Variance and bias for general loss functions

GARETH M. JAMES
Marshall School of Business, University of Southern California
gareth@usc.edu

February 18, 2003

Abstract

When using squared error loss, bias and variance and their decomposition of prediction error are well understood and widely used concepts. However, there is no universally accepted definition for other loss functions. Numerous attempts have been made to extend these concepts beyond squared error loss. Most approaches have focused solely on 0-1 loss functions and have produced significantly different definitions. These differences stem from disagreement as to the essential characteristics that variance and bias should display. This paper suggests an explicit list of rules that we feel any "reasonable" set of definitions should satisfy. Using this framework, bias and variance definitions are produced which generalize to any symmetric loss function. We illustrate these statistics on several loss functions with particular emphasis on 0-1 loss. We conclude with a discussion of the various definitions that have been proposed in the past as well as a method for estimating these quantities on real data sets.

Some key words: Bias, variance, prediction error, loss function.

1 Introduction

Over the last few years a great deal of research has been conducted on a family of classifiers known as *ensembles*. Examples of such classifiers include, the Error Correcting Output Coding method (ECOC) (Dietterich and Bakiri, 1995), Bagging (Breiman, 1996a) and AdaBoost (Freund and Schapire, 1996). Several theories have been proposed for the success of these classifiers. One involves the use of *margins* (Schapire *et al.*, 1998) while a second draws connections to additive logistic regression (Friedman and Rubin, 1967). A third theory postulates that the ensembles work because of the reduction in "variance", caused by the agglomeration of many classifiers into one classification rule. More recent work suggests that, while Bagging may indeed reduce variance, Boosting generates reductions in the error rate by decreasing both variance and bias. These results provide interesting insights and numerous attempts have been made to postulate why this effect occurs. Unfortunately, this work has been hampered by the fact that there are no universally accepted definitions for bias and variance when one moves away from squared error loss. There have been many suggestions to extend these concepts to classification problems. See for example Dietterich and Kong (1995), Kohavi and Wolpert (1996), Breiman (1996b), Tibshirani (1996), Friedman (1996), Wolpert (1997), Heskes (1998) and Domingos (2000). However, most of this work has concentrated on 0-1 loss functions and generally resulted in wildly differing definitions. The differences can be attributed to disagreement over the criteria that variance and bias should fulfill. Some definitions have clearly been constructed to produce an additive decomposition of the prediction error. While others attempt to reproduce the standard characteristics of variance and bias in a general setting. Unfortunately, for general loss functions it is not possible to provide a single bias/variance definition that will simultaneously achieve both criteria.

In this paper we introduce bias/variance definitions which have several key advantages over previously suggested approaches. One of the major difficulties in arriving at a satisfactory bias-variance decomposition is that there are several properties one would want that decomposition to have. Unfortunately, for general loss functions, there is no single definition that has all these properties and as a result different authors have dropped one or another, often without a particular rationale. In this paper we argue that these difficulties arise because there are two sets of quantities of interest, not just one. In the special case of squared error loss these quantities coincide but in general this will not be the case. In the standard setting there are two reasonable interpretations of variance. One is that it provides a measure of randomness of a quantity. We define this as the "variance". The other interpretation is that variance gives the increase in error rate caused by randomness which we call the "variance effect". Even though for squared error loss these quantities are numerically equal, in general they are clearly not conceptually the same thing. Further, either or both could be important depending on your objectives. Similarly there are two interpretations of bias. One is that it represents the systematic difference between a random variable and a particular value, e.g. the target, and the other is the degree to which that systematic difference contributes to error. We call these the "bias" and "systematic effect" respectively. Again these two quantities need not be the same and both may be important. In this paper we derive separate definitions for all four quantities. By providing two sets of definitions we ensure that all the important characteristics are captured and allow the user to specify which they are interested in.

Unlike previous work, the definitions in this paper are based on an explicitly stated set of criteria for variance and bias, ensuring that they possess what are generally considered the correct properties. For example the definitions attribute zero bias and variance to the Bayes classifier which several previous methods have not. The definitions we propose are not unique in fulfilling these criteria and reasonable people may disagree about what definitions should follow from them. However, to enable such a discussion one first needs to establish a set of reference criteria.

The third advantage of our approach is that the definitions are general to all symmetric loss functions. Most of the previously suggested definitions of bias and variance are specifically tailored to a small collection of loss functions such as 0-1 loss. It is obviously undesirable to have to derive new definitions for each new loss function.

The paper is structured as follows. In Section 2 we provide a set of criteria which we feel any definitions of bias and variance should fulfill. Based on these criteria we suggest the first set of definitions which are designed to replicate the natural properties of bias and variance. Section 3 develops the second set of definitions which provide an additive decomposition of the prediction error into bias and variance effects. For certain loss functions these two sets of definitions are identical but in general need not be. Section 4 gives examples where these general definitions are applied to some specific loss functions, in particular to 0-1 loss common in classification problems. In Section 5 we discuss some of the pros and cons of the previously suggested definitions. Finally, Section 6 provides examples of the definitions applied to simulated and real world data sets.

2 General definitions for bias and variance

In this section we briefly examine the dual roles of bias and variance for squared error loss. We then introduce three rules which we believe any reasonable definitions should satisfy and use these rules to extend the concepts of bias and variance to arbitrary symmetric loss functions.

2.1 Squared error loss

In the standard regression setting the variance of an estimator, $\hat{Y}$, is defined as $E_{\hat{Y}}(\hat{Y} - E\hat{Y})^2$ or equivalently

$$Var(\hat{Y}) = \min_{\mu} E_{\hat{Y}}(\hat{Y} - \mu)^2. \tag{1}$$

We define the systematic part of $\hat{Y}$ as

$$S\hat{Y} = \arg\min_{\mu} E_{\hat{Y}}(\hat{Y} - \mu)^2. \tag{2}$$

The notation $S\hat{Y}$ is used to emphasize that S is an operator acting on the distribution of $\hat{Y}$. In the standard squared error loss setting $S\hat{Y}$ will be equal to $E_{\hat{Y}}\hat{Y}$. Using this definition one can view $Var(\hat{Y})$ as a measure of the *expected distance*, in terms of squared error loss, of the random quantity ($\hat{Y}$) from its nearest non random number ($S\hat{Y}$).

The squared bias of $\hat{Y}$ in predicting a response Y is defined as

$$(E_{\hat{Y}}\hat{Y} - E_Y Y)^2 = (S\hat{Y} - SY)^2. \tag{3}$$

This means that squared bias can be viewed as a measure of the distance, in terms of squared error, between the systematic parts of $\hat{Y}$ and Y.

Finally we note, from the prediction error decomposition (Geman *et al.*, 1992),

$$\underbrace{E_{Y,\hat{Y}}(\hat{Y} - Y)^2}_{\text{prediction error}} = \underbrace{Var(Y)}_{\text{irreducible error}} + \underbrace{bias^2(\hat{Y}) + Var(\hat{Y})}_{\text{reducible error}}, \tag{4}$$

that the *expected loss* of using $\hat{Y}$ to predict Y is the sum of the variances of $\hat{Y}$ and Y plus the squared bias. The variance of Y is beyond our control and is thus known as the irreducible error. However, the bias and variance of $\hat{Y}$ are functions of our estimator and can therefore potentially be reduced.

This shows us that $Var(\hat{Y})$ serves two purposes.

1. Through (1) and (2) it provides a measure of the variability of $\hat{Y}$ about $S\hat{Y}$
2. and from (4) it indicates the effect of this variance on the prediction error.

Similarly $bias(\hat{Y})$ serves two purposes.

1. Through (3) it provides a measure of the distance between the systematic components of Y and $\hat{Y}$
2. and from (4) we see the effect of this bias on the prediction error.

This double role of both bias and variance is so automatic that one often fails to consider it. However, when these definitions are extended to arbitrary loss functions it will not, in general, be possible to define one statistic to serve both purposes. It is for this reason that we propose two sets of bias/variance definitions.

2.2 General loss functions

Squared error is a very convenient loss function to use. It possesses well known mathematical properties such as the bias/variance decomposition that make it very attractive to use. However, there are situations where squared error is clearly not the most appropriate loss function. This is especially true in classification problems where a loss function like 0-1 loss seems more realistic.

To extend the definitions of variance and bias in a systematic way we propose three simple rules which it seems reasonable that any definitions should follow.

❶ *When using squared error loss, any generalized definitions of bias and variance must reduce to the corresponding standard definitions.*

❷ *The "variance" must measure the variability of the estimator $\hat{Y}$. Hence it must not be a function of the distribution of the response variable, Y. Furthermore, it must be nonnegative and zero iff $\hat{Y}$ is constant for all training sets.*

❸ *The "bias" must measure the difference between the systematic parts of the response and predictor. In other words it must be a function of $\hat{Y}$ and Y only through $S\hat{Y}$ and SY. Furthermore, the bias should be zero if $S\hat{Y} = SY$.*

Based on an earlier version of this paper, Heskes (1998) develops his bias/variance decomposition using an almost identical set of requirements. The rules are also similar in spirit to those given in the desiderata of Wolpert (1997). The first of these rules is self evident. The second states that the variance of $\hat{Y}$ should only depend on the distribution of $\hat{Y}$ and not on Y. In other words it will depend on the training rather than the test data. This rule is desirable because it allows us to compare estimators across different response variables; a low variance estimator will be low variance for any test set of data. We also require that the variance be non-negative and zero iff $\hat{Y}$ does not deviate from $S\hat{Y}$. This is a natural requirement since variance is a measure of the average distance of $\hat{Y}$ from its systematic component. The third rule states that the bias of $\hat{Y}$ should depend only on the systematic components of $\hat{Y}$ and Y. Bias is viewed as a measure of the systematic difference between the response and predictor. Hence, variability of Y about SY and of $\hat{Y}$ about $S\hat{Y}$ will have no effect on the bias. A natural fourth requirement would be that the variance and bias provide an additive decomposition of the prediction error. Unfortunately, as has been mentioned earlier, for general loss functions, this requirement is inconsistent with the first three. We will return to a decomposition of the prediction error in the following section.

One might imagine that ❶ would be sufficient to provide a unique generalization. However, this is not the case because of the large number of definitions for variance and bias that are equivalent for squared error but not for other loss functions. For example, the following definitions of variance are all equivalent for squared error.

- $Var(\hat{Y}) = \min_\mu E_{\hat{Y}}(\hat{Y}-\mu)^2 = E_{\hat{Y}}(\hat{Y}-S\hat{Y})^2$
- $Var(\hat{Y}) = E_{\hat{Y}}(\hat{Y}-E_Y Y)^2 - (E_{\hat{Y}}\hat{Y}-E_Y Y)^2 = E_{\hat{Y}}(\hat{Y}-SY)^2 - (S\hat{Y}-SY)^2$
- $Var(\hat{Y}) = E_{Y,\hat{Y}}(Y-\hat{Y})^2 - E_Y(Y-E_{\hat{Y}}\hat{Y})^2 = E_{Y,\hat{Y}}(Y-\hat{Y})^2 - E_Y(Y-S\hat{Y})^2$

Note $E_{Y,\hat{Y}}$ indicates that the expectation is taken over the distribution of both the response and the predictor. Let L be an arbitrary symmetric loss function i.e. $L(a,b) = L(b,a)$. Then the above three definitions lead naturally to three possible generalized definitions,

$$Var(\hat{Y}) = \min_\mu E_{\hat{Y}} L(\hat{Y},\mu) = E_{\hat{Y}} L(\hat{Y},S\hat{Y}) \quad (5)$$

$$Var(\hat{Y}) = E_{\hat{Y}} L(SY,\hat{Y}) - L(SY,S\hat{Y}) \quad (6)$$

$$Var(\hat{Y}) = E_{Y,\hat{Y}} L(Y,\hat{Y}) - E_Y L(Y,S\hat{Y}), \quad (7)$$

where

$$S\hat{Y} = \arg\min_\mu L(\hat{Y},\mu) \quad (8)$$

$$SY = \arg\min_\mu L(Y,\mu). \quad (9)$$

For general loss functions these last three equations need not be consistent. This inconsistency accounts for some of the differences in the definitions that have been proposed. For example, Tibshirani bases his definition of variance on (5) while Dietterich and Kong base their's more closely on (7). We will see later that both (5) and (7) are useful for measuring different quantities. However, neither (6) nor (7) fulfill requirement ❷ which leaves (5) as a natural generalized definition of variance.

In a similar fashion there are several equivalent ways of defining the squared bias for squared error loss.

- $bias^2(\hat{Y}) = (E_Y Y - E_{\hat{Y}}\hat{Y})^2 = (SY - S\hat{Y})^2$
- $bias^2(\hat{Y}) = E_Y(Y - E_{\hat{Y}}\hat{Y})^2 - E_Y(Y - E_Y Y)^2 = E_Y(Y - S\hat{Y})^2 - E_Y(Y - SY)^2$
- $bias^2(\hat{Y}) = E_{\hat{Y}}(\hat{Y} - E_Y Y)^2 - Var(\hat{Y}) = E_{\hat{Y}}(\hat{Y} - SY)^2 - E(\hat{Y} - S\hat{Y})^2$
- $bias^2(\hat{Y}) = E_{Y,\hat{Y}}(Y - \hat{Y})^2 - Var(Y) - Var(\hat{Y}) = E_{Y,\hat{Y}}(Y - \hat{Y})^2 - E_Y(Y - SY)^2 - E_{\hat{Y}}(\hat{Y} - S\hat{Y})^2$

Therefore, requirement ❶ leads to four possible generalized definitions.

$$bias^2(\hat{Y}) = L(SY, S\hat{Y}) \quad (10)$$

$$bias^2(\hat{Y}) = E_Y L(Y, S\hat{Y}) - E_Y L(Y, SY) \quad (11)$$

$$bias^2(\hat{Y}) = E_{\hat{Y}} L(SY, \hat{Y}) - E_{\hat{Y}} L(\hat{Y}, S\hat{Y})$$

$$bias^2(\hat{Y}) = E_{Y,\hat{Y}} L(Y, \hat{Y}) - E_Y L(Y, SY) - E_{\hat{Y}} L(\hat{Y}, S\hat{Y})$$

Again these definitions will not be consistent for general loss functions. However, (10) is the only one that fulfills requirement ❸. Therefore, for an arbitrary symmetric loss function L we generalize the concepts of bias and variance in the following way.

	Loss Function	
	Squared Error	General
Variance	$E_{\hat{Y}}(\hat{Y} - E\hat{Y})^2$ $S\hat{Y} = \arg\min_\mu E_{\hat{Y}}(\hat{Y} - \mu)^2$	$E_{\hat{Y}} L(\hat{Y}, S\hat{Y})$ $S\hat{Y} = \arg\min_\mu E_{\hat{Y}} L(\hat{Y}, \mu)$
Bias^2	$(E_Y Y - E_{\hat{Y}}\hat{Y})^2$	$L(SY, S\hat{Y})$

In this formulation variance captures the average deviation between $\hat{Y}$ and its closest systematic value, measured relative to L. Similarly bias measures the distance between the systematic parts of $\hat{Y}$ and Y relative to L. Our definition of bias is equivalent to that of bias^2 for squared error. Note that while these definitions fulfill the natural requirements of bias and variance as defined by ❶ through ❸ they are not unique. For any given loss function there may be more than one set of definitions that satisfy these requirements. This is particularly true if the loss function is asymmetric. In this paper we have restricted to considering symmetric loss functions because it is not clear what interpretation to put on variance and bias in the asymmetric case. In addition to having the correct intuitive properties, for squared error loss, these definitions will also provide an additive decomposition of the prediction error. However, for general loss functions no such decomposition will be possible. In fact, we show in Section 4.1 that one may construct examples where the variance and bias of an estimator, $\hat{Y}$, are constant but the reducible prediction error changes as one alters the distribution of the response variable, Y. In the next section we provide a second set of definitions which form a decomposition of prediction error into variance and systematic effects.

3 Bias and variance effect

In this section we develop a second set of bias/variance definitions which provide an additive decomposition of the prediction error. In Section 3.2 we detail the theoretical and experimental relationships between these definitions and those proposed in the previous section.

3.1 An additive decomposition of prediction error

Often we will be interested in the *effect* of bias and variance. For example, in general, it is possible to have an estimator with high variance but for this variance to have little impact on the prediction error. It is even possible for increased variance to decrease the prediction error, as we show in Section 4.2. We call the change in error caused by variance the *variance effect* (VE) and the change in error caused by bias the *systematic effect* (SE). For squared error loss the variance effect is equal to the variance and the systematic effect is equal to the bias squared. However, in general the relationships will be more complicated.

Recall in the standard situation we can decompose the prediction error as follows.

$$E_{Y,\hat{Y}}(Y-\hat{Y})^2 = \underbrace{Var(Y)}_{\text{irreducible error}} + \underbrace{bias^2(\hat{Y}) + Var(\hat{Y})}_{\text{reducible error}} \tag{12}$$

However, note that

$$\begin{aligned} Var(Y) &= E_Y(Y-E_Y Y)^2 \\ bias^2(\hat{Y}) &= (E_Y Y - E_{\hat{Y}}\hat{Y})^2 = E_Y[(Y-E_{\hat{Y}}\hat{Y})^2 - (Y-E_Y Y)^2] \\ Var(\hat{Y}) &= E_{\hat{Y}}(\hat{Y}-E_{\hat{Y}}\hat{Y})^2 = E_{Y,\hat{Y}}[(Y-\hat{Y})^2 - (Y-E_{\hat{Y}}\hat{Y})^2] \end{aligned}$$

Hence, if L_S is squared error loss, then an equivalent decomposition to (12) is

$$\begin{aligned} \underbrace{E_{Y,\hat{Y}}L_S(Y,\hat{Y})}_{\text{Error}} &= \underbrace{E_Y L_S(Y,E_Y Y)}_{Var(Y)} + \underbrace{E_Y[L_S(Y,E_{\hat{Y}}\hat{Y}) - L_S(Y,E_Y Y)]}_{bias^2(\hat{Y})} \\ &\quad + \underbrace{E_{Y,\hat{Y}}[L_S(Y,\hat{Y}) - L_S(Y,E_{\hat{Y}}\hat{Y})]}_{Var(\hat{Y})} \end{aligned} \tag{13}$$

This decomposition does not rely on any special properties of squared error and will hold for any symmetric loss function. Notice that, in this formulation, $bias^2$ is simply the change in the error of predicting Y, when using $E_{\hat{Y}}\hat{Y}$, instead of $E_Y Y$; in other words it is the change in prediction error caused by bias. This is exactly what we have defined as the systematic effect. Similarly $Var(\hat{Y})$ is the change in prediction error when using $\hat{Y}$, instead of $E_{\hat{Y}}\hat{Y}$, to predict Y; in other words the change in prediction error caused by variance. This is what we have defined as the variance effect. Therefore (13) provides a natural approach to defining the systematic and variance effects for a general symmetric loss function, L. Namely

$$\begin{aligned} VE(\hat{Y},Y) &= E_{Y,\hat{Y}}[L(Y,\hat{Y}) - L(Y,S\hat{Y})] \\ SE(\hat{Y},Y) &= E_Y[L(Y,S\hat{Y}) - L(Y,SY)] \end{aligned}$$

Notice that the definitions of variance and systematic effects respectively correspond to (7) and (11). We now have a decomposition of prediction error into errors caused by variability in Y i.e. $Var(Y)$, bias between Y and $\hat{Y}$ i.e. $SE(\hat{Y},Y)$ and variability in $\hat{Y}$ i.e. $VE(\hat{Y},Y)$.

$$\begin{aligned} E_{Y,\hat{Y}}L(Y,\hat{Y}) &= \underbrace{E_Y L(Y,SY)}_{Var(Y)} + \underbrace{E_Y[L(Y,S\hat{Y}) - L(Y,SY)]}_{SE(\hat{Y},Y)} \\ &\quad + \underbrace{E_{Y,\hat{Y}}[L(Y,\hat{Y}) - L(Y,S\hat{Y})]}_{VE(\hat{Y},Y)} \\ &= Var(Y) + SE(\hat{Y},Y) + VE(\hat{Y},Y) \end{aligned}$$

As is the case with squared error loss, $Var(Y)$ provides a lower bound on the prediction error. In the case of 0-1 loss it is equivalent to the Bayes error rate. Different approaches have been taken to this term with some authors incorporating it (Tibshirani, 1996; Domingos, 2000) and others not (Dietterich and Kong, 1995). In practice $Var(Y)$ can either be estimated, as we demonstrate in Section 6.2, or assumed to be zero by setting $SY = Y$. This allows an individual to choose whether they wish to incorporate a Bayes error type term or not.

3.2 Relation between bias, variance and their effects

The following theorem summarizes the main theoretical relationships between variance and variance effect and between bias and systematic effect.

Theorem 1 *Provided the loss function is strictly convex;*

1. *Under squared error loss, the bias and systematic effects are identical. Similarly the variance and variance effects are identical.*
2. *The bias and systematic effect of an estimator will be identical if the Bayes error rate is zero i.e. $Y = SY$ for all inputs.*
3. *An estimator with zero bias will have zero systematic effect.*
4. *An estimator with zero variance will have zero variance effect.*

The proofs of these results are immediate from the fact that, for convex loss functions, $L(a,b) = 0$ implies $a = b$. The first result draws the connection between bias, variance, systematic and variance effects in the familiar setting of squared error loss. The second result is particularly important because it implies that, provided the noise level in the data is low, bias and the systematic effect will be very similar. When the noise level is high there is no guaranteed relationship between these two quantities. However, the median correlation between bias and systematic effect for the 6 data sets examined in Section 5 is 96.6%. This suggests that in practice there may be a strong relationship between the two quantities. This would mean that bias was a good predictor of its effect on the error rate. In other words, an estimator that tends to be low bias may also tend to have a small contribution from bias to the error rate.

Apart from the case where the variance is zero or the loss function is squared error there is also no theoretical relationship between variance and variance effect. However, the median correlation between these two numbers on the data sets of Section 5 is 81.1%, again suggesting a strong relationship may exist in practice. In other words an estimator that tends to be low variance may also tend to have a small contribution from variance to the error rate.

4 Alternatives to squared error loss

In this section we show how the definitions from Sections 2 and 3 can be applied to specific loss functions.

4.1 Polynomial loss

The squared error loss function can be generalized using $L(a,b) = |a-b|^p$. With $p = 2$ this gives squared error loss while $p = 1$ gives absolute loss. Using this loss function, the generalized definitions of variance

and bias from Section 2 become

$$\begin{aligned} Var(Y) &= E_Y L(Y, SY) = E_Y|Y - SY|^p && (14)\\ Var(\hat{Y}) &= E_{\hat{Y}} L(\hat{Y}, S\hat{Y}) = E_{\hat{Y}}|\hat{Y} - S\hat{Y}|^p && (15)\\ bias(\hat{Y}) &= L(SY, S\hat{Y}) = |SY - S\hat{Y}|^p && (16) \end{aligned}$$

where $SY = \arg\min_\mu E_Y|Y-\mu|^p$ and $S\hat{Y} = \arg\min_\mu E_{\hat{Y}}|\hat{Y}-\mu|^p$. While the systematic and variance effects of Section 3 reduce to

$$\begin{aligned} VE(\hat{Y}, Y) &= E_{Y,\hat{Y}}(|Y - \hat{Y}|^p - |Y - S\hat{Y}|^p) && (17)\\ SE(\hat{Y}, Y) &= E_Y(|Y - S\hat{Y}|^p - |Y - SY|^p) && (18) \end{aligned}$$

Notice that, with $p = 2$, $S\hat{Y} = E_{\hat{Y}}\hat{Y}$ and (14) through (18) reduce to the standard variance and bias (squared) definitions.

Alternatively, with $p = 1$, $S\hat{Y}$ becomes the median of $\hat{Y}$ and (14) through (18) become

$$\begin{aligned} Var(Y) &= E_Y|Y - med(Y)| \quad \text{(irreducible error)}\\ Var(\hat{Y}) &= E_{\hat{Y}}|\hat{Y} - med(\hat{Y})|\\ bias(\hat{Y}) &= |med(Y) - med(\hat{Y})|\\ VE(\hat{Y}, Y) &= E_{Y,\hat{Y}}(|Y - \hat{Y}| - |Y - med(\hat{Y})|)\\ SE(\hat{Y}, Y) &= E_Y(|Y - med(\hat{Y})| - |Y - med(Y)|) \end{aligned}$$

While it will often be the case that an estimator with large bias will have a large systematic effect and similarly an estimator with high variance will have a high variance effect, this is not necessarily the case. A simple example using absolute loss provides an illustration. Suppose Y is a random variable with the following distribution.

y	0	1	2
$P(Y=y)$	$a/4$	$1/2$	$(2-a)/4$

We choose to estimate Y using the constant $\hat{Y} = 2$. Note that both $Var(\hat{Y})$ and $VE(\hat{Y}, Y)$ are zero so the systematic effect is the only relevant quantity in this case. Clearly, for $0 < a < 2$, $med(Y) = 1$ and $med(\hat{Y}) = 2$ so $bias(\hat{Y}) = 1$ and

$$\begin{aligned} SE(\hat{Y}, Y) &= E_Y(|Y - med(\hat{Y})| - |Y - med(Y)|)\\ &= 2 \cdot \frac{a}{4} + 1 \cdot \frac{1}{2} + 0 \cdot \frac{2-a}{4} - (1 \cdot \frac{a}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{2-a}{4})\\ &= a/2 \end{aligned}$$

Notice that for any value of a between 0 and 2 the median of Y is equal to 1 so the bias remains constant as a changes. Clearly the systematic effect on the prediction error is not a function of the bias. In fact, as a approaches 0 so does the systematic effect. Hence, unlike squared error loss, it is possible to have bias which does not increase the prediction error.

4.2 Classification problems

When training a classifier, the most common loss function is $L(a, b) = I(a \neq b)$. We will now use the notation C and SC instead of $\hat{Y}$ and $S\hat{Y}$ to emphasize the fact that this is a classification problem so our predictor

typically takes on categorical values: $C \in \{1, 2, \ldots, K\}$ for a K class problem. We will also use the notation T and ST instead of Y and SY as the response variable in this situation is often referred to as the target. Further define

$$\begin{aligned} P_i^T &= P_T(T = i) \\ P_i^C &= P_C(C = i) \end{aligned}$$

where i runs from 1 to K. Note that P_i^C is based on averages over training sets. From (8) and (9) we note that, with this loss function, the systematic parts of T and C are defined as

$$\begin{aligned} ST &= \arg\min_i E_T(I(T \neq i)) = \arg\max_i P_i^T \\ SC &= \arg\max_i P_i^C \end{aligned}$$

ST is the Bayes classifier while SC is the mode of C. The variance and bias components are defined as

$$\begin{aligned} Var(T) &= P_T(T \neq ST) = 1 - \max_i P_i^T \\ Var(C) &= P_C(C \neq SC) = 1 - \max_i P_i^C \qquad (19) \\ bias(C) &= I(SC \neq ST) \qquad (20) \\ VE(C,T) &= P_{T,C}(T \neq C) - P_T(T \neq SC) \\ &= P_{SC}^T - \sum_i P_i^T P_i^C \qquad (21) \\ SE(C,T) &= P_T(T \neq SC) - P_T(T \neq ST) \\ &= P_{ST}^T - P_{SC}^T \qquad (22) \end{aligned}$$

Note that in general $Var(C)$ and $VE(C,T)$ will not be equal. In fact they may not even be related. Again we provide a simple example involving a three class problem to illustrate this point. Suppose T, at a fixed input X, has the following distribution.

t	0	1	2
$P_T(T = t\|X)$	0.5	0.4	0.1

Further suppose that we have two potential classifiers, C_1 and C_2, and, at the same input X, they have the following distributions over training samples.

c	0	1	2
$P_{C_1}(C_1 = c\|X)$	0.4	0.5	0.1
$P_{C_2}(C_2 = c\|X)$	0.1	0.5	0.4

For both classifiers the systematic part is, $SC = 1$. While the systematic part of T is $ST = 0$. Hence both classifiers are "biased" and the systematic effect is 0.1. In other words, if C_1 and C_2 had no variability, so they always classified to Class 1, their error rate would be 0.1 above the minimum i.e. Bayes error rate. The two classifiers have identical distributions except for a permutation of the class labels. Since the labels have no ordering any reasonable definition of variance should assign the same number to both classifiers. Using (19) we do indeed get the same variance for both.

$$Var(C_1) = Var(C_2) = 1 - 0.5 = 0.5$$

However, the effect of this variance is certainly not the same.

$$\begin{aligned} VE(C_1,T) &= P_{T,C_1}(T \neq C_1) - P_T(T \neq SC_1) = 0.59 - 0.6 = -0.01 \\ VE(C_2,T) &= P_{T,C_2}(T \neq C_2) - P_T(T \neq SC_2) = 0.71 - 0.6 = 0.11 \end{aligned}$$

The variance of C_1 has actually caused the error rate to decrease while the variance of C_2 has caused it to increase. This is because the variance in C_1 is a result of more classifications being made to Class 0 which is the Bayes class while the variance in C_2 is a result of more classifications being made to Class 2. Friedman (1996) noted, that for 0-1 loss functions, increasing the variance can actually cause a reduction in the error rate as we have seen with this example.

5 A comparison of definitions

Dietterich and Kong (1995), Kohavi and Wolpert (1996), Breiman (1996b), Tibshirani (1996), Heskes (1998) and Domingos (2000) have all provided alternative definitions for the bias and variance of a classifier. In this section we discuss these definitions and compare them with the more general ones provided in Sections 2 and 3.

Kohavi and Wolpert define bias and variance of a classifier in terms of the squared error when comparing P_i^C to P_i^T. For a two class problem they define the squared bias as $(P_1^T - P_1^C)^2$ and the variance as $P_1^C(1-P_1^C)$ which are as one would expect for squared error. These definitions suffer from the fact that they attempt to assess how closely the distribution of the classifier P_i^C matches the distribution of the target P_i^T. However, in general one does not necessarily want these probabilities to match. For example a classifier that assigns probability 1 to $\arg\max_i P_i^T$ i.e. $P_{ST}^C = 1$ will produce the lowest expected (Bayes) error rate yet the Kohavi and Wolpert approach would assign a non-zero bias to this classifier. In general one is more interested in whether $C = T$ than whether $P_i^T = P_i^C$. In fact, the Bayes classifier will always have $P_i^T \neq P_i^C$ unless $P_i^T = 1$.

Dietterich and Kong define $bias = I(P(C \neq T) \geq 1/2)$ and $var = P(C \neq T) - bias$. This gives a decomposition of the prediction error into

$$P(C \neq T) = var + bias$$

From these definitions we note the following.

- Although not immediately apparent, this definition of bias coincides with (20) for the 2 class situation.
- No allowance is made for any noise or Bayes error term. As a result the bias estimate will tend to be greatly exaggerated.
- For $K > 2$ the two definitions are not consistent which can be seen from the fact that for our definition of bias the Bayes classifier will have zero bias while for Dietterich and Kong's it is possible for the Bayes classifier to have positive bias.
- The variance term will be negative whenever the bias is non zero.

Breiman's definitions are in terms of an "aggregated" classifier which is the equivalent of SC for a 0-1 loss function. He defines a classifier as unbiased, at a given input, X, if $ST = SC$ and lets U be the set of all X at which C is unbiased. He also defines the complement of U as the bias set and denotes it by B. He then defines the bias and variance over the entire test set as

$$bias(C) = P_X(C \neq T, X \in B) - P_X(ST \neq T, X \in B)$$
$$var(C) = P_X(C \neq T, X \in U) - P_X(ST \neq T, X \in U)$$

This is equivalent to defining bias and variance at a fixed X as

$$bias = \begin{cases} P(C \neq T) - P(ST \neq T) & ST \neq SC \\ 0 & ST = SC \end{cases}$$

$$var = \begin{cases} P(C \neq T) - P(ST \neq T) & ST = SC \\ 0 & ST \neq SC \end{cases}$$

This definition has the following appealing properties.

- Bias and variance are always non-negative.
- If C is deterministic then its variance is zero (hence SC has zero variance).
- The bias and variance of the Bayes classifier (ST) is zero.

However, this approach has a couple of significant problems. First, it can not easily be extended to loss functions other than 0-1. Second, at any fixed input X the entire reducible error, i.e. total error rate less Bayes error rate, is either assigned to variance, if C is unbiased at X, or to bias, if C is biased at X. It is reasonable to assign all the reducible error to variance if C is unbiased because in this case if C did not vary it would be equal to the Bayes classifier. In fact for these inputs Breiman's definitions coincide with those of this paper. However, when C is biased it is not reasonable to assign all reducible error to bias. Even when C is biased, variability can cause the error rate to increase or decrease (as illustrated in Section 4.2) and this is not reflected in the definition.

Tibshirani defines variance, bias and a prediction error decomposition for *classification rules (categorical data)*. Within this class of problems his definition of variance is identical to (19). He defines a quantity AE (Aggregation Effect), which is equal to (21), and for most common loss functions his definition of bias is equivalent to (22). This gives the following decomposition of prediction error,

$$P(C \neq T) = P(T \neq ST) + Bias(C) + AE(C)$$

which is identical to ours. However, it should be noted that although these definitions are generalizable to any symmetric loss function they do not easily extend beyond the class of "classification rules" to general random variables, e.g. real valued. It is comforting that when we restrict ourselves to this smaller class the two sets of definitions are almost identical.

Based on an earlier version of this paper, Heskes suggests using the rules from Section 2.2 to construct a bias/variance decomposition for a Kullback-Leibler class of loss functions. This measures the error between the target density $q(t)$ and the classifier's density $\hat{p}(t)$. He defines variance and bias terms which are an exact analog of those presented in this paper when measuring error in densities. It is shown that the error can be decomposed into an additive combination of the two terms. Interestingly, for this particular class of loss functions, as with squared error, it is possible to define a single quantity to serve the purpose of bias and systematic effect and a single quantity to serve the purpose of variance and variance effect. This is an elegant approach but, unfortunately, does not apply to all loss functions. For example it does not directly apply to 0-1 loss functions. An attempt is made to extend these results to zero-one loss by taking the limit case of a log-likelihood-type error. When this is performed the following decomposition is produced for a fixed value of T.

$$P(C \neq T) = (P(C \neq T) - P(C \neq SC)) + P(C \neq SC)$$

The last term is defined as variance and is identical to the variance definitions in this paper and that of Tibshirani. The first term, when summed over the distribution of T, is defined as a combination of bias and intrinsic noise. Unfortunately, as the author points out, in taking the limiting case the first term can no longer be considered the error of the systematic (or average) classifier. Hence it losses any natural interpretation as bias.

Domingos provides definitions of bias and variance which are identical to those in this paper. He then suggests the following decomposition of the error term into

$$E_{T,C}L(T,C) = c_1\text{Noise} + \text{Bias} + c_2\text{Variance} \tag{23}$$

where c_1 and c_2 are factors that depend on the loss function. For example for zero one loss, with two classes, $c_1 = 2P(T = ST) - 1$ and $c_2 = \pm 1$. While (23) appears to provide an additive decomposition of the error rate c_1 and c_2 are in fact functions of bias and variance. It can be shown that $c_1 = (1 - 2\text{Bias})(1 - 2\text{Variance})$ and $c_2 = 1 - 2\text{Bias}$ so that the decomposition can be rewritten in several alternative forms. For example

$$E_{T,C}L(T,C) = (1 - 2\text{Bias})(1 - 2\text{Variance})\text{Noise} + \text{Bias} + (1 - 2\text{Bias})\text{Variance}$$

Thus, the decomposition is multiplicative. It would be interesting to study the relationship between Friedman's and Domingos' theories, both of which suggest a multiplicative decomposition. See Friedman (1996) for further discussion of several of these definitions.

6 Experiments

In this section we demonstrate the definitions of Sections 2 and 3 on several data sets. In Section 6.1 we use simulated data to give a numerical comparison of many of the bias/variance definitions that were discussed in Section 5. Then in Section 6.2 we illustrate how to estimate variance, bias, variance effect and systematic effect on real data sets. These experiments demonstrate how one might implement the various bias/variance quantities on real data sets. They are not intended to provide any empirical justification for the definitions themselves.

6.1 Experimental study of different definitions

To provide an experimental comparison of some of the definitions for variance and bias that have been suggested, we performed two simulation studies. The first simulation consisted of a classification problem with 26 classes, each distributed according to a standard bivariate normal with identity covariance matrix but differing means. Many independent training sets with 10 observations per class were chosen. On each of these training sets 7 different classifiers were trained and their classifications, on a large set of test points, were recorded. Based on the classifications from the different training sets estimates for bias and variance, averaged over the input space, were calculated for each of the classifiers. The 7 different classifiers were linear discriminant analysis (LDA) (Fisher, 1936), ECOC (Dietterich and Bakiri, 1995), bagging (Breiman, 1996a), a tree classifier (Breiman *et al.*, 1984), and $1, 5$ and 11 nearest neighbors (Fix and Hodges, 1951; Cover and Hart, 1967; Stone, 1977). The ECOC and bagging classifiers were both produced using decision trees as the base classifier. On the first 4 classifiers 100 training sets were used. However, it was discovered that the estimates of bias for nearest neighbors were inaccurate for this number so 1000 training sets were used for the last 3 classifiers. Estimates for bias and variance were made using Dietterich, Breiman and Kohavi & Wolpert's definitions as well as those given in this paper. The results are shown in Table 1.

Notice that LDA performs exceptionally well. This is not surprising because LDA is asymptotically optimal for mixtures of normals as we have in this case. Both Breiman's bias estimate and the systematic effect indicate no effect from bias. This is comforting since we know that LDA has the correct model for this data set. The estimate of bias from (20) is slightly above zero, 1.6%. This is due to the relatively low number of training samples. It can be shown that this estimate will converge to zero as the number of training sets increases. Since the bias estimate is averaged over the input space, one can interpret it as the proportion of the space where the classifier tends more often to classify to a class other than the Bayes class. For example

the ECOC method can be seen to classify to a class other than the Bayes class over about 5% of the input space, indicating a fairly low level of bias.

Also notice that Breiman's bias and variance estimates are very similar to the systematic and variance effect estimates from (22) and (21). His estimate of the bias contribution seems to be consistently below or equal to that of the systematic effect. This slight difference between the two definitions is due to the fact that, at any given test point, all the reducible error is attributed to either bias or variance (see Section 5). Dietterich's definitions produce quite different estimates. They tend to attribute almost all the error rate to bias rather than variance. This is partly due to the fact that no allowance is made for the positive Bayes error of 23.1%. However, even when this is subtracted off there are still some anomalies such as LDA having a negative bias. Kohavi and Wolpert provide their own definition for noise which does not correspond to the standard Bayes error. In general it produces a lower estimate. In turn this tends to cause their variance term to be inflated relative to the variance effect defined in this paper or Breiman's definition of variance. The Kohavi and Wolpert definition of bias provides roughly similar estimates to that of Breiman and of the systematic effect from this paper.

It is well known that the nearest neighbors classifier tends to experience decreased variance in its classifications when the number of neighboring points are increased. One can gain an idea of the effect on variance and bias by examining the three nearest neighbor classifiers. As one would expect, the variance, and variance effect, decrease as the number of neighbors increase. However, the bias estimate also decreases slightly which is not what we would expect. This happens with most of the definitions. In fact the bias is not decreasing. There is a tendency to overestimate bias if it is very low because of the skewed nature of the statistic. 11-nearest neighbors averages each of its classifications over 11 points for each training data set so is using $11,000$ data points. This produces a good estimate for bias. However, 1-nearest neighbors is only using $1,000$ data points which gives a less accurate estimate. It is likely in both cases that the true bias is almost zero. This is evidenced by the fact that the systematic effect is zero.

The ECOC and Bagging classifiers are both examples of so called ensembles or majority vote classifiers. They are constructed by combining the classifications from 100 of the tree classifiers from Table 1. According to the theories of Section 1 these methods should produce lower variance classifiers. However, while both methods do reduce the variance, and variance effect, they also reduce the bias, and systematic effects. Clearly the reason for the success of these methods is more complicated than simply a reduction in variance. Finally note that, while in theory there need not be any relationship between bias and systematic effect and between variance and variance effect, in this particular example there is a strong relationship. For example, the correlation between variance and variance effect, among the 7 classifiers, is 99%. As observed earlier there is some evidence that in practice bias and variance may be good predictors for systematic and variance effects.

The second data set is similar to the first except that in this one there were only 10 classes and 5 training data points per class. For this data set eight classifiers were used. They were LDA, ECOC, bagging, tree classifiers with $5, 8$ and 13 terminal nodes, 1-nearest neighbor and 11-nearest neighbor. The results are presented in Table 2. Again LDA performs extremely well, with a very low bias and Breiman's definitions produce similar results to those of (21) and (22). Notice that Dietterich's definition can result in a negative variance. Also note that while in theory the variance effect can be negative it is not for any of the examples we examine. As the number of terminal nodes in a tree increases we would expect its bias to decrease and variance to increase. In this example the bias, and systematic effect, do decrease. However, the variance, and variance effect, also decrease. This can happen if, by increasing the number of terminal nodes, we average over less variable data points.

In summary, it appears that Dietterich's definitions assign far too high a proportion of the prediction error to bias rather than variance. His definitions do not take into account the Bayes error rate. As this error rate increases, the bias, by his definition, will also tend to increase which does not seem sensible. This definition

Classifier	LDA	ECOC	Bagging	Tree	1NN	5NN	11NN
Bias (K and W)	2.5	0.7	1.7	1.8	0.1	0.3	0.7
Variance (K and W)	7.4	13.8	13.1	16.8	16.0	12.4	11.0
Noise (K and W)	14.9	14.9	14.9	14.9	14.9	14.9	14.9
Bias (Dietterich)	21.5	27.4	27.7	32.2	27.7	25.3	24.1
Bias less Bayes Error	−1.6	4.3	4.6	9.1	4.6	2.2	1.0
Variance (Dietterich)	3.3	1.9	2.0	1.3	3.3	2.3	2.4
Bias (Breiman)	0.0	0.4	1.1	1.6	0.1	0.0	0.0
Variance (Breiman)	1.7	5.9	5.5	8.8	7.8	4.5	3.4
Bias	1.6	5.2	6.1	8.5	1.6	1.2	0.9
Variance	10.5	20.6	19.3	25.1	24.8	18.8	16.3
Systematic Effect	0.0	0.5	1.5	2.2	0.0	0.0	0.0
Variance Effect	1.7	5.8	5.1	8.2	7.9	4.5	3.5
Bayes Error	23.1	23.1	23.1	23.1	23.1	23.1	23.1
Prediction Error	24.8	29.3	29.7	33.5	31.0	27.6	26.5

Table 1: Bias and variance for various definitions calculated on a simulated data set with 26 classes.

of bias may work better for a two class problem. Alternatively the definitions of Kohavi and Wolpert, while allowing for a noise term, tend to underestimate the Bayes error rate and and as a consequence overestimate the variance component. Both Breiman's definitions and those presented in this paper seem to produce reasonable estimates with Breiman's tending to put slightly more weight on variance.

6.2 Estimating bias and variance on real-world data sets

Finally we illustrate how to estimate bias, variance, systematic effect, variance effect and Bayes error on four real world data sets from the UCI repository. The data sets were glass, breast cancer, vowel and dermatology. Estimating bias, systematic effect and Bayes error rate is difficult when the underlying distribution is unknown. Previous authors (Kohavi and Wolpert, 1996; Domingos, 2000) have solved this problem by assuming the noise level to be zero. This allows bias to be estimated but unfortunately tends to result in it being drasticly overestimated because any variability in the target is added to the bias term. This is less of a problem if one is only interested in the change in bias or variance for given classifiers. However, often one is attempting to decide whether bias or variance are the primary cause of error in a classifier. If bias is the main problem one should use a more flexible classifier while the reverse is true if variance is the problem. By assuming a noise level of zero and hence overestimating bias one could easily be lead to fit a more flexible classifier when in reality a less flexible classifier is required. Hence we take an alternative approach to the problem by estimating the noise level using the following method. If multiple targets are observed at each input then the noise level can be estimated by calculating the proportion of targets that differ from the most common class. In practice it is rare to observe multiple targets at each input so we use targets at nearby inputs. Hence a decision must be made as to the size of the neighborhood to be used. In this study it was found that using targets at the 3 closest inputs to each given target provided reasonable estimates of noise level and hence bias. This procedure implicitly assumes that the probability distribution of targets is continuous over the input space. However, this does not seem to be an unreasonable assumption in most situations and is likely to produce more accurate estimates than setting the noise level to zero.

We were able to calculate *ST* by taking the most common class among the 3 nearest neighbors to each input. This allowed the Bayes error rate to be estimated. To calculate the other quantities we used a Boot-

Classifier	LDA	ECOC	Bagging	Tree_5	Tree_8	Tree_{13}	1NN	11NN
Bias (K and W)	1.2	0.8	0.9	15.2	2.8	1.1	0.1	1.9
Variance (K and W)	8.5	15.0	13.3	27.7	22.3	17.7	16.3	17.4
Noise (K and W)	14.2	14.2	14.2	14.2	14.2	14.2	14.2	14.2
Bias (Dietterich)	12.5	24.8	20.8	73.3	41.0	28.8	24.5	18.3
Bias less Bayes Error	−8.0	4.3	0.3	52.8	20.5	8.3	4.0	−2.2
Variance (Dietterich)	11.4	5.2	7.4	−16.1	−2.8	11.7	6.1	15.3
Bias (Breiman)	0.0	0.6	0.6	17.3	2.7	0.8	0.0	0.1
Variance (Breiman)	3.3	8.8	7.2	19.3	13.9	11.7	10.1	12.9
Bias	1.3	5.5	5.0	33.0	13.3	5.8	1.0	1.5
Variance	12.0	21.4	19.3	43.8	30.1	26.0	23.7	25.4
Systematic Effect	0.0	0.8	0.9	17.5	3.3	1.0	0.0	0.1
Variance Effect	3.3	8.7	7.0	19.1	13.4	11.5	10.1	12.9
Bayes Error	20.5	20.5	20.5	20.5	20.5	20.5	20.5	20.5
Prediction Error	23.9	30.0	28.4	57.1	37.2	33.0	30.6	33.5

Table 2: Bias and variance for various definitions calculated on a simulated data set with 10 classes.

strap approach (Efron and Tibshirani, 1993). This involves resampling the original data to produce new data sets with similar distributions to the original. We produced 50 so called bootstrap data sets and fit the various classifiers to each one in turn. From these fits we estimated *SC*, the most commonly classified class for each input. This in turn meant that bias, variance, systematic effect and variance effect could be estimated. A cross-validation procedure was used to provide estimates for a new test data set. Cross-validation is performed by removing a portion of the data (e.g. 10%), training the classifier on the remainder and producing predictions on the left out data. This procedure is then repeated by removing another portion of the data until a prediction has been produced for each input. Five fold cross-validation, i.e. leaving out 20% of the data at each step, was performed on all the data sets except the Vowel data which already had a separate test set. The 7 classifiers used were LDA, Bagging (using 20 trees each with 5 terminal nodes), 5 and 10 terminal node trees and 1, 5 and 11 nearest neighbors. The results are shown in Tables 3 through 6.

We examine the glass data first. The Bayes error is high for this data. If it was assumed to be zero the systematic effect terms would all increase by about 17% causing a possibly undue focus on bias as the cause of errors. Despite the adjustment for the Bayes error the bias and systematic effect terms for LDA are still large. This indicates that the data does not follow a Gaussian distribution. However, we note that the variance and variance effect terms are very low indicating that a large reduction in error may be caused by fitting a more flexible classifier. Such a classifier would likely increase the error from variance but more than offset this by a decrease in the error from bias. The tree classifier results are as we would predict. Namely decreasing bias and systematic effects and increasing variance and variance effects as the number of terminal nodes increases. The nearest neighbor results are also as we would predict with increasing bias and decreasing variance as the number of neighbors increases. The lone exception is the variance term for 5 nearest neighbors which increases over 1 nearest neighbors. This is likely a result of ties among multiple classes which nearest neighbors breaks at random. Such ties are not possible with 1 nearest neighbors. This effect was apparent in all four data sets. Finally we note that even after taking account of the large Bayes error rate all 7 classifiers clearly exhibit considerably more bias than variance indicating reductions in the error rate could be achieved by fitting a more flexible classifier such as a neural network.

The cancer data set has a much lower Bayes error and all 7 classifiers perform fairly well. Again LDA has very low variance and variance effect but also low bias indicating that, in this case, its model assumptions are

Classifier	LDA	Bagging	Tree 5	Tree 10	1NN	5NN	11NN
Bias	54.6	30.2	33.1	28.5	11.6	18.7	31.4
Variance	16.4	19.2	19.5	23.2	12.1	20.4	15.7
Systematic Effect	39.0	18.3	20.2	12.4	9.2	13.0	22.2
Variance Effect	1.2	2.8	2.1	6.5	3.7	5.4	0
Bayes Error	17.0	17.0	17.0	17.0	17.0	17.0	17.0
Prediction Error	57.2	38.1	39.3	35.9	29.9	35.4	39.2

Table 3: Estimates for bias, variance, systematic effect and variance effect on glass data set.

Classifier	LDA	Bagging	Tree 5	Tree 10	1NN	5NN	11NN
Bias	2.6	2.8	2.6	2.2	2.2	1.4	1.7
Variance	0.4	3.2	3.8	3.7	1.7	1.8	1.1
Systematic Effect	2.2	1.6	1.5	0.9	2.0	0.6	0.9
Variance Effect	0	1.5	2.4	2.4	0.1	0.9	0.3
Bayes Error	2.0	2.0	2.0	2.0	2.0	2.0	2.0
Prediction Error	5.0	5.1	5.9	5.3	4.1	3.5	3.2

Table 4: Estimates for bias, variance, systematic effect and variance effect on breast cancer data set.

more reasonable. As we would expect the 10 node tree classifier has a reduced bias over the 5 node version. Interestingly this does not appear to come at the expense of increased variance. The Bagging classifier causes a small reduction in variance and variance effect over the 5 node tree classifier. The variance and systematic effects are very low for all three nearest neighbor classifiers which makes them difficult to compare.

Interestingly, given the difficulty of the Vowel data set, the Bayes error is extremely low. All the classifiers have very high bias relative to their variance. Clearly more flexible classifiers are required to fit the test data. This is highlighted by the fact that the 10 node tree has a significantly lower bias than the 5 node tree. Notice the extremely strong relationship between bias and systematic effect for all classifiers (over a 99.99% correlation). This is because the Bayes error rate is close to zero so that statement 2 of Theorem 1 applies to this data.

The dermatology data set also has a fairly low Bayes error rate. However, notice the effect that including this term has on the variance and systematic effect terms for the 10 node tree. When the Bayes error rate is taken account of we notice that the variance effect is larger than the systematic one suggesting that a classifier with fewer nodes may produce a lower error rate. However, if the Bayes error rate is assumed to be zero the systematic effect increases to 4.3 which is significantly larger than the variance effect and incorrectly implies that a tree with more nodes should be fit.

7 Conclusion

In this paper we have produced two sets of bias/variance definitions. The first satisfies natural properties such as measuring variability of an estimator and distance between the systematic parts of the estimator and response. The second provide an additive decomposition of the prediction error in terms of the "effects" of bias and variance. The definitions apply to all symmetric loss functions and types of predictors/classifiers, i.e. real valued or categorical. When squared error loss is used with real valued random variables they reduce to the standard definitions but in general allow a much more flexible class of loss functions to be used. While, in

Classifier	LDA	Bagging	Tree 5	Tree 10	1NN	5NN	11NN
Bias	55.6	64.3	64.9	58.0	43.5	39.8	40.3
Variance	21.0	51.6	51.4	33.7	12.4	22.3	26.0
Systematic Effect	55.6	63.9	65.0	58.0	43.5	39.8	40.3
Variance Effect	−1.1	7.1	5.9	4.7	0.9	5.2	6.6
Bayes Error	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Prediction Error	54.7	71.2	71.1	62.9	44.6	45.2	47.1

Table 5: Estimates for bias, variance, systematic effect and variance effect on vowel data set.

Classifier	LDA	Bagging	Tree 5	Tree 10	1NN	5NN	11NN
Bias	3.0	11.9	11.1	6.0	2.2	1.8	2.9
Variance	2.2	5.1	5.4	5.4	2.0	2.6	2.3
Systematic Effect	1.0	9.4	8.8	2.3	1.5	0.4	1.5
Variance Effect	1.3	−0.6	−0.5	2.9	0.8	1.6	0.7
Bayes Error	2.0	2.0	2.0	2.0	2.0	2.0	2.0
Prediction Error	4.3	10.8	10.3	7.1	4.3	4.0	4.2

Table 6: Estimates for bias, variance, systematic effect and variance effect on dermatology data set.

theory it is possible for there to be little or no relationship between variance and variance effect and between bias and systematic effect, in practice there is some evidence that these quantities are highly correlated. This means that one may be able to use variance and bias to predict the systematic and variance effects of an estimator/classifier on a new data set.

Acknowledgments

The author wishes to thank Trevor Hastie for valuable discussions. He would also like to thank the editor and referees for numerous suggestions.

References

Breiman, L. (1996a). Bagging predictors. *Machine Learning 26, No. 2* , 2, 123–140.

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

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regresion Trees*. Wadsworth (Since 1993 this book has been published by Chapman & Hall, New York.).

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **IT-13**, 21–27.

Dietterich, T. and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research 2* 263–286.

Dietterich, T. G. and Kong, E. B. (1995). Error-correcting output coding corrects bias and variance. *Proceedings of the 12th International Conference on Machine Learning* 313–321 Morgan Kaufmann.

Domingos, P. (2000). A unified bias-variance decomposition and its applications. *Proceedings of the 17th International Conference on Machine Learning* .

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London : Chapman and Hall.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Fix, E. and Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. *Technical Report, Randolph Field, Texas: USAF School of Aviation Medicine* .

Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. *Machine Learning : Proceedings of the Thirteenth International Conference* .

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association* **62**, 1159–1178.

Friedman, J. H. (1996). On bias, variance, 0/1-loss, and the curse of dimensionality. *Technical Report, Department of Statistics, Stanford University* .

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

Heskes, T. (1998). Bias/variance decompositions for likelihood-based estimators. *Neural Comuptation* **10**, 1425–1433.

Kohavi, R. and Wolpert, D. (1996). Bias plus variance decomposition for zero-one loss functions. *Machine Learning : Proceedings of the Thirteenth International Conference* .

Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics* **26**, 1651–1686.

Stone, C. (1977). Consistent nonparametric regression (with discussion). *Annals of Statistics* **5**, 595–645.

Tibshirani, R. (1996). Bias, variance and prediction error for classification rules. *Technical Report, Department of Statistics, University of Toronto* .

Wolpert, D. (1997). On bias plus variance. *Neural Computation* **9**, 1211–1243.