Journal of Machine Learning Research 5 (2004) 1089–1105 Submitted 05/03; Revised 9/03; Published 9/04

No Unbiased Estimator of the Variance of K-Fold Cross-Validation

Yoshua Bengio BENGIOY@IRO.UMONTREAL.CA
Dept. IRO, Université de Montréal
C.P. 6128, Montreal, Qc, H3C 3J7, Canada

Yves Grandvalet YVES.GRANDVALET@UTC.FR
Heudiasyc, UMR CNRS 6599
Université de Technologie de Compiègne, France

Editor: Dana Ron

Abstract

Most machine learning researchers perform quantitative experiments to estimate generalization error and compare the performance of different algorithms (in particular, their proposed algorithm). In order to be able to draw statistically convincing conclusions, it is important to estimate the uncertainty of such estimates. This paper studies the very commonly used K-fold cross-validation estimator of generalization performance. The main theorem shows that there exists no universal (valid under all distributions) unbiased estimator of the variance of K-fold cross-validation. The analysis that accompanies this result is based on the eigen-decomposition of the covariance matrix of errors, which has only three different eigenvalues corresponding to three degrees of freedom of the matrix and three components of the total variance. This analysis helps to better understand the nature of the problem and how it can make naive estimators (that don't take into account the error correlations due to the overlap between training and test sets) grossly underestimate variance. This is confirmed by numerical experiments in which the three components of the variance are compared when the difficulty of the learning problem and the number of folds are varied.

Keywords: cross-validation, variance estimators, k-fold cross-validation, statistical comparisons of algorithms

1. Introduction

In machine learning, the standard measure of accuracy for trained models is the prediction error (PE), i.e. the expected loss on future examples. Learning algorithms themselves are often compared according to their average performance, which is formally defined as the expected value of prediction error (EPE) over training sets.

When the data distribution is unknown, PE and EPE cannot be computed. If the amount of data is large enough, PE can be estimated by the mean error over a hold-out test set. The usual variance estimates for means of independent samples can then be computed to derive error bars on the estimated prediction error, and to assess the statistical significance of differences between models.

The hold-out technique does not account for the variance with respect to the training set, and may thus be considered inappropriate for the purpose of algorithm comparison (Dietterich, 1999). Moreover, it makes an inefficient use of data which forbids its application to small sample sizes. In this situation, one rather uses computer intensive resampling methods such as cross-validation or bootstrap to estimate PE or EPE.

We focus here on K-fold cross-validation. While it is known that cross-validation provides an unbiased estimate of EPE, it is also known that its variance may be very large (Breiman, 1996). This variance should be estimated to provide faithful confidence intervals on PE or EPE, and to test the significance of observed differences between algorithms. This paper provides theoretical arguments showing the difficulty of this estimation.

The difficulties of the variance estimation have already been addressed (Dietterich, 1999; Kohavi, 1995; Nadeau and Bengio, 2003). Some distribution-free bounds on the deviations of cross-validation are available, but they are specific to some locally defined decision rules, such as nearest neighbors (Devroye et al., 1996). This paper builds upon the work of Nadeau and Bengio (2003), which investigated in detail the theoretical and practical merits of several estimators of the variance of cross-validation. Our analysis departs from this work in the sampling procedure defining the cross-validation estimate. While Nadeau and Bengio (2003) consider K independent training and test splits, we focus on the standard K-fold cross-validation procedure, where there is no overlap between test sets: each example of the original data set is used once and only once as a test example.

This paper is organized as follows. Section 2 defines the measures of performance for algorithms, their estimation by K-fold cross-validation and similar procedures such as delete-*m* jackknife. Our theoretical findings are summarized in Sections 3–6. They are followed in Section 7 by experiments illustrating the effect of experimental conditions on the total variance and its decomposition in three components, and confirming the underestimation of variance obtained by the naive estimator commonly used by researchers.

2. General Framework

In machine learning, the performance measure differs according to the experimenter's viewpoint. In applications, we are interested in finding the best algorithm for solving the particular task at hand, specified by one particular training set and some information about the data generating process. In algorithm evaluation, we want to compare several learning algorithms for different learning tasks, and we care about the sensitivity of the learning algorithm to the choice of training examples.

2.1 Measures of Performance

Formally, we have a training set $D = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, with $\mathbf{z}_i \in \mathcal{Z}$, independently sampled from an unknown distribution P. We also have a learning algorithm A, which maps a data set of (almost) arbitrary size to a function $A : \mathcal{Z}^* \to \mathcal{F}$. Throughout this paper, we consider symmetric algorithms, i.e. A is insensitive to the ordering of examples in the training set D. The discrepancy between the prediction and the observation $\mathbf{z}$ is measured by a loss functional $L : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$. Typically, L is the quadratic loss in regression ($L(f,(x,y)) = (f(x)-y)^2$) and the misclassification $\{0,1\}$-loss in classification ($(L(f,(x,y)) = 1_{f(x)\neq y}$).

Let $f = A(D)$ be the function returned by algorithm A on the training set D. In application based evaluation, the goal of learning is usually stated as the minimization of the prediction error, i.e. the expected loss on future test examples

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

where the expectation is taken with respect to $\mathbf{z}$ sampled from P.[1]

In algorithm based evaluation, we are not really interested in performances on a specific training set; we would like comparisons on a more general basis. In this context, the lowest level of generality can be stated as “training sets of size n sampled from P”, and the performance of learning algorithm A can be measured by the expected performance of the functions returned in this situation

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

where the expectation is taken with respect to D sampled from P^n and $\mathbf{z}$ independently sampled from P.

Note that other types of performances measure can be proposed, based for example on parameters, or defined by the predictability in other frameworks, such as the prequential analysis (Dawid, 1997).

When the data distribution is unknown, PE and EPE cannot be computed. They have to be estimated, and it is often crucial to assess the uncertainty attached to this estimation:

- in application-oriented experiments, to give a confidence interval on PE;
- in algorithm-oriented experiments, to take into account the stability of a given algorithm. For comparisons between algorithms, it is essential to assess the statistical significance of observed differences in the estimate $\widehat{\text{EPE}}$.

Although this point is often overlooked, estimating the variance of the estimates $\widehat{\text{PE}}$ and $\widehat{\text{EPE}}$ requires caution.

2.2 Hold-Out Estimates of Performance

If the amount of data is large enough, PE can be estimated by the mean error over a hold-out test set, and the usual variance estimate for means of independent variables can then be computed. However, even in the ideal situation where several independent training and test sets would be available, this estimate should not be applied to compute the variance of $\widehat{\text{EPE}}$: even though training and test examples are independent, the test errors are correlated, since many test errors are computed for each training set, now considered as a random variable.

Figure 1 illustrates how crucial it is to take these correlations into account. The mean of two variance estimators is reported *vs.* the empirical variance of the hold-out estimate, in an ideal situation where 10 independent training and test sets are available. The variance of $\widehat{\text{EPE}}(n)$ (estimated on 100,000 independent experiments) is displayed for reference by the dotted line. The average of $\widehat{\theta}_1$, the variance estimator ignoring correlations, shows that this estimate is highly biased, even for large sample sizes, whereas the variance estimator $\widehat{\theta}_2$, taking into account correlations, is unbiased. The details of this experiment are given below.

Experiment 1 *Ideal hold-out estimate of* EPE.

We have $K = 10$ independent training sets $D_1, \ldots, D_K$ of n independent examples $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, where $\mathbf{x_i} = (x_{i1}, \ldots, x_{id})'$ is a d-dimensional centered, unit covariance Gaussian variable ($d = 30$),

1. Note that we are using the same notation for random variables and their realization. The intended meaning will be specified when not clear from the context.

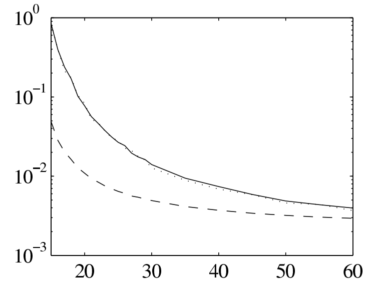

Figure 1: Estimates of the variance of $\widehat{\mathrm{EPE}}(n)$ vs. empirical variance of $\widehat{\mathrm{EPE}}(n)$ (shown by bold curve) on 100,000 experiments. The average of the variance estimators $\widehat{\theta}_1$ (ignoring correlations, dashed curve) and $\widehat{\theta}_2$ (taking into account correlations, dotted curve) are displayed for different training sample size n.

$y_i = \sqrt{3/d}\sum_{k=1}^{d} x_{ik} + \varepsilon_i$ *with* ε_i *being independent, centered, unit variance Gaussian variables.*[2] *We also have* K *independent test sets* $T_1, \ldots, T_K$ *of size n sampled from the same distribution.*

The learning algorithm consists in fitting a line by ordinary least squares, and the estimate of EPE *is the average quadratic loss on test examples* $\widehat{\mathrm{EPE}} = \bar{L} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{\mathbf{z}_i \in T_k} L_{ki}$, *where* $L_{ki} = L(A(D_k), \mathbf{z}_i)$.

The first estimate of variance of $\widehat{\mathrm{EPE}}$ *is* $\widehat{\theta}_1 = \frac{1}{Kn(Kn-1)}\sum_{k=1}^{K}\sum_i (L_{ki} - \bar{L})^2$, *which is unbiased provided there is no correlation between test errors. The second estimate is* $\widehat{\theta}_2 = \frac{1}{K(K-1)n^2}\sum_{k=1}^{K}\sum_{i,j}(L_{ki} - \bar{L})(L_{kj} - \bar{L})$, *which takes into account correlations between test errors.*

Looking at Figure 1 suggests that asymptotically the naive estimator of variance converges to the true variance. This can be shown formally by taking advantage of the results in this paper, as long as the learning algorithm converges as the amount of training data goes to infinity (i.e. as $n \to \infty$ the function $A(D)$ obtained does not depend on the particular training set D). In that limit, the correlations between test errors converge to 0. The rate of convergence will depend on the stability of the learning algorithm as well as on the nature of the data distribution (e.g., the presence of thick tails and outliers will slow down convergence).

The hold-out technique makes an inefficient use of data which forbids its application in most real-life applications with small samples. Then, K-fold cross-validation can provide estimates of PE or EPE.

2.3 K-Fold Cross-Validation Estimates of Performance

Cross-validation is a computer intensive technique, using all available examples as training and test examples. It mimics the use of training and test sets by repeatedly training the algorithm K times with a fraction $1/K$ of training examples left out for testing purposes. This kind of hold-out estimate of performance lacks computational efficiency due to the repeated training, but the latter are meant to lower the variance of the estimate (Stone, 1974).

2. The $\sqrt{3/d}$ factor provides an R^2 of approximately $3/4$.

In practice, the data set D is first chunked into K disjoint subsets (or *blocks*) of the same size[3] $m \triangleq n/K$. Let us 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 defined as the average of the errors on test block T_k obtained when the training set is deprived from T_k:

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

Does CV estimate PE or EPE? Such a question may seem pointless considering that $\mathrm{PE}(D)$ is an estimate of $\mathrm{EPE}(n)$, but it becomes relevant when considering the variance of CV: does it inform us of the uncertainty about PE or EPE?

On the one hand, only one training set, D, enters the definition of CV, which can be, up to an approximation, an unbiased estimate of $\mathrm{PE}(D)$ (Hastie and Tibshirani, 1990).[4] Some distribution-free bounds on the expected deviations of $|\mathrm{CV}(D) - PE(D)|$ are available for leave-one-out cross-validation applied to specific algorithms A such as nearest neighbors (Devroye et al., 1996). In a more general context, it has also been proved that, under suitable stability assumptions on the algorithm A, $\mathrm{CV}(D)$ estimates $\mathrm{PE}(D)$ at least as accurately as the training error (Kearns and Ron, 1996; Anthony and Holden, 1998). A more appealing result states that CV is a more accurate estimate of PE than hold-out testing (Blum et al., 1999). However, this statement does not apply to $\mathrm{PE}(D)$, but to the prediction error of a randomized algorithm picking solutions uniformly within $\{A(D_k)\}_{k=1}^{K}$.

On the other hand, CV is explicitly defined from the learning algorithm A, and not from the function $f = A(D)$. The inner average in the definition of CV (3) is an average test loss for $A(D_k)$ which thus estimates unbiasedly $\mathrm{PE}(D_k)$. The training sets $D_1, \ldots, D_K$ are clearly not independent, but they are sampled from P^{n-m}. Hence, the outer average of (3) estimates unbiasedly $\mathrm{EPE}(n-m)$.[5] Here, following Dietterich (1999) and Nadeau and Bengio (2003), we will adopt this latter point of view.

The variance estimate of $\widehat{\mathrm{EPE}}$ provided by the hold-out estimate has to account for test error dependencies due to the choice of training set, which cannot be estimated using a single training/test experiment. Here, the situation is more complex, since there are additional dependencies due to the overlapping training sets $D_1, \ldots, D_K$. Before describing this situation in detail and summarizing the results of our theoretical analysis in Sections 3–6, we detail some procedures similar to K-fold cross-validation, for which the forthcoming analysis will also hold.

2.4 Other Estimates of the K-Fold Cross-Validation Type

One of the main use of variance estimates of $\widehat{\mathrm{EPE}}$ is to compare learning algorithms. The analysis presented in this paper also applies to the version of cross-validation dedicated to this purpose: if we want to compare the performances of algorithms A_1 and A_2, cross-validation with matched pairs

3. To simplify the analysis below we assume that n is a multiple of K.
4. More precisely, following Hastie and Tibshirani (1990), when L is the quadratic loss, and writing $f = A(D)$, $f^{-k} = A(D_k)$, assuming that for $(\mathbf{x}_i, y_i) = \mathbf{z}_i \in T_k$, $\frac{1}{K}\sum_{k=1}^{K} f^{-k}(\mathbf{x}_i) \approx f(\mathbf{x}_i)$ (which is weaker than $f^{-k} \approx f$) yields $E[\mathrm{CV}] \approx E[\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2]$, where the expectation is taken with respect to $y_1, \ldots, y_n$.
5. Note that leave-one-out cross-validation is known to fail to estimate EPE for unsmooth statistics (e.g. Breiman, 1996; Efron and Tibshirani, 1993). This failure is due to the similarity of the training sets $D_1, \ldots, D_K$ which are far from being representative samples drawn from P^{n-m}.

should be the method of choice

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

Compared to the difference of two independent cross-validation estimates, ΔCV avoids the additional variability due to train/test splits.

In application oriented experiments, we would like to estimate PE(D), the expected error when training with the given D. We have seen in Section 2.3 that under stability assumptions, CV can be used to estimate PE. Alternatively, we may resort to the jackknife or the delete-m jackknife (see e.g. Efron and Tibshirani (1993)) to estimate the optimism (i.e. the bias of the mean error on training examples, when the latter is used to estimate PE(D)). Ideally, the estimate of optimism should be an average over all subsets of size $n-m$, but a less computationally intensive alternative is

$$(K-1)\left(\frac{1}{K(n-m)}\sum_{k=1}^{K}\sum_{\mathbf{z}_i\in D_k} L(A(D_k),\mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^{n} L(A(D),\mathbf{z}_i)\right). \tag{5}$$

The link with cross-validation is exhibited more clearly by the following expression of the (debiased) jackknife estimate of PE

$$\mathrm{JK} = \mathrm{CV} + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(L(A(D),\mathbf{z}_i) - L(A(D_k),\mathbf{z}_i)\right). \tag{6}$$

For additional information about jackknife estimates and clues on the derivation of (5) and (6), the reader is referred to Efron and Tibshirani (1993).

2.5 Generic Notations

This paper studies the variance of statistics such as CV, ΔCV or JK. In what follows, these statistics will be denoted by $\hat{\mu}$, a generic notation for means of observations e_i split in K groups.

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} e_i \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{i\in T_k} e_i,\end{aligned}$$

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

$$\forall i\in T_k,\ e_i = \begin{cases} L(A(D_k),\mathbf{z}_i) & \text{for } \hat{\mu}=\mathrm{CV}, \\ L(A_1(D_k),\mathbf{z}_i) - L(A_2(D_k),\mathbf{z}_i) & \text{for } \hat{\mu}=\Delta\mathrm{CV}, \\ KL(A(D),\mathbf{z}_i) - \sum_{\ell\neq k} L(A(D_\ell),\mathbf{z}_i) & \text{for } \hat{\mu}=\mathrm{JK}. \end{cases}$$

Note that $\hat{\mu}$ is the average of identically distributed (dependent) variables. Thus, it asymptotically converges to a normally distributed variable, which is completely characterized by its expectation $E[\hat{\mu}]$ and its variance $\mathrm{Var}[\hat{\mu}] = E[\hat{\mu}^2] - E[\hat{\mu}]^2$.

3. Structure of the Covariance Matrix

The variance of $\hat{\mu}$ is defined as follows

$$\theta = \frac{1}{n^2} \sum_{i,j} \text{Cov}(e_i, e_j),$$

where $\text{Cov}(e_i, e_j) = E[e_i e_j] - E[e_i]E[e_j]$ is the covariance between variables e_i and e_j.

By using symmetry arguments over permutations of the examples in D, we show that many distributions on e_i and pairwise joint distributions on (e_i, e_j) are identical. As a result, the covariance matrix Σ has a very particular block structure, with only three possible values for $\Sigma_{ij} = \text{Cov}(e_i, e_j)$, and the expression of θ is thus a linear combination of these three values.

Lemma 1 *Using the notation introduced in section 2.5,*

1. *all e_i are identically distributed:*
   there exists f such that, $\forall i,\ P(e_i = u) = f(u)$.
2. *all pairs (e_i, e_j) belonging to the same test block are jointly identically distributed:*
   there exists g such that, $\forall (i, j) \in T_k^2 : j \neq i,\ P(e_i = u, e_j = v) = g(u, v)$.
3. *all pairs (e_i, e_j) belonging to different test blocks are jointly identically distributed:*
   there exists h such that, $\forall i \in T_k,\ \forall j \in T_\ell : \ell \neq k,\ P(e_i = u, e_j = v) = h(u, v)$.

Proof
These results are derived immediately from the permutation-invariance of $P(D)$ and the symmetry of A.

- *invariance with respect to permutations within test blocks:*
  1. $\forall (i, i') \in T_k^2,\ P(e_i = u) = P(e_{i'} = u) = f_k(u)$;
     $\forall (i, i') \in T_k^2,\ \forall j \in T_\ell$:
     $P(e_i = u, e_j = v) = P(e_{i'} = u, e_j = v)$
     hence:
  2. $\forall (i, j) \in T_k^2 : j \neq i,\ P(e_i = u, e_j = v) = g_k(u, v)$.
  3. $\forall i \in T_k,\ \forall j \in T_\ell : \ell \neq k,\ P(e_i = u, e_j = v) = h_{k\ell}(u, v)$.
- *invariance with respect to permutations between test blocks.*
  1. $\forall (k, k'),\ f_k(u) = f_{k'}(u) = f(u)$;
  2. $\forall (k, k'),\ g_k(u, v) = g_{k'}(u, v) = g(u, v)$;
  3. $\forall (k, k'),\ \forall (\ell, \ell') : \ell \neq k, \ell \neq k', \ell' \neq k, \ell' \neq k',\ h_{k\ell}(u, v) = h_{k\ell'}(u, v) = h_{k'\ell'}(u, v) = h_{k'\ell}(u, v) = h(u, v)$.

Q.E.D.

Corollary 2 *The covariance matrix Σ of cross-validation errors $\mathbf{e} = (e_1, \ldots, e_n)'$ has the simple block structure depicted in Figure 2:*

1. *all diagonal elements are identical*
   $\forall i,\ \mathrm{Cov}(e_i, e_i) = \mathrm{Var}[e_i] = \sigma^2$;
2. *all the off-diagonal entries of the K $m \times m$ diagonal blocks are identical*
   $\forall (i, j) \in T_k^2 : j \neq i,\ \mathrm{Cov}(e_i, e_j) = \omega$;
3. *all the remaining entries are identical*
   $\forall i \in T_k,\ \forall j \in T_\ell : \ell \neq k,\ \mathrm{Cov}(e_i, e_j) = \gamma$.

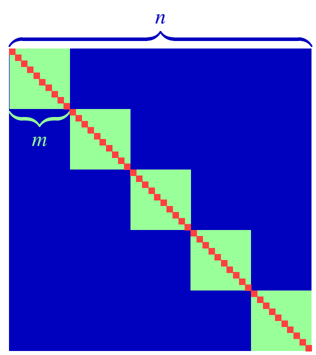

Figure 2: Structure of the covariance matrix.

Corollary 3 *The variance of the cross-validation estimator is a linear combination of three moments:*

$$\begin{aligned}\theta &= \frac{1}{n^2} \sum_{i,j} \mathrm{Cov}(e_i, e_j) \\ &= \frac{1}{n}\sigma^2 + \frac{m-1}{n}\omega + \frac{n-m}{n}\gamma \qquad (7)\end{aligned}$$

Hence, the problem of estimating θ does not involve estimating $n(n+1)/2$ covariances, but it cannot be reduced to that of estimating a single variance parameter. Three components intervene, which may be interpreted as follows when $\hat{\mu}$ is the K-fold cross-validation estimate of EPE:

1. The variance σ^2 is the average (taken over training sets) variance of errors for "true" test examples when algorithm A is fed with training sets of size $m(K-1)$.
2. The within-block covariance ω would also apply to "true" test examples; it arises from the dependence of test errors stemming from the common training set.
3. The between-blocks covariance γ is due to the dependence of training sets (which share $n(K-2)/K$ examples) and the fact that test block T_k appears in all the training sets D_ℓ for $\ell \neq k$.

The forthcoming section makes use of this structure to show that there is no universal unbiased estimator of θ.

4. No Unbiased Estimator of $\mathrm{Var}[\hat{\mu}]$ Exists

Consider a generic estimator $\hat{\theta}$ that depends on the sequence of cross-validation errors $\mathbf{e} = (e_1, e_2, \ldots, e_n)'$. Let us assume that $\hat{\theta}$ is an analytic function of the errors, so that we can write its Taylor expansion

$$\hat{\theta} = \alpha_0 + \sum_i \alpha_1(i) e_i + \sum_{i,j} \alpha_2(i,j) e_i e_j + \sum_{i,j,k} \alpha_3(i,j,k) e_i e_j e_k + \ldots. \tag{8}$$

We first show that for unbiased variance estimates (i.e. $E[\hat{\theta}] = \mathrm{Var}[\hat{\mu}]$), all the α_i coefficients must vanish except for the second order coefficients $\alpha_{2,i,j}$.

Lemma 4 *There is no universal unbiased estimator of* $\mathrm{Var}[\hat{\mu}]$ *that involves the* e_i *in a non-quadratic way.*

Proof

Take the expected value of $\hat{\theta}$ *expressed as in (8), and equate it with* $\mathrm{Var}[\hat{\mu}]$ *(7):*

$$\begin{cases} E[\hat{\theta}] = \alpha_0 + \sum_i \alpha_1(i) E[e_i] + \sum_{i,j} \alpha_2(i,j) E[e_i e_j] + \sum_{i,j,k} \alpha_3(i,j,k) E[e_i e_j e_k] + \ldots \\ \theta = \frac{1}{n}\sigma^2 + \frac{m-1}{n}\omega + \frac{n-m}{n}\gamma. \end{cases}$$

For having $E[\hat{\theta}] = \theta$ *for all possible values of the moments of* $\mathbf{e}$, *one must have* $\alpha_0 = 0$ *because* θ *has no such constant term, not depending on any of the moments of* $\mathbf{e}$. *Similarly,* $\alpha_1(\cdot)$ *must be zero because* θ *has no term in* $E[e_i] = \mu$. *Finally, the third and higher order coefficients* $\alpha_\ell(\ldots)$, $\ell > 2$ *must also be zero because* θ *has only quantities depending on the second order moments* σ^2, ω *and* γ.

Q.E.D.

Since estimators that include moments other than the second moments in their expectation are biased, we now focus on the class of estimators which are quadratic forms of the errors, i.e.

$$\hat{\theta} = \mathbf{e}'\mathbf{W}\mathbf{e} = \sum_{i,j} W_{ij} e_i e_j. \tag{9}$$

Lemma 5 *The expectation of quadratic estimators* $\hat{\theta}$ *defined as in (9) is a linear combination of only three terms*

$$E[\hat{\theta}] = a(\sigma^2 + \mu^2) + b(\omega + \mu^2) + c(\gamma + \mu^2), \tag{10}$$

where (a, b, c) *are defined as follows:*

$$\begin{cases} a & \triangleq & \sum_{i=1}^n W_{ii}, \\ b & \triangleq & \sum_{k=1}^K \sum_{i \in T_k} \sum_{j \in T_k : j \neq i} W_{ij}, \\ c & \triangleq & \sum_{k=1}^K \sum_{\ell \neq k} \sum_{i \in T_k} \sum_{j \in T_\ell} W_{ij}. \end{cases}$$

A "trivial" representer of estimators with this expected value is

$$\hat{\theta} = a s_1 + b s_2 + c s_3, \tag{11}$$

where (s_1, s_2, s_3) are the only quadratic statistics of $\mathbf{e}$ *that are invariants to the within blocks and between blocks permutations described in Lemma 1:*

$$\left\{\begin{array}{lcl} s_1 & \stackrel{\Delta}{=} & \frac{1}{n}\sum_{i=1}^{n} e_i^2, \\ s_2 & \stackrel{\Delta}{=} & \frac{1}{n(m-1)}\sum_{k=1}^{K}\sum_{i\in T_k}\sum_{j\in T_k: j\neq i} e_i e_j, \\ s_3 & \stackrel{\Delta}{=} & \frac{1}{n(n-m)}\sum_{k=1}^{K}\sum_{\ell\neq k}\sum_{i\in T_k}\sum_{j\in T_\ell} e_i e_j. \end{array}\right. \tag{12}$$

Proof
This result is obtained exploiting Corollary 2 and grouping the terms of $\hat{\theta}$ in Equation (9) that have the same expected values.

$$\begin{aligned} E[\hat{\theta}] &= \sum_{k=1}^{K}\sum_{i\in T_k}\left(W_{ii}E[e_i^2] + \sum_{j\in T_k: j\neq i} W_{ij}E[e_ie_j] + \sum_{\ell\neq k}\sum_{j\in T_\ell} W_{ij}E[e_ie_j]\right) \\ &= (\sigma^2+\mu^2)\sum_{i=1}^{n} W_{ii} + (\omega+\mu^2)\sum_{k=1}^{K}\sum_{i\in T_k}\sum_{j\in T_k: j\neq i} W_{ij} + \\ &\quad (\gamma+\mu^2)\sum_{k=1}^{K}\sum_{\ell\neq k}\sum_{i\in T_k}\sum_{j\in T_\ell} W_{ij} \\ &= a(\sigma^2+\mu^2) + b(\omega+\mu^2) + c(\gamma+\mu^2) \\ &= aE[s_1] + bE[s_2] + cE[s_3], \end{aligned}$$

which is recognized as the expectation of the estimator defined in Equation (11).

Q.E.D.

We now use Lemma 5 to prove that there is no *universally* unbiased estimator of $\mathrm{Var}[\hat{\mu}]$, i.e. there is no estimator $\hat{\theta}$ such that $E[\hat{\theta}] = \mathrm{Var}[\hat{\mu}]$ for all possible distributions of $\mathbf{e}$.

Theorem 6 *There exists no universally unbiased estimator of* $\mathrm{Var}[\hat{\mu}]$.
Proof
Because of Lemma 4 and 5, it is enough to prove the result for estimators that are quadratic forms expressed as in Equation (11). To obtain unbiasedness, the expected value of that estimator must be equated with $\mathrm{Var}[\hat{\mu}]$ *(7):*

$$a(\sigma^2+\mu^2) + b(\omega+\mu^2) + c(\gamma+\mu^2) = \frac{1}{n}\sigma^2 + \frac{m-1}{n}\omega + \frac{n-m}{n}\gamma. \tag{13}$$

For this equality to be satisfied for all distributions of cross-validation errors, it must be satisfied for all admissible values of μ, σ^2, ω, and γ. This imposes the following unsatisfiable constraints on (a,b,c):

$$\left\{\begin{array}{lcl} a & = & \frac{1}{n}, \\ b & = & \frac{m-1}{n}, \\ c & = & \frac{n-m}{n}, \\ a+b+c & = & 0. \end{array}\right. \tag{14}$$

Q.E.D.

5. Eigenanalysis of the Covariance Matrix

One way to gain insight on the origin of the negative statement of Theorem 6 is via the eigenanalysis of Σ, the covariance of $\mathbf{e}$. This decomposition can be performed analytically thanks to the very particular block structure displayed in Figure 2.

Lemma 7 *Let $\mathbf{v}_k$ be the binary vector indicating the membership of each example to test block k. The eigensystem of Σ is as follows:*

- $\lambda_1 = \sigma^2 - \omega$ *with multiplicity $n-K$ and eigenspace defined by the orthogonal of basis $\{\mathbf{v}_k\}_{k=1}^K$;*
- $\lambda_2 = \sigma^2 + (m-1)\omega - m\gamma$ *with multiplicity $K-1$ and eigenspace defined in the orthogonal of $\mathbf{1}$ by the basis $\{\mathbf{v}_k\}_{k=1}^K$;*
- $\lambda_3 = \sigma^2 + (m-1)\omega + (n-m)\gamma$ *with eigenvector $\mathbf{1}$.*

Proof
From Corollary 2, the covariance matrix $\Sigma = E[\mathbf{e}\mathbf{e}'] - E[\mathbf{e}]E[\mathbf{e}]'$ can be decomposed as

$$\Sigma = (\sigma^2 - \omega)\Sigma_1 + m(\omega - \gamma)\Sigma_2 + n\gamma\Sigma_3,$$

where $\Sigma_1 = \mathbf{I}$, $\Sigma_2 = \frac{1}{m}(\mathbf{v}_1 \dots \mathbf{v}_K)(\mathbf{v}_1 \dots \mathbf{v}_K)'$ and $\Sigma_3 = \frac{1}{n}\mathbf{1}\mathbf{1}'$.

Σ_1, Σ_2 and Σ_3 share the same eigenvectors, with eigenvalues being equal either to zero or one:

- *the eigenvector $\mathbf{1}$ has eigenvalue 1 for Σ_1, Σ_2 and Σ_3;*
- *the eigenspace defined in the orthogonal of $\mathbf{1}$ by the basis $\{\mathbf{v}_k\}_{k=1}^K$ defines $K-1$ eigenvectors with eigenvalues 1 for Σ_1 and Σ_2 and 0 for Σ_3;*
- *all remaining eigenvectors have eigenvalues 1 for Σ_1 and 0 for Σ_2 and Σ_3.*

Q.E.D.

Lemma 7 states that the vector $\mathbf{e}$ can be decomposed into three uncorrelated parts: $n-K$ projections to the subspace orthogonal to $\{\mathbf{v}_k\}_{k=1}^K$, $K-1$ projections to the subspace spanned by $\{\mathbf{v}_k\}_{k=1}^K$ in the orthogonal of $\mathbf{1}$, and one projection on $\mathbf{1}$. A single vector example with n independent elements can be seen as n independent examples. Similarly, these projections of $\mathbf{e}$ can be equivalently represented by respectively $n-K$, $K-1$ and one uncorrelated one-dimensional examples, corresponding to the coordinates of $\mathbf{e}$ in these subspaces.

In particular, for the projection on $\mathbf{1}$, with only a single one-dimensional point, the sample variance is null, resulting in the absence of an unbiased variance estimator of λ_3. The projection of $\mathbf{e}$ on the eigenvector $\frac{1}{n}\mathbf{1}$ is precisely $\hat{\mu}$. Hence there is no unbiased estimate of $\mathrm{Var}[\hat{\mu}] = \frac{\lambda_3}{n}$ when we have only one realization of the vector $\mathbf{e}$. For the same reason, even with simple parametric assumptions on $\mathbf{e}$ (such as $\mathbf{e}$ Gaussian), the maximum likelihood estimate of θ is not defined. Only λ_1 and λ_2 can be estimated unbiasedly. Note that this problem cannot be addressed by performing multiple K-fold splits of the data set. Such a procedure would not provide independent realizations of $\mathbf{e}$.

6. Possible Values for ω and γ

Theorem 6 states that no estimator is unbiased, and in its demonstration, it is shown that the bias of any quadratic estimator is a linear combination of μ^2, σ^2, ω and γ. Regarding estimation, it is thus interesting to see what constraints restrict the possible range of these quantities.

Lemma 8 *For $\hat{\mu} = \mathrm{CV}$ and $\hat{\mu} = \Delta\mathrm{CV}$, the following inequalities hold:*

$$\begin{cases} 0 & \leq \omega \leq & \sigma^2 \\ -\frac{1}{n-m}(\sigma^2 + (m-1)\omega) & \leq \gamma \leq & \frac{1}{m}(\sigma^2 + (m-1)\omega) \end{cases}$$
$$\Rightarrow \begin{cases} 0 & \leq \omega \leq & \sigma^2 \\ -\frac{m}{n-m}\sigma^2 & \leq \gamma \leq & \sigma^2. \end{cases}$$

The shape of the admissible (ω, γ) region corresponding to the first set of (tighter) inequalities is displayed in Figure 3.

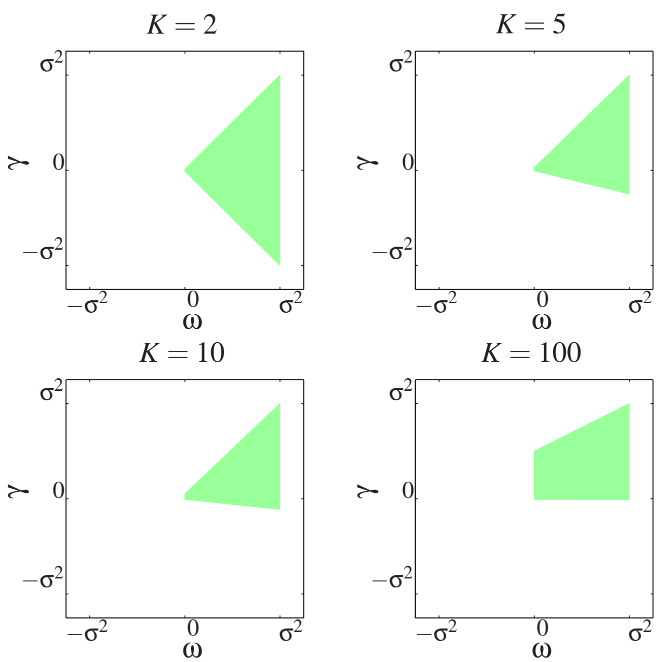

Figure 3: Possible values of (ω, γ) according to σ^2 for $n = 200$ and $K = \{2, 5, 10, 100\}$.

Proof
The constraints on ω result from the Cauchy-Schwartz inequality which provides $\mathrm{Cov}(u, v)^2 \leq \mathrm{Var}[u]\mathrm{Var}[v]$, hence

$$-\sigma^2 \leq \omega \leq \sigma^2.$$

Moreover, the following reasoning shows that, for $\hat{\mu} = \mathrm{CV}$ and $\hat{\mu} = \Delta\mathrm{CV}$, ω is non-negative: ω is the covariance of (differences in) test errors for training sets of size $n-m$ and test sets of size $\ell = m$. The variance of the average test error is given by the mean of covariances $\frac{1}{\ell}(\sigma^2 + (\ell-1)\omega)$. The variance σ^2 and covariance ω of test errors are not affected by ℓ, and the variance of the average *test error should be non-negative for any test set size ℓ. Hence ω is bound to be non-negative. When this type of reasoning cannot be used, as for $\hat{\mu} = \mathrm{JK}$, ω can only be proved to be greater than $-\sigma^2/(m-1)$.*

The constraints on γ simply rephrase that the eigenvalues λ_2 and λ_3 of the covariance matrix Σ should be non-negative. The simpler (and looser) form is obtained by using $\omega \leq \sigma^2$.

Q.E.D.

The admissible (ω, γ) region obtained in Lemma 8 is very large. Furthermore, there is no constraint linking μ and σ^2, the mean and variance of e_i. Hence we cannot propose a variance estimate with universally small bias.

7. Experiments

We already mentioned that the bias of any quadratic estimator is a linear combination of μ^2, σ^2, ω and γ. The admissible values provided in the preceding section suggest that ω and γ cannot be proved to be negligible compared to σ^2. This section illustrates that in practice, the contribution to the variance of $\hat{\mu}$ due to ω and γ (see Equation (7)) can be of same order than the one due σ^2. It therefore suggests that the estimators of θ should indeed take into account the correlations of e_i.

Experiment 2 *True variance of K-fold cross-validation.*

We repeat the experimental setup of Experiment 1, except that now, we are in the more realistic situation where only one sample of size n is available. Since cross-validation is known to be sensitive to the instability of algorithms, in addition to this standard setup, we also consider another one with outliers:

The input $\mathbf{x_i} = (x_{i1}, \ldots, x_{id})'$ is still 30-dimensional, but it is now a mixture of two centered Gaussian variables: let t_i be a binary variable, with $P(t_i = 1) = p = 0.95$; when $t_i = 1$, $x_i \sim \mathcal{N}(0, \mathbf{I})$; when $t_i = 0$, $x_i \sim \mathcal{N}(0, 100\mathbf{I})$; $y_i = \sqrt{3/(d(p+100(1-p)))} \sum_{k=1}^{d} x_{ik} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, 1/(p+100(1-p)))$ when $t_i = 1$ and $\varepsilon_i \sim \mathcal{N}(0, 100/(p+100(1-p)))$ when $t_i = 0$.

We now look at the variance of K-fold cross-validation ($K = 10$), and decompose in the three orthogonal components σ^2, ω and γ. The results are shown in Figure 4.

When there are no outliers, the contribution of γ is very important for small sample sizes. For large sample sizes, the overall variance is considerably reduced and is mainly caused by σ^2. In these situations, the learning algorithm returns very similar answers for all training sets. When there are outliers, ω has little effect, but the contribution of γ is of same order as the one of σ^2, even when the ratio of examples to free parameters is large (here up to 20). Thus, in difficult situations, where $A(D)$ varies according to the realization of D, neglecting the effect of ω and γ can be expected to introduce a bias of the order of the true variance.

It is also interesting to see how these quantities are affected by the number of folds K. The decomposition of θ in σ^2, ω and γ (7) does not imply that K should be set either to n or to 2 (according to the sign of $\omega - \gamma$) in order to minimize the variance of $\hat{\mu}$. Modifying K affects σ^2, ω and γ through the size and overlaps of the training sets $D_1, \ldots, D_K$, as illustrated in Figure 5. For a fixed sample size, the variance of $\hat{\mu}$ and the contribution of σ^2, ω and γ effects varies smoothly with K.[6] The experiments with and without outliers illustrate that there is no general trend either in

6. Of course, the mean of $\hat{\mu}$ is also affected in the process.

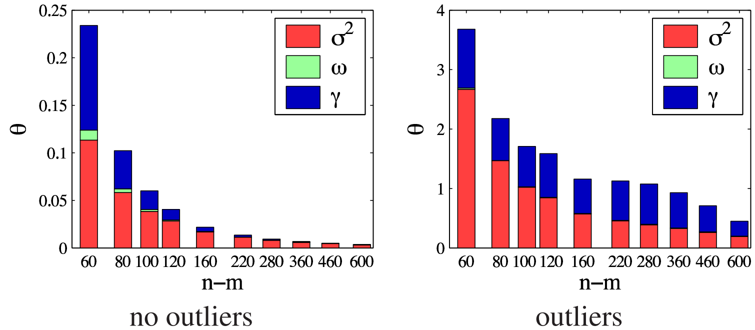

Figure 4: Bar plots of the contributions to total variance Var[CV] due to σ^2, ω and γ *vs.* the number of training examples $n-m$ for Experiment 2.

variance or decomposition of the variance in its σ^2, ω and γ components. The minimum variance can be reached for $K = n$ or for an intermediate value of K.

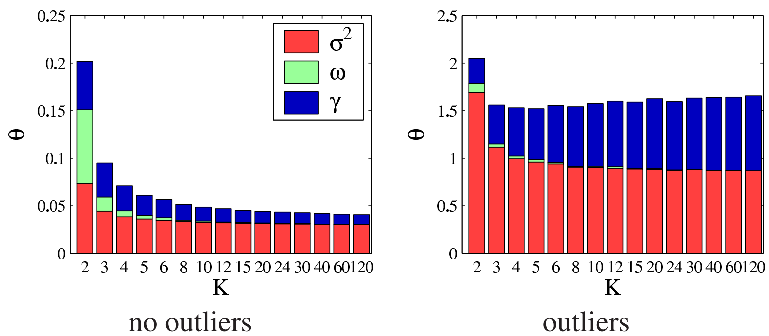

Figure 5: Bar plots of contributions of σ^2, ω and γ to θ *vs.* K for $n = 120$ for Experiment 2.

We also report an experiment illustrating that the previous observations also apply to classification on real data. The variance of K-fold cross-validation ($K = 10$), decomposed in the three orthogonal components σ^2, ω and γ is displayed in Figure 6.

Experiment 3 *Classification with trees on the Letter data set.*

The Letter data set comprises 20,000 examples described by 16 numeric features. The original setup considers 26 categories reprensenting the letters of the roman alphabet. Here, we used a simplified setup with 2 classes (A to M) vs. *(N to Z) in order to obtain sensible results for small sample sizes.*

Accurate estimates of σ^2, ω and γ require many independent training samples. This was achieved by considering the set of 20,000 examples to be the population, from which independent training samples were drawn by uniform sampling with replacement.

Here again, the variance of CV is mainly due to σ^2 and γ. According to the number of training examples, σ^2 is only responsible for 50 to 70 % of the total variance, so that a variance estimate based solely on σ^2 has a negative bias of the order of magnitude of the variance itself.

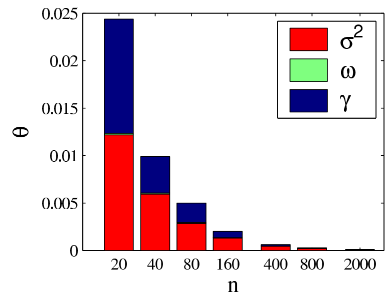

Figure 6: Bar plots of the contributions to total variance Var[CV] due to σ^2, ω and γ *vs.* the number of training examples n for Experiment 3.

8. Special Cases

This section addresses how our main result can be transposed to hold-out estimates of generalization error. We also detail how it applies to two specific instances of the general K-fold cross-validation scheme: two-fold and leave-one-out cross validation.

8.1 Hold-Out Estimate of EPE

When having *K independent training and test sets*, the structure of hold-out errors resemble the one of cross-validation errors, except that we know (from the independence of training and test sets) that $\gamma = 0$. This knowledge allows to build the unbiased variance estimate $\widehat{\theta_2}$ described in 2.2. This can be seen directly in the proof of Theorem 6: knowing that $\gamma = 0$ removes the third equation in the linear system (14). In practice, one is often restricted to $K = 1$ (ordinary hold-out test), which allows to estimate the variance due to the finite test set but not due to the particular choice of training set.

8.2 Two-Fold Cross-Validation

Two-fold cross-validation has been advocated to perform hypothesis testing (Dietterich, 1999; Alpaydin, 1999). It is a special case of K-fold cross-validation since the training blocks are mutually independent since they do not overlap. However, this independence does not modify the structure of **e** in the sense that γ is not null. The between-block correlation stems from the fact that the training block D_1 is the test block T_2 and vice-versa.

8.3 Leave-One-Out Cross-Validation

Leave-one-out cross-validation is a particular case of K-fold cross-validation, where $K = n$. The structure of the covariance matrix is simplified, without diagonal blocks: $\Sigma = (\sigma^2 - \gamma)\Sigma_1 + n\gamma\Sigma_3$. The estimation difficulties however remain: even in this particular case, there is no unbiased estimate of variance. From the definition of b (Lemma 5), we have $b = 0$, and with $m = 1$ the linear system (14) reads

$$\begin{cases} a & = & \frac{1}{n}, \\ c & = & \frac{n-1}{n}, \\ a+c & = & 0, \end{cases}$$

which still admits no solution.

9. Conclusions

It is known that K-fold cross-validation may suffer from high variability, which can be responsible for bad choices in model selection and erratic behavior in the estimated expected prediction error.

In this paper, we show that estimating the variance of K-fold cross-validation is difficult. Estimating a variance can be done from independent realizations or from dependent realizations whose correlation is known. K-fold cross-validation produces dependent test errors. Our analysis shows that although the correlations are structured in a very simple manner, their values cannot be estimated unbiasedly. Consequently, there is no unbiased estimator of the variance of K-fold cross-validation.

Our experimental section shows that in very simple cases, the bias incurred by ignoring the dependencies between test errors will be of the order of the variance itself. These experiments illustrate thus that the assessment of the significance of observed differences in cross-validation scores should be treated with much caution. The problem being unveiled, the next step of this study consists in building and comparing variance estimators dedicated to the very specific structure of the test error dependencies.

References

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

M. Anthony and S. B. Holden. Cross-validation for binary classification by real-valued functions: Theoretical analysis. In *Proceedings of the International Conference on Computational Learning Theory*, pages 218–229, 1998.

A. Blum, A. Kalai, and J. Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the International Conference on Computational Learning Theory*, pages 203–208, 1999.

L. Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996.

A. P. Dawid. Prequential analysis. In S. Kotz, C. B. Read, and D. L. Banks, editors, *Encyclopedia of Statistical Sciences, Update Volume 1*, pages 464–470. Wiley-Interscience, 1997.

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

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1999.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1993.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1990.

M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1996.

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

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

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