A Unified Bias-Variance Decomposition for Zero-One and Squared Loss

Pedro Domingos
Department of Computer Science and Engineering
University of Washington
Seattle, Washington 98195, U.S.A.
pedrod@cs.washington.edu
http://www.cs.washington.edu/homes/pedrod

Abstract

The bias-variance decomposition is a very useful and widely-used tool for understanding machine-learning algorithms. It was originally developed for squared loss. In recent years, several authors have proposed decompositions for zero-one loss, but each has significant shortcomings. In particular, all of these decompositions have only an intuitive relationship to the original squared-loss one. In this paper, we define bias and variance for an arbitrary loss function, and show that the resulting decomposition specializes to the standard one for the squared-loss case, and to a close relative of Kong and Dietterich's (1995) one for the zero-one case. The same decomposition also applies to variable misclassification costs. We show a number of interesting consequences of the unified definition. For example, Schapire et al.'s (1997) notion of "margin" can be expressed as a function of the zero-one bias and variance, making it possible to formally relate a classifier ensemble's generalization error to the base learner's bias and variance on training examples. Experiments with the unified definition lead to further insights.

Introduction

For the better part of the last two decades, machine-learning research has concentrated mainly on creating ever more flexible learners using ever more powerful representations. At the same time, very simple learners were often found to perform very well in experiments, sometimes better than more sophisticated ones (e.g., Holte (1993), Domingos & Pazzani (1997)). In recent years the reason for this has become clear: predictive error has two components, and while more powerful learners reduce one (bias) they increase the other (variance). The optimal point in this trade-off varies from application to application. In a parallel development, researchers have found that learning ensembles of models very often outperforms learning a single model (e.g., Bauer & Kohavi (1999)). That complex ensembles would outperform simple single models contradicted many existing intuitions about the relationship between simplicity and accuracy. This finding, apparently at odds with the one above about the value of simple learners, also becomes easier to understand in light of a bias-variance decomposition of error: while allowing a more intensive search for a single model is liable to increase variance, averaging multiple models will often (though not always) reduce it. As a result of these developments, the bias-variance decomposition of error has become a cornerstone of our understanding of inductive learning.

Although machine-learning research has been mainly concerned with classification problems, using zero-one loss as the main evaluation criterion, the bias-variance insight was borrowed from the field of regression, where squared-loss is the main criterion. As a result, several authors have proposed bias-variance decompositions related to zero-one loss (Kong & Dietterich, 1995; Breiman, 1996b; Kohavi & Wolpert, 1996; Tibshirani, 1996; Friedman, 1997). However, each of these decompositions has significant shortcomings. In particular, none has a clear relationship to the original decomposition for squared loss. One source of difficulty has been that the decomposition for squared-loss is purely additive (i.e., loss = bias + variance), but it has proved difficult to obtain the same result for zero-one loss using definitions of bias and variance that have all the intuitively necessary properties. Here we take the position that instead of forcing the bias-variance decomposition to be purely additive, and defining bias and variance so as to make this happen, it is preferable to start with a single consistent definition of bias and variance for all loss functions, and then investigate how loss varies as a function of bias and variance in each case. This should lead to more insight and to a clearer picture than a collection of unrelated decompositions. It should also make it easier to extend the bias-variance decomposition to further loss functions. Intuitively, since a bias-variance trade-off exists in any generalization problem, it should be possible and useful to apply a bias-variance analysis to any "reasonable" loss function. We believe the unified decomposition we propose here is a step towards this goal.

We begin by proposing unified definitions of bias and variance, and showing how squared-loss, zero-one loss and variable misclassification costs can be decomposed according to them. This is followed by the derivation of a number of properties of the new decomposition, in particular relating it to previous results. We then describe experiments with the new decomposition and discuss related work.

A Unified Decomposition

Given a training set $\{(x_1, t_1), \ldots, (x_n, t_n)\}$, a learner produces a model f. Given a test example x, this model produces a prediction $y = f(x)$. (For the sake of simplicity, the fact that y is a function of x will remain implicit throughout this paper.) Let t be the true value of the predicted variable for the test example x. A *loss function* $L(t, y)$ measures the cost of predicting y when the true value is t. Commonly used loss functions are squared loss ($L(t, y) = (t-y)^2$), absolute loss ($L(t, y) = |t - y|$), and zero-one loss ($L(t, y) = 0$ if $y = t$, $L(t, y) = 1$ otherwise). The goal of learning can be stated as producing a model with the smallest possible loss; i.e, a model that minimizes the average $L(t, y)$ over all examples, with each example weighted by its probability. In general, t will be a nondeterministic function of x (i.e., if x is sampled repeatedly, different values of t will be seen). The *optimal prediction* y_* for an example x is the prediction that minimizes $E_t[L(t, y_*)]$, where the subscript t denotes that the expectation is taken with respect to all possible values of t, weighted by their probabilities given x. The optimal model is the model for which $f(x) = y_*$ for every x. In general, this model will have non-zero loss. In the case of zero-one loss, the optimal model is called the *Bayes classifier*, and its loss is called the *Bayes rate*.

Since the same learner will in general produce different models for different training sets, $L(t, y)$ will be a function of the training set. This dependency can be removed by averaging over training sets. In particular, since the training set size is an important parameter of a learning problem, we will often want to average over all training sets of a given size. Let D be a set of training sets. Then the quantity of interest is the expected loss $E_{D,t}[L(t, y)]$, where the expectation is taken with respect to t and the training sets in D (i.e., with respect to t and the predictions $y = f(x)$ produced for example x by applying the learner to each training set in D). Bias-variance decompositions decompose the expected loss into three terms: bias, variance and noise. A standard such decomposition exists for squared loss, and a number of different ones have been proposed for zero-one loss.

In order to define bias and variance for an arbitrary loss function we first need to define the notion of main prediction.

Definition 1 *The* main prediction *for a loss function L and set of training sets D is $y_m^{L,D} = \text{argmin}_{y'} E_D[L(y, y')]$.*

When there is no danger of ambiguity, we will represent $y_m^{L,D}$ simply as y_m. The expectation is taken with respect to the training sets in D, i.e., with respect to the predictions y produced by learning on the training sets in D. Let Y be the multiset of these predictions. (A specific prediction y will appear more than once in Y if it is produced by more than one training set.) In words, the main prediction is the value y' whose average loss relative to all the predictions in Y is minimum (i.e., it is the prediction that "differs least" from all the predictions in Y according to L). The main prediction under squared loss is the mean of the predictions; under absolute loss it is the median; and under zero-one loss it is the mode (i.e., the most frequent prediction). For example, if there are k possible training sets of a given size, we learn a classifier on each, $0.6k$ of these classifiers predict class 1, and $0.4k$ predict 0, then the main prediction under zero-one loss is class 1. The main prediction is not necessarily a member of Y; for example, if $Y = \{1, 1, 2, 2\}$ the main prediction under squared loss is 1.5.

We can now define bias and variance as follows.

Definition 2 *The* bias *of a learner on an example x is $B(x) = L(y_*, y_m)$.*

In words, the bias is the loss incurred by the main prediction relative to the optimal prediction.

Definition 3 *The* variance *of a learner on an example x is $V(x) = E_D[L(y_m, y)]$.*

In words, the variance is the average loss incurred by predictions relative to the main prediction. Bias and variance may be averaged over all examples, in which case we will refer to them as *average bias* $E_x[B(x)]$ and *average variance* $E_x[V(x)]$.

It is also convenient to define noise as follows.

Definition 4 *The* noise *of an example x is $N(x) = E_t[L(t, y_*)]$.*

In other words, noise is the unavoidable component of the loss, that is incurred independently of the learning algorithm.

Definitions 2 and 3 have the intuitive properties associated with bias and variance measures. y_m is a measure of the "central tendency" of a learner. (What "central" means depends on the loss function.) Thus $B(x)$ measures the systematic loss incurred by a learner, and $V(x)$ measures the loss incurred by its fluctuations around the central tendency in response to different training sets. The bias is independent of the training set, and is zero for a learner that always makes the optimal prediction. The variance is independent of the true value of the predicted variable, and is zero for a learner that always makes the same prediction regardless of the training set. However, it is not necessarily the case that the expected loss $E_{D,t}[L(t, y)]$ for a given loss function L can be decomposed into bias and variance as defined above. Our approach will be to propose a decomposition and then show that it applies to each of several different loss functions. We will also exhibit some loss functions to which it does not apply. (However, even in such cases it may still be worthwhile to investigate how the expected loss can be expressed as a function of $B(x)$ and $V(x)$.)

Consider an example x for which the true prediction is t, and consider a learner that predicts y given a training set in D. Then, for certain loss functions L, the following decomposition of $E_{D,t}[L(t, y)]$ holds:

$$\begin{aligned} & E_{D,t}[L(t, y)] \\ & \quad = c_1 E_t[L(t, y_*)] + L(y_*, y_m) + c_2 E_D[L(y_m, y)] \\ & \quad = c_1 N(x) + B(x) + c_2 V(x) \qquad (1) \end{aligned}$$

c_1 and c_2 are multiplicative factors that will take on different values for different loss functions. We begin by showing that this decomposition reduces to the standard one for squared loss.

Theorem 1 *Equation 1 is valid for squared loss, with* $c_1 = c_2 = 1$.

Proof. Substituting $L(a,b) = (a-b)^2$, $y_* = E_t[t]$, $y_m = E_D[y]$ and $c_1 = c_2 = 1$, Equation 1 becomes:

$$\begin{aligned} E_{D,t}[(t-y)^2] &= E_t[(t-E_t[t])^2] + (E_t[t]-E_D[y])^2 \\ &\quad +E_D[(E_D[y]-y)^2] \end{aligned} \tag{2}$$

This is the standard decomposition for squared loss, as derived in (for example) Geman et al. (1992). $y_* = E_t[t]$ because $E_t[(t-y)^2] = E_t[(t-E_t[t])^2] + (E_t[t]-y)^2$ (also shown in Geman et al. (1992), etc.), and therefore $E_t[(t-y)^2]$ is minimized by making $y = E_t[t]$. □

Some authors (e.g., Kohavi and Wolpert, 1996) refer to the $(E_t[t]-E_D[y])^2$ term as "bias squared." Here we follow the same convention as Geman et al. (1992) and others, and simply refer to it as "bias." This makes more sense given our goal of a unified bias-variance decomposition, since the square in $(E_t[t]-E_D[y])^2$ is simply a consequence of the square in squared loss.

We now show that the same decomposition applies to zero-one loss in two-class problems, with c_1 reflecting the fact that on noisy examples the non-optimal prediction is the correct one, and c_2 reflecting that variance increases error on biased examples but decreases it on biased ones. Let $P_D(y = y_*)$ be the probability over training sets in D that the learner predicts the optimal class for x.

Theorem 2 *Equation 1 is valid for zero-one loss in two-class problems, with* $c_1 = 2P_D(y = y_*) - 1$ *and* $c_2 = 1$ *if* $y_m = y_*$, $c_2 = -1$ *otherwise.*

Proof. $L(a,b)$ represents zero-one loss throughout this proof. We begin by showing that

$$E_t[L(t,y)] = L(y_*,y) + c_0 E_t[L(t,y_*)] \tag{3}$$

with $c_0 = 1$ if $y = y_*$ and $c_0 = -1$ if $y \neq y_*$. If $y = y_*$ Equation 3 is trivially true with $c_0 = 1$. Assume now that $y \neq y_*$. Given that there are only two classes, if $y \neq y_*$ then $t \neq y_*$ implies that $t = y$ and vice-versa. Therefore $P_t(t = y) = P_t(t \neq y_*)$, and

$$\begin{aligned} E_t[L(t,y)] &= P_t(t \neq y) \;=\; 1 - P_t(t = y) \\ &= 1 - P_t(t \neq y_*) \;=\; 1 - E_t[L(t,y_*)] \\ &= L(y_*,y) - E_t[L(t,y_*)] \\ &= L(y_*,y) + c_0 E_t[L(t,y_*)] \end{aligned} \tag{4}$$

with $c_0 = -1$, proving Equation 3. We now show in a similar manner that

$$E_D[L(y_*,y)] = L(y_*,y_m) + c_2 E_D[L(y_m,y)] \tag{5}$$

with $c_2 = 1$ if $y_m = y_*$ and $c_2 = -1$ if $y_m \neq y_*$. If $y_m = y_*$ Equation 5 is trivially true with $c_2 = 1$. If $y_m \neq y_*$ then $y_m \neq y$ implies that $y_* = y$ and vice-versa, and

$$\begin{aligned} E_D[L(y_*,y)] &= P_D(y_* \neq y) \;=\; 1 - P_D(y_* = y) \\ &= 1 - P_D(y_m \neq y) \;=\; 1 - E_D[L(y_m,y)] \\ &= L(y_*,y_m) - E_D[L(y_m,y)] \\ &= L(y_*,y_m) + c_2 E_D[L(y_m,y)] \end{aligned} \tag{6}$$

with $c_2 = -1$, proving Equation 5. Using Equation 3,

$$\begin{aligned} E_{D,t}[L(t,y)] &= E_D[E_t[L(t,y)]] \\ &= E_D[L(y_*,y) + c_0 E_t[L(t,y_*)]] \end{aligned} \tag{7}$$

Since $L(t,y_*)$ does not depend on D,

$$E_{D,t}[L(t,y)] = E_D[c_0]E_t[L(t,y_*)] + E_D[L(y_*,y)] \tag{8}$$

and since

$$\begin{aligned} E_D[c_0] &= P_D(y = y_*) - P_D(y \neq y_*) \\ &= 2P_D(y = y_*) - 1 \;=\; c_1 \end{aligned} \tag{9}$$

we finally obtain Equation 1, using Equation 5. □

This decomposition for zero-one loss is closely related to that of Kong and Dietterich (1995). The main differences are that Kong and Dietterich ignored the noise component $N(x)$ and defined variance simply as the difference between loss and bias, apparently unaware that the absolute value of that difference is the average loss incurred relative to the most frequent prediction. A side-effect of this is that Kong and Dietterich incorporate c_2 into their definition of variance, which can therefore be negative. Kohavi and Wolpert (1996) and others have criticized this fact, since variance for squared loss must be positive. However, our decomposition shows that the subtractive effect of variance follows from a self-consistent definition of bias and variance for zero-one and squared loss, even if the variance itself remains positive. The fact that variance is additive in unbiased examples but subtractive in biased ones has significant consequences. If a learner is biased on an example, increasing variance decreases loss. This behavior is markedly different from that of squared loss, but is obtained with the same definitions of bias and variance, purely as a result of the different properties of zero-one loss. It helps explain how highly unstable learners like decision-tree and rule induction algorithms can produce excellent results in practice, even given very limited quantities of data. In effect, when zero-one loss is the evaluation criterion, there is a much higher tolerance for variance than if the bias-variance decomposition was purely additive, because the increase in average loss caused by variance on unbiased examples is partly offset (or more than offset) by its decrease on biased ones. The average loss over all examples is the sum of noise, the average bias and what might be termed the *net variance*, $E_x[(2B(x)-1)V(x)]$:

$$\begin{aligned} &E_{D,t,x}[L(t,y)] \\ &\quad = E_x[c_1 N(x)] + E_x[B(x)] + E_x[(2B(x)-1)V(x)] \end{aligned} \tag{10}$$

by averaging Equation 1 over all test examples x.

The c_1 factor (see Equation 9) also points to a key difference between zero-one and squared loss. In squared loss, increasing noise always increases error. In zero-one loss, for training sets and test examples where $y \neq y_*$, increasing noise decreases error, and a high noise level can therefore in principle be beneficial to performance.

The same decomposition applies in the more general case of multiclass problems, with correspondingly generalized coefficients c_1 and c_2.

Theorem 3 *Equation 1 is valid for zero-one loss in multiclass problems, with* $c_1 = P_D(y = y_*) - P_D(y \neq y_*)\, P_t(y = t \mid y_* \neq t)$ *and* $c_2 = 1$ *if* $y_m = y_*$, $c_2 = -P_D(y = y_* \mid y \neq y_m)$ *otherwise.*

Proof. The proof is similar to that of Theorem 2, with the key difference that now $y \neq y_*$ and $t \neq y_*$ no longer imply that $t = y$, and $y_m \neq y_*$ and $y_m \neq y$ no longer imply that $y = y_*$. Given that $y \neq y_*$ implies $P_t(y = t \mid y_* = t) = 0$,

$$\begin{aligned} P_t(y=t) &= P_t(y_* \neq t)\, P_t(y = t \mid y_* \neq t) \\ &\quad + P_t(y_* = t)\, P_t(y = t \mid y_* = t) \\ &= P_t(y_* \neq t)\, P_t(y = t \mid y_* \neq t) \qquad (11) \end{aligned}$$

and Equation 4 becomes

$$\begin{aligned} E_t[L(t,y)] &= P_t(y \neq t) \;=\; 1 - P_t(y = t) \\ &= 1 - P_t(y_* \neq t)\, P_t(y = t \mid y_* \neq t) \\ &= L(y_*, y) + c_0 E_t[L(t, y_*)] \qquad (12) \end{aligned}$$

with $c_0 = -P_t(y = t \mid y_* \neq t)$. When $y = y_*$ Equation 3 is trivially true with $c_0 = 1$, as before. A similar treatment applies to Equation 5, leading to $c_2 = -P_D(y = y_* \mid y \neq y_m)$ if $y_m \neq y_*$, etc. Given that

$$E_D[c_0] = P_D(y = y_*) - P_D(y \neq y_*) P_t(y = t | y_* \neq t) = c_1 \qquad (13)$$

we obtain Theorem 3. □

Theorem 3 means that in multiclass problems not all variance on biased examples contributes to reducing loss; of all training sets for which $y \neq y_m$, only some have $y = y_*$, and it is in these that loss is reduced. This leads to an interesting insight: when zero-one loss is the evaluation criterion, the tolerance for variance will decrease as the number of classes increases, other things being equal. Thus the ideal setting for the "bias-variance trade-off" parameter in a learner (e.g., the number of neighbors in k-nearest neighbor) may be more in the direction of high variance in problems with fewer classes.

In many classification problems, zero-one loss is an inappropriate evaluation measure because misclassification costs are asymmetric; for example, classifying a cancerous patient as healthy is likely to be more costly than the reverse. Consider the two class case with $\forall_y\ L(y, y) = 0$ (i.e., there is no cost for making the correct prediction), and with any nonzero real values for $L(y_1, y_2)$ when $y_1 \neq y_2$. The decomposition above also applies in this case, with the appropriate choice of c_1 and c_2.

Theorem 4 *In two-class problems, Equation 1 is valid for any real-valued loss function for which* $\forall_y\ L(y, y) = 0$ *and* $\forall_{y_1 \neq y_2}\ L(y_1, y_2) \neq 0$, *with* $c_1 = P_D(y = y_*) - \frac{L(y_*,y)}{L(y,y_*)} P_D(y \neq y_*)$ *and* $c_2 = 1$ *if* $y_m = y_*$, $c_2 = -\frac{L(y_*,y_m)}{L(y_m,y_*)}$ *otherwise.*

We omit the proof in the interests of space; see Domingos (2000). Theorem 4 essentially shows that the loss-reducing effect of variance on biased examples will be greater or smaller depending on how asymmetric the costs are, and on which direction they are greater in. Whether this decomposition applies in the multiclass case is an open problem. It does not apply if $L(y, y) \neq 0$; in this case the decomposition contains an additional term corresponding to the cost of the correct predictions.

Properties of the Unified Decomposition

One of the main concepts Breiman (1996a) used to explain why the bagging ensemble method reduces zero-one loss was that of an *order-correct* learner.

Definition 5 *(Breiman, 1996a) A learner is* order-correct *on an example x iff* $\forall_{y \neq y_*}\ P_D(y) < P_D(y_*)$.

Breiman showed that bagging transforms an order-correct learner into a nearly optimal one. An order-correct learner is an unbiased one according to Definition 2:

Theorem 5 *A learner is order-correct on an example x iff $B(x) = 0$ under zero-one loss.*

The proof is immediate from the definitions, considering that y_m for zero-one loss is the most frequent prediction.

Schapire et al. (1997) have proposed an explanation for why the boosting ensemble method works in terms of the notion of *margin*. For algorithms like bagging and boosting, that generate multiple hypotheses by applying the same learner to multiple training sets, their definition of margin can be stated as follows.

Definition 6 *(Schapire et al., 1997) In two-class problems, the* margin *of a learner on an example x is* $M(x) = P_D(y = t) - P_D(y \neq t)$.

A positive margin indicates a correct classification by the ensemble, and a negative one an error. Intuitively, a large margin corresponds to a high confidence in the prediction. D here is the set of training sets to which the learner is applied. For example, if 100 rounds of boosting are carried out, $|D| = 100$. Further, for algorithms like boosting where the different training sets (and corresponding predictions) have different weights that sum to 1, $P_D(.)$ is computed according to these weights. Definitions 1–4 apply unchanged in this situation. In effect, we have generalized the notions of bias and variance to apply to any training set selection scheme, not simply the traditional one of "all possible training sets of a given size, with equal weights."

Schapire et al. (1997) showed that it is possible to bound an ensemble's generalization error (i.e., its zero-one loss on test examples) in terms of the distribution of margins on training examples and the VC dimension of the base learner. In particular, the smaller the probability of a low margin, the lower the bound on generalization error. The following theorem shows that the margin is closely related to bias and variance as defined above.

Theorem 6 *The margin of a learner on an example x can be expressed in terms of its zero-one bias and variance as* $M(x) = \pm[2B(x) - 1][2V(x) - 1]$, *with positive sign if* $y_* = t$ *and negative sign otherwise.*

Proof. When $y_* = t$, $M(x) = P_D(y = y_*) - P_D(y \neq y_*) = 2P_D(y = y_*) - 1$. If $B(x) = 0$, $y_m = y_*$ and $M(x) =$

$2P_D(y = y_m) - 1 = 2[1 - V(x)] - 1 = -[2V(x) - 1]$. If $B(x) = 1$ then $M(x) = 2V(x) - 1$. Therefore $M(x) = [2B(x) - 1][2V(x) - 1]$. The demonstration for $y_* \neq t$ is similar, with $M(x) = P_D(y \neq y_*) - P_D(y = y_*)$. $\square$

Conversely, it is possible to express the bias and variance in terms of the margin: $B(x) = \frac{1}{2}[1 \pm \text{sign}(M(x))]$, $V(x) = \frac{1}{2}[1 \pm |M(x)|]$, with positive sign if $y_* \neq t$ and negative sign otherwise. The relationship between margins and bias/variance expressed in Theorem 6 implies that Schapire et al.'s theorems can be stated in terms of the bias and variance on training examples. Bias-variance decompositions relate a learner's loss on an example to its bias and variance on that example. However, to our knowledge this is the first time that *generalization* error is related to bias and variance on *training* examples.

Theorem 6 also sheds light on the polemic between Breiman (1996b, 1997) and Schapire et al. (1997) on how the success of ensemble methods like bagging and boosting is best explained. Breiman has argued for a bias-variance explanation, while Schapire et al. have argued for a margin-based explanation. Theorem 6 shows that these are two faces of the same coin, and helps to explain why the bias-variance explanation sometimes seems to fail when applied to boosting. Maximizing margins is a combination of reducing the number of biased examples, decreasing the variance on unbiased examples, and increasing it on biased ones (for examples where $y_* = t$; the reverse, otherwise). Without differentiating between these effects it is hard to understand how boosting affects bias and variance.

Unfortunately, there are many loss functions to which the decomposition in Equation 1 does not apply. For example, it does not apply to $L(t, y) = (t - y)^m$ with arbitrary m; in particular, it does not apply to absolute loss. (See Domingos (2000).) An important direction for future work is determining general properties of loss functions that are necessary and/or sufficient for Equation 1 to apply. Here we show that, as long as the loss function is a metric, it can be bounded from above and below by simple functions of the bias, variance and noise.

Theorem 7 *The following inequalities are valid for any metric loss function:*

$$\begin{aligned} E_{D,t}[L(t,y)] &\leq N(x) + B(x) + V(x) \\ E_{D,t}[L(t,y)] &\geq \max(\{N(x) - B(x) - V(x), \\ &\qquad B(x) - V(x) - N(x), \\ &\qquad V(x) - B(x) - N(x)\}) \end{aligned}$$

Proof. Recall that a function of two arguments $d(a_1, a_2)$ is a metric iff $\forall_{a,b}\ d(a, b) \geq d(a, a) = 0$ (minimality), $\forall_{a,b}\ d(a, b) = d(b, a)$ (symmetry), and $\forall_{a,b,c}\ d(a, b) + d(b, c) \geq d(a, c)$ (triangle inequality). Using the triangle inequality,

$$\begin{aligned} L(t,y) &\leq L(t, y_*) + L(y_*, y) \\ &\leq L(t, y_*) + L(y_*, y_m) + L(y_m, y) \end{aligned} \quad (14)$$

Taking the expected value of this equation with respect to D and t and simplifying produces the upper bound. Using the triangle inequality and symmetry,

$$\begin{aligned} L(y_*, y_m) &\leq L(y_*, t) + L(t, y) + L(y, y_m) \\ &\leq L(t, y_*) + L(t, y) + L(y_m, y) \end{aligned} \quad (15)$$

Rearranging terms, taking the expectation wrt D and t and simplifying leads to $E_{D,t}[L(t, y)] \geq B(x) - V(x) - N(x)$. The remaining components of the lower bound are obtained in a similar manner. $\square$

Experiments

We used the bias-variance decomposition for zero-one loss proposed here in numerous experiments on a large suite of benchmark datasets (Blake & Merz, 2000). While space limitations preclude a full description of the experiments (see Domingos (2000)), some of the main observations made are:

- Surprisingly, varying C4.5's pruning parameter (Quinlan, 1993) has only a minor effect on bias and variance.
- Varying the maximum number of levels in C4.5's decision trees produces more interesting results. Bias typically decreases very rapidly at first (one to three levels) and then stabilizes. Net variance increases steadily but slowly, largely because variance on biased examples significantly offsets variance on unbiased ones. The minimum loss is often found at one extreme (one level or unlimited levels).
- Boosting C4.5 tends to slightly reduce bias and strongly reduce variance. The bulk of bias reduction occurs in the first few rounds, after which bias stabilizes. The variance curves are more irregular.
- In k-nearest neighbor bias increase with k dominates variance reduction. However, increasing k has the "ideal" effect of reducing variance on unbiased examples while increasing it on biased ones.
- Compared to the results of Kohavi and Wolpert (1996) with their decomposition, variance is typically a smaller contributor to error. Again, this can be largely traced to the conflicting effects of variance on biased and unbiased examples.
- There are exceptions to every one of the previous observations. In general, it is not always the case that variance increases as bias decreases, or that both vary monotonically with the "bias-variance" parameter.

Related Work

The first bias-variance decomposition for zero-one loss was proposed by Kong and Dietterich (1995). Although they proposed it in a purely *ad hoc* manner and only applied it to one ensemble learner in one artificial, noise-free domain, our results show that it is in fact a well-founded and useful decomposition, even if incomplete. Breiman (1996b) proposed a decomposition for the average zero-one loss over all examples, leaving bias and variance for a specific example x undefined. As Tibshirani (1996) points out, Breiman's definitions of bias and variance have some undesirable properties, seeming artificially constructed to produce a purely additive decomposition. Tibshirani's (1996) definitions do not suffer from these problems; on the other hand, he makes no use of the variance, instead decomposing zero-one loss into bias and an unrelated quantity he calls the "aggregation effect." Kohavi and Wolpert (1996) defined bias and variance in terms of quadratic functions of $P_t(t)$ and $P_D(y)$. Although the resulting decomposition is purely additive, it suffers from the serious problem that it does not assign zero bias to the Bayes classifier. Also, although Kohavi and Wolpert emphasize the fact that their definition of zero-one bias is not restricted to taking on the values 0 or 1, it would seem that a binary-valued bias is the natural consequence of a binary-valued loss function. In practice, Kohavi and Wolpert's method produces biased estimates of bias and variance; although their estimators can be debiased, this obscures their meaning (for example, the corrected bias can be negative). Friedman (1997) studied the relationship between zero-one loss and the bias and variance of class probability estimates. He emphasized that the effect of bias and variance is strongly non-additive; increasing variance can reduce error. In this paper we obtain similar results directly in terms of the bias and variance of class predictions, and without Friedman's restrictive assumptions (only two classes, Gaussian probabilities).

Conclusion

In this paper we proposed general definitions of bias and variance applicable to any loss function, and derived the corresponding decompositions for squared loss, zero-one loss and variable misclassification costs. We also showed that margins can be expressed as a function of zero-one bias and variance, and that a simple relationship between loss, bias and variance exists for any metric loss function. Experiments on benchmark datasets illustrated the utility of our decomposition. Directions for future work include applying the decomposition to further loss functions, and conducting further experiments. C functions implementing the bias-variance decomposition proposed in this paper are available at http://www.cs.washington.edu/homes/pedrod/bvd.c.

Acknowledgments

This research was partly supported by a PRAXIS XXI grant. The author is grateful to all those who provided the datasets used in the experiments.

References

Bauer, E., and Kohavi, R. 1999. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning* 36:105–142.

Blake, C., and Merz, C. J. 2000. UCI repository of machine learning databases. Machine-readable data repository, Department of Information and Computer Science, University of California at Irvine, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L. 1996a. Bagging predictors. *Machine Learning* 24:123–140.

Breiman, L. 1996b. Bias, variance and arcing classifiers. Technical Report 460, Statistics Department, University of California at Berkeley, Berkeley, CA.

Breiman, L. 1997. Arcing the edge. Technical Report 486, Statistics Department, University of California at Berkeley, Berkeley, CA.

Domingos, P., and Pazzani, M. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29:103–130.

Domingos, P. 2000. A unified bias-variance decomposition. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle, WA.

Friedman, J. H. 1997. On bias, variance, 0/1 - loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* 1:55–77.

Geman, S.; Bienenstock, E.; and Doursat, R. 1992. Neural networks and the bias/variance dilemma. *Neural Computation* 4:1–58.

Holte, R. C. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning* 11:63–91.

Kohavi, R., and Wolpert, D. H. 1996. Bias plus variance decomposition for zero-one loss functions. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 275–283. Bari, Italy: Morgan Kaufmann.

Kong, E. B., and Dietterich, T. G. 1995. Error-correcting output coding corrects bias and variance. In *Proceedings of the Twelfth International Conference on Machine Learning*, 313–321. Tahoe City, CA: Morgan Kaufmann.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Schapire, R. E.; Freund, Y.; Bartlett, P.; and Lee, W. S. 1997. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 322–330. Nashville, TN: Morgan Kaufmann.

Tibshirani, R. 1996. Bias, variance and prediction error for classification rules. Technical report, Department of Preventive Medicine and Biostatistics and Department of Statistics, University of Toronto, Toronto, Canada.