\hat{\mu}] \leq \mathrm{Var}[\widehat{\mu_k}]$ and thus $\lim_{n\to\infty} \mathrm{Var}[\hat{\mu}] = 0$.

It then follows from the developments of $\mathrm{Var}[\widehat{\mu_k}]$ and $\mathrm{Var}[\hat{\mu}]$ ($\mathrm{Var}[\widehat{\mu_k}] = \frac{K}{n}\sigma^2 + (1-\frac{K}{n})\omega$ and $\mathrm{Var}[\hat{\mu}] = \frac{1}{n}\sigma^2 + (1-\frac{K}{n})\omega + \frac{K-1}{K}\gamma$) that we have $\lim_{n\to\infty}\omega = \lim_{n\to\infty}\gamma = 0$ *Q.E.D.*

3 Estimators built from cross-validation residuals

The statement that there exists no universal unbiased estimator of $\mathrm{Var}[\hat{\mu}]$ means that, for any estimator, there are distributions on cross-validation residuals e_i, such that the estimation is biased. However, by itself, this negative result does not prohibit the quest for sensible estimators. One way to derive estimates of $\mathrm{Var}[\hat{\mu}]$ consists in stating restrictive assumptions for which an unbiased estimator can be built.

We consider the series of simplifying assumptions listed on the left-hand side of Table 1, each one providing a unique unbiased estimate of variance of $\hat{\mu}$ expressed as a combination of s_1, s_2 and s_3. Unbiasedness applies to the whole set of distributions satisfying the assumption and is not guaranteed elsewhere. Hence, we also report the universal bias, computed thanks to Lemma 2, in the last column of Table 1.

Table 1: Non-universal unbiased variance estimates

Assumption	Estimate	Bias
$\mu = 0$	$\widehat{\theta_1} = \frac{1}{n}s_1 + \frac{m-1}{n}s_2 + \frac{n-m}{n}s_3$	μ^2
$\omega = 0$	$\widehat{\theta_2} = \frac{1}{n}s_1 - \frac{n+1-m}{n}s_2 + \frac{n-m}{n}s_3$	$-\omega$
$\gamma = 0$	$\widehat{\theta_3} = \frac{1}{n}s_1 + \frac{m-1}{n}s_2 - \frac{m}{n}s_3$	$-\gamma$
$\gamma = -\frac{m}{n-m}\omega$	$\widehat{\theta_4} = \frac{1}{n}s_1 - \frac{1}{n}s_2$	$-\frac{m}{n}\omega - \frac{n-m}{n}\gamma$

The first assumption, $\mu = 0$, corresponds to the null hypothesis when comparing the expected prediction error of two learning algorithms. All the other assumptions aim at simplifying the structure of the covariance matrix of $\mathbf{e}$, either regarding its entries ($\omega = 0$ and $\gamma = 0$) or regarding its eigendecomposition ($\gamma = 0$, $\gamma = -\frac{m}{n-m}\omega$) [3]. These simplifications push the e_i toward independence: some correlations are assumed to vanish, or the e_i are "sphered" by equalizing the eigenvalues of the covariance matrix).

Supposing that e_i are uncorrelated, i.e. $\omega = \gamma = 0$, is more restrictive than the assumptions listed in Table 1. In this situation, many variance estimates are unbiased. However, in the absence of correlation,

the standard unbiased estimator of the variance of $\hat{\mu}$ is

$$\begin{aligned}\widehat{\theta_5} &= \frac{1}{n(n-1)}\sum_i (e_i - \hat{\mu})^2 \\ &= \frac{1}{n}s_1 - \frac{m-1}{(n-1)n}s_2 - \frac{n-m}{(n-1)n}s_3 \ ,\end{aligned}$$

whose universal bias is $-\frac{m-1}{n-1}\omega - \frac{n-m}{n-1}\gamma$.

The variance estimates $\widehat{\theta_1}$, $\widehat{\theta_2}$, $\widehat{\theta_3}$, $\widehat{\theta_4}$ and $\widehat{\theta_5}$ are discussed in more details in the following subsections. Note that Proposition 1 implies that $\widehat{\theta_2}$–$\widehat{\theta_5}$ are asymptotically unbiased. A better understanding of some of these estimators is provided by rephrasing them as functions of $\hat{\mu}$, $\widehat{\mu_k}$ and e_i.

3.1 Assuming null mean

The first variance estimate, $\widehat{\theta_1}$, is an unbiased estimate of the variance of $\hat{\mu}$, when μ is assumed to be zero. This assumption is used to define the distribution of $\hat{\mu} = \Delta\text{CV}$ (4), under the null hypothesis that two algorithms A_1 and A_2 have an identical expected prediction error.

This estimate is expressed more simply as $\widehat{\theta_1} = \hat{\mu}^2$. Obviously, $\widehat{\theta_1}$ cannot be used to compute the usual pivotal t-statistic for testing the null hypothesis, since $t = \hat{\mu}/\sqrt{\widehat{\theta_1}} = \text{sign}(\hat{\mu})$. The distribution of $(\hat{\mu} - \mu)/\sqrt{\widehat{\theta_1}} = \text{sign}(\hat{\mu})(1 - \mu/\hat{\mu})$ explicitly depends on μ, hence this statistic is not pivotal and not appropriate for hypothesis testing. Thus, in spite of the relevance of the assumption $\mu = 0$ in algorithm comparison, $\widehat{\theta_1}$ is useless.

3.2 Assuming no correlation within test blocks

The estimator $\widehat{\theta_2}$ is based on the assumption that there is no correlation within test blocks. The following rewriting,

$$\widehat{\theta_2} = \frac{1}{n(m-1)}\sum_{k=1}^{K}\sum_{i\in T_k}\left\{(e_i - \hat{\mu})^2 + \sum_{\ell\neq k}\sum_{j\in T_\ell}(e_i - \hat{\mu})(e_j - \hat{\mu})\right\} \ ,$$

shows that $\widehat{\theta_2}$ is proportional to a truncated sample covariance. It can be shown that this truncation does not preserve the positivity of the estimator. This major defect rules $\widehat{\theta_2}$ out.

3.3 Assuming no correlation between test blocks

The estimator $\widehat{\theta_3}$ assumes that the between block covariance γ is null: for $\ell \neq k$, $\text{Cov}(\widehat{\mu_k}, \widehat{\mu_\ell}) = 0$. One can rewrite $\widehat{\theta_3}$ in the compact form

$$\widehat{\theta_3} = \frac{1}{K}\frac{1}{K-1}\sum_{k=1}^{K}(\widehat{\mu_k} - \hat{\mu})^2 \ ,$$

where one recognizes the sample variance of $\widehat{\mu_k}$, divided by K to account for that we are interested in estimating the variance of the mean $\hat{\mu} = \frac{1}{K}\sum_k \widehat{\mu_k}$ and not the variance of $\widehat{\mu_k}$ themselves.

Since the training examples $\mathbf{z}_i$ are independent, $\widehat{\mu_k}$ are uncorrelated when $A(D_1), \ldots, A(D_K)$ are identical. The present assumption should thus provide a good approximation of variance when the algorithm is stable. When the answers of the algorithm vary a lot, e.g. when the training samples D are small compared to the capacity of A, or when they are expected to contain some outliers, γ may reach high values, and $\widehat{\theta_3}$ will be consequently biased.

3.4 Assuming canceling correlations

The estimate $\widehat{\theta_4}$ assumes a linear dependence between ω and γ which implies that the within-block and between-block covariances approximately cancel out. It results that $\mathbf{e}$ can be decomposed in $n - K + 1$ uncorrelated compounds that allow to estimate θ. One can show that

$$\widehat{\theta_4} = \frac{1}{n}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{m-1}\sum_{i\in T_k}(e_i - \widehat{\mu_k})^2 \ ,$$

where the inner averages are the within-block sample variances. These variances, known to be identical for all blocks, are averaged over the K test blocks, and the factor n^{-1} transforms the variance of e_i into the variance of $\hat{\mu}$.

This estimator considers that the problem of estimating the variance of $\hat{\mu}$ and $\widehat{\mu_k}$ are orthogonal. As a result, there is no term in s_3, i.e. no $e_i e_j$ terms with $i \in T_k,\ j \in T_\ell,\ \ell \neq k$. The assumed linear dependence between ω and γ is unlikely, but it can be a good approximation to reality when ω and γ are small. When the correlations are larger, there is little chance that $\widehat{\theta_4}$ will perform well.

3.5 Assuming no correlation

As already mentioned, many unbiased estimates of variance exist when assuming a total absence of correlation $\omega = \gamma = 0$. The last estimator is motivated by the usual sample variance for uncorrelated examples

$$\widehat{\theta_5} = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \hat{\mu})^2 \ ,$$

where the factor the factor n^{-1} transforms the variance of e_i into the variance of $\hat{\mu}$.

Again, the assumption of absence of correlation between errors is only realistic when the algorithm is stable. The range of possible bias for $\widehat{\theta_5}$ is similar to the one of $\widehat{\theta_3}$ and $\widehat{\theta_4}$.

4 A t-statistic for K-fold cross-validation

4.1 Motivation

The variability of $\hat{\mu}$ is due to the variance and the covariance of cross-validation residuals. The previous section illustrates that obtaining unbiased estimates of θ requires assumptions on the covariance matrix which are unsupprted for small sample sizes.

The major concern is that γ may be positive; the consequences of this positivity have been motivating the present thread of research, starting from [2]. When γ is negative, $\widehat{\theta_3}$ over-estimates variance; this is much less problematic than under-estimating variance since it leads to a provable conservative test. Theory

does not preclude negative γ values, but, in a modified cross-validation setup, where the assignment of examples to the training or the test set is performed independently on each fold, Nadeau and Bengio prooved that γ is non-negative [4]. In our simulations, negative γ values are atypical, but they were observed in regression problems with gross outliers.

To go beyond speculations on γ, we should estimate its value, but the estimation process from a single cross-validation trial is clearly hopeless. We observed the result of cross-validation on a single training set, and we are willing to estimate a covariance reflecting the variability with respect to changes in the training set. Circumventing this problem requires additional information, such as the one provided by several cross-validations operating on independent training sets.

Dietterich showed experimentally that Student's t-test (uncorrected for correlations) is not reliable in the context of K-fold cross-validation [2]. He then proposed the 5×2 cross-validation procedure, leading to the 5×2 cross-validation t-test, which was later modified by Alpaydin's 5×2 cross-validation F-test [5]. Besides assumptions of the kind of the ones leading to the variance estimates given in Table 1, these tests are based on independance assumptions which are in conflict with the definitions of the variables (in the form of "X is independant of $X + Y$"). Nadeau and Bengio proposed another procedure providing an unbiased estimate of variance, but achieve this goal at the expense of a 10-fold increase in computation, without variance reduction [4].

Here, we take a different stance. Despite the negative results of [3], we aim at providing some guidance for concluding at the significance of the observed differences in performance between two algorithms. As in the 5×2 cross-validation procedure, only 10 trainings are necessary, but unlike the latter, the standard 10-fold cross-validation is performed. Also, the derived statistic is, up to an unknown factor, proven to really follow a t-distribution.

4.2 Derivation

The t-test is designed to answer questions about the mean of a random sample of independent observations from a normal distribution when the variance is unknown. In the context of algorithm comparison via K-fold cross-validation, it is routinely though improperly used to test whether two algorithms perform equally, i.e. if the expected value of ΔCV (4) is zero. The analysis presented in [3] can be used to analyze how the between-block covariances affect the t-statistic.

Proposition 2 *Under the assumption that $\widehat{\mu_k}$ are normally distributed* [2]*,*

$$t_{K-1} = \sqrt{K(K-1)(1-\rho)} \frac{\hat{\mu} - \mu}{\sqrt{\sum_{k=1}^{K} \left(\widehat{\mu_k} - \widehat{\mu}\right)^2}} \ , \tag{9}$$

where $\rho = \frac{\gamma}{\theta}$ is the between-block correlation, follows a Student t-distribution with $K-1$ degrees of freedom.

Proof *We first recall that, $\forall\ k\ , \forall\ \ell \neq k$*

$$\left\{ \begin{array}{rcl} E[\widehat{\mu_k}] & = & \mu \\ \mathrm{Var}[\widehat{\mu_k}] & = & \frac{1}{m}\sigma^2 + \frac{m-1}{m}\omega \\ \mathrm{Cov}(\widehat{\mu_k}, \widehat{\mu_\ell}) & = & \gamma \ . \end{array} \right.$$

Let $\mathbf{C} = \left(\frac{1}{m}\sigma^2 + \frac{m-1}{m}\omega - \gamma\right)\mathbf{I} + \gamma\mathbf{1}\mathbf{1}^T$ denote the covariance matrix of $\widehat{\boldsymbol{\mu}} = (\widehat{\mu_1}, \ldots, \widehat{\mu_K})^T$. The eigen-decomposition of $\mathbf{C}$ yields $\mathbf{C} = \mathbf{P}^T\mathbf{\Lambda}\mathbf{P}$, where $\mathbf{P} = (\mathbf{\Gamma}, \sqrt{1/K}\mathbf{1})$ is an orthonormal matrix and $\mathbf{\Lambda} =$

[2] This assumption is rather weak since $\widehat{\mu_k}$ are averages of identically distributed variables. It is thus a good approximation of the true distribution even for moderately large n/K.

$\mathrm{diag}(\nu_1, \ldots, \nu_1, \nu_2)$, *with* $\nu_1 = \frac{1}{m}\sigma^2 + \frac{m-1}{m}\omega - \gamma$ *appearing* $(K-1)$ *times and* $\nu_2 = \frac{1}{m}\sigma^2 + \frac{m-1}{m}\omega + (K-1)\gamma$. *We define* $\tilde{\boldsymbol{\mu}} = \boldsymbol{\Lambda}^{-1/2}\mathbf{P}\widehat{\boldsymbol{\mu}}$, *which has the following mean and variance*

$$E[\tilde{\boldsymbol{\mu}}] = \left(\mathbf{0}, \sqrt{\frac{K}{\nu_2}}\mu\right)^T \ , \ \mathrm{Var}[\tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^T] = \mathbf{I} \ .$$

Under the assumption that $\widehat{\mu_k}$ *are normally distributed,* $\sum_{k=1}^{K-1} \tilde{\mu}_k^2$ *is distributed according to a* χ^2 *distribution with* $K-1$ *degrees of freedom. This sum can be computed as follows:*

$$\begin{aligned}
\sum_{k=1}^{K-1} \tilde{\mu}_k^2 &= \tilde{\boldsymbol{\mu}}^T\tilde{\boldsymbol{\mu}} - \tilde{\mu}_K^2 \\
&= \widehat{\boldsymbol{\mu}}^T\mathbf{C}^{-1}\widehat{\boldsymbol{\mu}} - \frac{1}{K\nu_2}\widehat{\boldsymbol{\mu}}^T\mathbf{1}\mathbf{1}^T\widehat{\boldsymbol{\mu}} \\
&= \widehat{\boldsymbol{\mu}}^T\left(\frac{1}{\nu_1}\mathbf{I} + \left(\frac{\nu_1-\nu_2}{K\nu_1\nu_2} - \frac{1}{K\nu_2}\right)\mathbf{1}\mathbf{1}^T\right)\widehat{\boldsymbol{\mu}}
\end{aligned}$$

$$\begin{aligned}
\sum_{k=1}^{K-1} \tilde{\mu}_k^2 &= \frac{1}{\nu_1}\left(\widehat{\boldsymbol{\mu}}^T\widehat{\boldsymbol{\mu}} - \frac{1}{K}\widehat{\boldsymbol{\mu}}^T\mathbf{1}\mathbf{1}^T\widehat{\boldsymbol{\mu}}\right) \\
&= \frac{1}{\nu_1}\left(\sum_{k=1}^{K}\widehat{\mu_k}^2 - \frac{1}{K}\left(\sum_{k=1}^{K}\widehat{\mu_k}\right)^2\right) \\
&= \frac{1}{\nu_1}\sum_{k=1}^{K}(\widehat{\mu_k} - \widehat{\mu})^2 \ .
\end{aligned}$$

Furthermore, $\tilde{\mu}_K - \sqrt{\frac{K}{\nu_2}}\mu = \sqrt{\frac{K}{\nu_2}}(\hat{\mu} - \mu)$ *is normally distributed* $\mathcal{N}(0,1)$ *and independent of* $\tilde{\mu}_k$, $k \neq K$, *hence*

$$\sqrt{\frac{K(K-1)\nu_1}{\nu_2}}\frac{\hat{\mu} - \mu}{\sqrt{\sum_{k=1}^{K}(\widehat{\mu_k} - \widehat{\mu})^2}}$$

is t-distributed with $K-1$ *degrees of freedom. The proposition follows as* $\nu_1/\nu_2 = 1 - \rho$. *Q.E.D.*

Note that Proposition 2 does not provide a means to build a principled t-test, because the correlation ρ is unknown. We however believe that it is worth being stated because of its consequences. First, approaches based on calibrating t-tests should be targeted at finding a substitute for ρ instead of adjusting the number of degrees of freedom of the test statistic [6].

Second, constructing a conservative t-test requires to upper-bound $(1-\rho)^{\frac{1}{2}}$ by some positive constant. Overall, in our experiments, the maximum observed ρ value was 0.7. With this value, when comparing two algorithms with identical performances, assuming $\rho = 0$ will result in a type-I error of range 26% at an alleged 5% significance level. In other words, by using the usual t-statistic, we will conclude in 26% of cases that the algorithms differ while we would have like to limit the frequency of this type of erroneous judgment to 5%. Looking at the other side of the same problem, an alleged p-value of 0.2% in fact represents a true p-value of 5%.

Finally, one cannot compute t_{K-1} (9) since ρ is unknown, but for a given significance level α, we can compute the maximum value ρ_α such that the null hypothesis would be accepted. This figure is not exactly what we would ideally like (a p-value), but it is objective, does not rely on wrong nor unverifiable assumptions, and can be used to quantify the confidence in the rejection of the hypothesis test, since it is a monotonically decreasing function of the true p-value. More rigorously, if one believes that correlations above 0.7 are extremely unlikely in 10-fold cross-validation, then it is sensible to compute t_{K-1} with this ρ value to have a conservative test. Then, rejecting the null hypothesis amounts to state "Provided that $\rho \leq 0.7$, the difference in ΔCV is statistically significant at the 5% level."

5 Experimental results

We have conducted experiments for classification and regression problems, for artificial and real data, using trees and linear predictors. In all experiments, the variance estimates $\widehat{\theta_3}$, $\widehat{\theta_4}$ and $\widehat{\theta_5}$ behaved similarly. This is due to the fact that covariances ω and γ are alike. We report here the results for the t-test based on $\widehat{\theta_3}$, for which the Gaussian assumption is more grounded than for $\widehat{\theta_4}$ or $\widehat{\theta_5}$.

For lack of space, we only report results with trees for the Letter and Forest data. The Letter data set comprises 20 000 examples described by 16 numeric features. The Forest data set comprises 581 012 examples described by 54 numeric features. Both problems were transformed in approximately balanced binary classification problems to obtain sensible results for small sample sizes.

Accurate estimates of the true expected cross-validation error (and its variance) require many independent training samples. This was achieved by considering the whole data set to be the population, from which 10 000 independent training samples of size n were drawn by uniform sampling with replacement. Results obtained for different training set sizes are summarized in Table 2.

Table 2: Experimental results over 10 000 independent trainings, for the significance level $\alpha = 5\%$

Letter	n	20	40	80	160	400	800	2000
	ρ (%)	52.45	43.43	42.16	34.04	28.23	25.69	22.68
Type-I error, $\hat{\rho} = 0$ (%)		16.4	12.8	12.4	9.9	8.8	8.1	7.8
Type-I error, $\hat{\rho} = 0.7$ (%)		3.1	1.5	1.3	1.0	0.7	0.7	0.5
Forest	n	20	40	80	160	400	800	2000
	ρ (%)	53.22	49.85	51.94	47.15	45.44	42.61	40.79
Type-I error, $\hat{\rho} = 0$ (%)		16.3	14.1	15.9	14	12.9	12.3	11.6
Type-I error, $\hat{\rho} = 0.7$ (%)		3.1	2.3	2.4	1.7	1.5	1.3	1.1

¿From the definition of $\widehat{\theta_3}$, the correlation ρ is the downward relative bias of the estimator, that is $\rho = \frac{\theta - \mathbb{E}[\widehat{\theta_3}]}{\theta}$. The figures obtained for $\widehat{\theta_4}$ and $\widehat{\theta_5}$ are very similar, thus not represented. One sees that, for small sample sizes, the variance estimates can be less than half of the total variance.

This underestimation of variance causes the usual t-test, computed with $\hat{\rho} = 0$, to erroneously reject the null hypothesis in more than 16% of cases, while this type-I error should normally be limited to $\alpha = 5\%$. To build our t-test, we chose $\rho = 0.7$, which is the maximum value we experimentally observed (including results not shown here). The test is conservative (as it should be, since ρ is always below 0.53),

but is highly skeptical for large sample sizes. Obtaining a better upper-bound for ρ without additional computation is still an open issue.

6 Conclusions

K-fold cross-validation is known to be variable, and its variance cannot be unbiasedly estimated. Although many estimators of variance are asymptotically unbiased (such as $\widehat{\theta_3}$, $\widehat{\theta_4}$ and $\widehat{\theta_5}$), for small sample sizes, these estimates are almost systematically down-biased, leading to tests which reject unduly the null hypothesis. Based on the derivation of the form of a t-statistic parametrized by the between-block correlation, we propose a test which is consistently conservative. The t-statistic can also provide an interpretable confidence score. For calibrated tests, this role is played by the p-value, but here, the latter is meaningless. On the other hand, knowing the maximum correlation for which the test is accepted is likely to be more interpretable.

If extra computation is permitted, the unbiased variance estimate of Nadeau and Bengio [4] could be used to estimate the between-block correlation on training sets of size approximately $n/2$. In our experiments, this correlation decreases smoothly with sample size, so that a good estimate at $n/2$ would provide a tight upper-bound for sample size n.

References

[1] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1143, 1995.

[2] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1998.

[3] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

[4] C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003.

[5] E. Alpaydin. Combined 5×2 cv F test for comparing supervised classification learning algorithms. *Neural Computation*, 11(8):1885–1892, 1999.

[6] R. R. Bouckaert. Choosing between two learning algorithms based on calibrated tests. In *ICML 2003: Proc. of the Twentieth International Conference on Machine Learning*, 2003.