Generalizations of the Bias/Variance Decomposition for Prediction Error

GARETH JAMES*
and
TREVOR HASTIE
Dept. of Statistics, Stanford University

January 29, 1997

Abstract

The bias and variance of a real valued random variable, using squared error loss, are well understood. However because of recent developments in classification techniques it has become desirable to extend these concepts to general random variables and loss functions. The 0-1 (misclassification) loss function with categorical random variables has been of particular interest. We explore the concepts of variance and bias and develop a decomposition of the prediction error into functions of the systematic and variable parts of our predictor. After providing some examples we conclude with a discussion of the various definitions that have been proposed.

1 Introduction

A lot of work has recently been conducted on developing general definitions for bias and variance, in particular for misclassification errors. However in attempting this task two questions arise. Namely

- what do these quantities measure?
- and why are they useful?

*The authors were partially supported by grant DMS-9504495 from the National Science Foundation, and grant ROI-CA-72028-01 from the National Institutes of Health.

1.1 Bias and Variance

In the regression setting the variance of an estimator $\hat{Y}$ is defined as $E(\hat{Y} - E\hat{Y})^2$. An equivalent definition is

$$Var(\hat{Y}) = \min_a E(\hat{Y} - a)^2$$

where a is non random. If we define

$$S\hat{Y} = \arg\min_a E(\hat{Y} - a)^2$$

then $Var(\hat{Y})$ is a measure of the *expected distance*, in terms of squared error loss, of the random quantity $(\hat{Y})$ from its nearest non random number $(S\hat{Y})$. We call $S\hat{Y}$ the systematic part of $\hat{Y}$ and use the notation $S\hat{Y}$ to emphasize that S is an operator acting on the distribution of $\hat{Y}$. In this case $S\hat{Y}$ will be equal to $E\hat{Y}$.

If we use $\hat{Y}$ to estimate a parameter θ then the squared bias of $\hat{Y}$ is $(S\hat{Y} - \theta)^2$. The bias when using $\hat{Y}$ to predict Y, where $\hat{Y}$ and Y are independent random variables, is less well defined. However from the decomposition,

$$\begin{aligned} E(Y - \hat{Y})^2 &= E(Y - SY)^2 + E(\hat{Y} - SY)^2 \\ PE(Y, \hat{Y}) &= Var(Y) + MSE(\hat{Y}, SY) \end{aligned}$$

where $SY = \arg\min_a E(Y - a)^2$, we can see that the problem of predicting Y is equivalent to one of estimating SY. This is because $Var(Y)$ is independent of $\hat{Y}$ so the mean squared error between $\hat{Y}$ and SY is the only quantity that we have control over. This shows us that we can think of bias as a measure of the distance between the systematic parts of $\hat{Y}$ and Y ($S\hat{Y}$ and SY).

By writing $\hat{Y} = S\hat{Y} + \epsilon$ we see that it is possible to decompose our random variable into systematic $(S\hat{Y})$ and random (ϵ) parts. Both parts contribute to any error we may make in estimation but their causes and cures can differ markedly.

1.2 Standard Prediction Error Decomposition

It is well known that we can decompose the expected squared error of $\hat{Y}$ from Y as follows

$$E(\hat{Y} - Y)^2 = \underbrace{Var(Y)}_{\text{irreducible error}} + \underbrace{bias^2(\hat{Y}, SY) + Var(\hat{Y})}_{\text{reducible error}} \tag{1}$$

so the *expected loss* of using $\hat{Y}$ is the sum of the variance of $\hat{Y}$ and Y plus the squared distance between their systematic components. 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. it provides a measure of the variability of $\hat{Y}$ about $S\hat{Y}$
2. and it indicates the effect of this variance on the prediction error.

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

1. it provides a measure of the distance between the systematic components of Y and $\hat{Y}$
2. and by squaring it we see the effect of this bias on the prediction error.

This double role of both bias and variance is so automatic that we often fail to consider it. However when we extend these definitions to arbitrary loss functions it will not, in general, be possible to define one statistic to serve both purposes.

2 Generalizing the definitions

Often squared error is a very convenient loss function to use. It possesses well known mathematical properties such as the bias/variance decomposition (1) 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 much more realistic.

So how might we extend these concepts of variance and bias to general loss functions? There is one obvious requirement that it seems natural for any generalization to fulfill.

❶ *When using squared error loss our general definitions must reduce to the standard ones.*

Unfortunately ❶ is not a strong enough requirement to ensure a unique generalization. This is a result 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 are all equivalent for squared error.

- $Var(\hat{Y}) = \min_a E(\hat{Y} - a)^2 = E(\hat{Y} - S\hat{Y})^2$
- $Var(\hat{Y}) = MSE(\hat{Y}, SY) - bias^2(\hat{Y}, SY) = E(\hat{Y} - SY)^2 - (S\hat{Y} - SY)^2$
- $Var(\hat{Y}) = PE(Y, \hat{Y}) - E(Y - S\hat{Y})^2 = E(Y - \hat{Y})^2 - E(Y - S\hat{Y})^2$

These lead naturally to three possible generalized definitions.

I. $Var(\hat{Y}) = \min_a EL(\hat{Y}, a) = EL(\hat{Y}, S\hat{Y})$

II. $Var(\hat{Y}) = EL(\hat{Y}, SY) - L(S\hat{Y}, SY)$

III. $Var(\hat{Y}) = EL(Y, \hat{Y}) - EL(Y, S\hat{Y})$

where L is a general loss function, $S\hat{Y} = \arg\min_a L(\hat{Y}, a)$ and $SY = \arg\min_a L(Y, a)$

For general loss functions these last three definitions certainly need not be consistent. This inconsistency accounts for some of the differences in the definitions that have been proposed. For example Tibshirani (1996) bases his definition of variance on I while Dietterich and Kong (1995) base their's more closely on III. We will see later that both I and III are useful for measuring different quantities.

What other requirements should a definition of variance fulfill? We can think of $\hat{Y} \sim g(F_{Tr})$ where g is a function that depends on the method used to obtain $\hat{Y}$ from the observations and F_{Tr} is the distribution of the observations or *training data (Tr)*. While $Y \sim F_{Te}$ where F_{Te} is the distribution of the *test data*. Often F_{Tr} and F_{Te} are assumed the same but in general they need not be. So we see that variance is a function of g and F_{Tr} but is not a function of F_{Te}. This is desirable because it allows us to compare estimators across different test sets (a low variance estimator for one test set will also be low variance for a second test set). So another natural requirement is

❷ *The variance must not be a function of the distribution of the test data,* F_{Te}.

This requirement rules out II and III given above which in general will be a function of F_{Te}.

There is a similar requirement on the bias. Since bias is a measure of the distance between the systematic components of $\hat{Y}$ and Y we require that

❸ *The bias must be a function of $\hat{Y}$ and Y through $S\hat{Y}$ and SY only (ie bias must be a function of $S\hat{Y}$ and SY)*

With ❶,❷ and ❸ in mind the most natural definitions of bias and variance are

	Loss Function	
	Squared Error	General
Variance	$E(\hat{Y} - S\hat{Y})^2$ $S\hat{Y} = \arg\min_a E(\hat{Y} - a)^2$	$EL(\hat{Y}, S\hat{Y})$ $S\hat{Y} = \arg\min_a EL(\hat{Y}, a)$
Bias^2	$(SY - S\hat{Y})^2$	$L(SY, S\hat{Y})$

This definition of variance is identical to that given in Tibshirani (1996) but our definition of bias differs from his. Our definition of bias is equivalent to that of bias^2 for squared error. It should be noted that, even with the restrictions we have listed, these definitions by no means represent a unique generalization of the concepts of bias and variance. However, as we will see in the next section, these statistics may not be our primary concern.

3 Bias and Variance effect

While these definitions fulfill the stated conditions they have one major drawback. Namely in general there is no way to decompose the error into any function of bias and variance, as is the case for squared error loss (1). In fact it is possible to construct examples (see section 4.1) where the variance and bias are constant but the reducible prediction error changes as we alter the test distribution.

Often we will be interested in the *effect* of bias and variance. For example it is possible to have an estimator with high variance but for this variance to have little impact on the error rate. It is even possible for increased variance to cause a lower error rate (see 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 the variance effect is just the variance and the systematic effect is the bias squared. However in general this will not be the case.

Recall in the standard situation we can decompose the expected squared error as follows,

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

but note

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

Remember for squared error $SY = EY$ and $S\hat{Y} = E\hat{Y}$. This gives the following decomposition

$$\underbrace{EL_S(Y,\hat{Y})}_{PE} = \underbrace{EL_S(Y,SY)}_{Var(Y)} + \underbrace{E[L_S(Y,S\hat{Y}) - L_S(Y,SY)]}_{bias^2(\hat{Y},SY)} + \underbrace{E[L_S(Y,\hat{Y}) - L_S(Y,S\hat{Y})]}_{Var(\hat{Y})}$$

where L_S is squared error loss. Note that everything is defined in terms of prediction error of Y with respect to L_S.

Notice that, in this formulation, $bias^2$ is simply the change in prediction error when using $S\hat{Y}$, instead of SY, to predict 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 $S\hat{Y}$, to predict Y; in other words the change in prediction error caused by variance. This is what we have defined as the Variance Effect.

This decomposition will hold for any loss function so in general we define

$$\begin{aligned} SE(\hat{Y}, Y) &= E[L(Y, S\hat{Y}) - L(Y, SY)] \\ \text{and} \qquad VE(\hat{Y}, Y) &= E[L(Y,\hat{Y}) - L(Y, S\hat{Y})] \end{aligned}$$

Notice that the definition of VE corresponds to III in section 2. We now have a decomposition of prediction error into errors caused by variability in Y ($Var(Y)$), bias between Y and $\hat{Y}$ ($SE(\hat{Y}, Y)$) and variability in $\hat{Y}$ ($VE(\hat{Y}, Y)$)

$$\begin{aligned} EL(Y, \hat{Y}) &= \underbrace{EL(Y, SY)}_{Var(Y)} + \underbrace{E[L(Y, S\hat{Y}) - L(Y, SY)]}_{SE(\hat{Y}, Y)} + \underbrace{E[L(Y, \hat{Y}) - L(Y, S\hat{Y})]}_{VE(\hat{Y}, Y)} \\ &= Var(Y) + SE(\hat{Y}, Y) + VE(\hat{Y}, Y) \qquad (2) \end{aligned}$$

Now in general there is no reason for $Var(\hat{Y})$ to equal $VE(\hat{Y}, Y)$ or for $bias(\hat{Y}, SY)$ to equal $SE(\hat{Y}, Y)$. *Often it will be the variance and bias effects that we are more interested in rather than the variance and bias itself.* One of the nice properties of squared error loss is that $VE = Var$ so the variance effect, like the variance, is constant over test sets. In general this will not be the case.

Note that due to the fact L is a loss function, $Var(Y) \geq 0$, and by the definition of SY, $SE(\hat{Y}, Y) \geq 0$. However the only restriction on $VE(\hat{Y}, Y)$ is, $VE(\hat{Y}, Y) \geq -SE(\hat{Y}, Y)$. Indeed we will see examples where the variance effect is negative.

4 Examples

All calculations in the following two examples are performed at a fixed input X. I have not included X in the notation to avoid confusion.

4.1 Absolute Loss

Suppose we use the loss function $L(a, b) = |a - b|$. What decomposition does this give?

$$\begin{aligned} EL(\hat{Y}, Y) &= Var(Y) + SE(\hat{Y}, Y) + VE(\hat{Y}, Y) \\ &= EL(Y, SY) + E[L(Y, S\hat{Y}) - L(Y, SY)] \\ &+ E[L(Y, \hat{Y}) - L(Y, S\hat{Y})] \end{aligned}$$

$$
\begin{aligned}
\Rightarrow E|Y-\hat{Y}| &= E|Y-SY| + E(|Y-S\hat{Y}| - |Y-SY|) \\
&+ \ E(|Y-\hat{Y}| - |Y-S\hat{Y}|) \\
&= E|Y-med(Y)| + E(|Y-med(\hat{Y})| - |Y-med(Y)|) \\
&+ \ E(|Y-\hat{Y}| - |Y-med(\hat{Y})|)
\end{aligned}
$$

where $med(Y)$ is the median of Y.

This gives

$$
\begin{aligned}
VE(\hat{Y},Y) &= E(|Y-\hat{Y}| - |Y-med(\hat{Y})|) \\
Var(\hat{Y}) &= EL(\hat{Y},S\hat{Y}) = E|\hat{Y}-med(\hat{Y})| \\
SE(\hat{Y},Y) &= E(|Y-med(\hat{Y})| - |Y-med(Y)|) \\
bias(\hat{Y},SY) &= L(SY,S\hat{Y}) = |med(Y)-med(\hat{Y})| \\
Var(Y) &= \text{irreducible error} = E|Y-med(Y)|
\end{aligned}
$$

A simple example illustrates the concepts involved. Suppose Y is a random variable with the following distribution

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

We will start with $a=1$. Suppose our estimator is simply the constant $\hat{Y}=2$. Then clearly $med(Y)=1$ and $med(\hat{Y})=2$ so $bias(\hat{Y},SY)=1$. 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.

$$
\begin{aligned}
SE(\hat{Y},Y) &= E(|Y-med(\hat{Y})| - |Y-med(Y)|) \\
&= E(|Y-2| - |Y-1|) \\
&= 1-1/2 = 1/2
\end{aligned}
$$

So the SE is not equal to the bias. We can show that SE is not a function of the bias by altering a. Notice that for $0<a<2$ the median of Y remains unchanged at 1. So the bias is also constant. However.

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

So as a approaches 0 so does the SE! In other words it is possible to have an estimator that is systematically wrong but with an arbitrarily low reducible loss associated with it.

4.2 0-1 Loss

Now suppose our 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 loss function is normally used in classification problems so our predictor typically takes on categorical values: $C \in \{1, 2, \ldots, K\}$ for a K class problem.

Further define

$$\begin{aligned} P_i^Y &= Pr(Y = i) \\ \text{and} \quad P_i^C &= Pr(C = i) \end{aligned}$$

where i runs from 1 to K. Recall that $C \sim g(F_{Tr})$, and hence P_i^C are based on averages over training sets. With this loss function we see

$$\begin{aligned} SY &= \arg\min_i E(I(Y \neq i)) \\ &= \arg\min_i \sum_{j \neq i} P_j^Y \\ &= \arg\max_i P_i^Y \quad \text{ie the bayes classifier} \\ \text{and} \quad SC &= \arg\max_i P_i^C \quad \text{ie the mode of } C \end{aligned}$$

We now get

$$\begin{aligned} VE(C, Y) &= E(I(Y \neq C) - I(Y \neq SC)) \\ &= P(Y \neq C) - P(Y \neq SC) \\ &= \sum_i P_i^Y (1 - P_i^C) - (1 - P_{SC}^Y) \\ Var(C) &= \min_a EI(C \neq a) \\ &= P(C \neq SC) \\ &= 1 - \max_i P_i^C \end{aligned}$$

$$
\begin{aligned}
SE(C,Y) &= E(I(Y \neq SC) - I(Y \neq SY)) \\
&= P(Y \neq SC) - P(Y \neq SY) \\
&= P^Y_{SY} - P^Y_{SC} \\
&= \max_i P^Y_i - P^Y_{SC} \\
bias(C,SY) &= I(SC \neq SY) \\
Var(Y) &= EI(Y \neq SY) \\
&= P(Y \neq SY) \\
&= 1 - \max_i P^Y_i
\end{aligned}
$$

A simple example will again provide some illumination. Suppose Y has the following distribution.

y	0	1	2
$Pr(Y = y)$	0.5	0.4	0.1

Now we compare two classifier random variables (at a fixed predictor X) with the following distributions.

c	0	1	2
$Pr(C_1 = c)$	0.4	0.5	0.1
$Pr(C_2 = c)$	0.1	0.5	0.4

$SY = 0$, $SC_1 = SC_2 = 1$ and $SE(C,Y)$ equals 0.1 for both classifiers. These two classifiers have identical distributions except for a permutation of the class labels. Since the labels have no ordering we would hope that both classifiers have the same variance. In fact $Var(C_1) = Var(C_2) = 1 - 0.5 = 0.5$. However the effect of this variance is certainly not the same for each classifier.

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

In fact 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 0 which is the bayes group while the variance in C_2 is a result of more classifications being made to 2 which is a very unlikely group to occur. Therefore, we see that it does not necessarily follow that increasing the variance of C_1 would cause a further reduction in $VE(C_1, Y)$ or that decreasing the variance of C_2 would cause a reduction in $VE(C_2, Y)$. 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 Conclusion

The definitions, given in this paper, apply to all loss functions and types of predictors/classifiers (ie real valued or categorical). They reduce to the standard definitions in the case of squared error with real valued random variables but allow a much more flexible class of loss functions to be used.

We have seen that it is possible to construct a definition of variance that is independent from F_{Te} but in general the variance effect will be a function of the test distribution. In practice it may often be the case that variance and variance effect are fairly well correlated so we can use variance to predict the effect of an estimator/classifier on a new data set.

When we move away from squared error loss functions we not only need to consider the bias and variance of an estimator but also the effect of these quantities on a specific test set distribution. Hence all four statistics, $Var(\hat{Y})$, $VE(\hat{Y}, Y)$, $Bias(\hat{Y}, SY)$ and $SE(\hat{Y}, Y)$, are important. If we wish to minimize prediction error, for a fixed test distribution, then it is $VE(\hat{Y}, Y)$ and $SE(\hat{Y}, Y)$ that we should concentrate on. However if we are instead interested in predicting the systematic and variance effects for general test distributions it is $Var(\hat{Y})$ and $bias(\hat{Y}, SY)$ that we will be concerned with.

6 Discussion

Dietterich and Kong (1995), Kohavi and Wolpert (1996), Breiman (1996b), Tibshirani (1996), and Friedman (1996) have all recently written papers on the topic of bias and variance for classification rules.

Kohavi and Wolpert (1996) define bias and variance of a classifier in terms of the squared error when comparing P_i^C to P_i^Y. For a two class problem they define the squared bias as $(P_1^Y - P_1^C)^2$ and the variance as $P_1^C(1 - P_1^C)$ which are as one would expect for squared error. As a result the Bayes Classifier will have a positive squared bias unless $P_1^Y \in \{0, 1\}$.

The definitions of Dietterich and Kong (1995), Breiman (1996b), and Tibshirani (1996) are more similar in spirit to those in this paper. Dietterich and Kong (1995) define $bias = I(Pr(C \neq Y) \geq 1/2)$ and $var = Pr(C \neq Y) - bias$. This gives the decomposition

$$Pr(C \neq Y) = var + bias$$

From these definitions we can note the following

- although not immediately apparent, this definition of bias coincides with ours ($I(SY = S\hat{Y})$) for the 2 class situation.
- 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 (1995) 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 (1996b) definitions are in terms of an "aggregated" classifier which is the equivalent of SC for a 0-1 loss function. He defines bias as $Pr(Y \neq SC) - Pr(Y \neq SY)$ and variance as $Pr(Y \neq C) - Pr(Y \neq SY)$. These terms are what we call the Systematic and Variance Effects for 0-1 loss. Again this has the property that the variance can be negative.

Tibshirani (1996) 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 that given in this paper. He defines a quantity AE (Aggregation Effect), which is equal to the Variance Effect we have defined, and for most common loss functions his definition of Bias will be equivalent to our Systematic Effect. This gives a decomposition of

$$Pr(C \neq Y) = Pr(Y \neq SY) + Bias(C) + AE(C)$$

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

Friedman (1996) provides a good comparison of the different definitions.

In large part all this work has been spurred on by the success of the *plug in classification techniques* that have recently been developed. A plug in classification technique (PaCT) is one where we take a standard classifier and *plug* it into a new algorithm. Examples include the Error-Correcting method (Dietterich and Bakiri (1995)), the Bagging method (Breiman (1996a)) and the Arcing method (Freund and Schapire (1995)).

References

Breiman, L. (1996a) Bagging predictors, in press, Machine Learning

Breiman, L. (1996b) Bias, Variance, and Arcing Classifiers, Dept. of Statistics, University of California Berkeley, Technical Report

Dietterich, T.G. and Bakiri G. (1995) Solving Multiclass Learning Problems via Error-Correcting Output Codes, Journal of Artificial Intelligence Research 2 (1995) 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 pp. 313-321 Morgan Kaufmann

Efron, B. (1978), Regression and anova with zero-one data, J. Amer. Statist. Assoc. pp. 113-121.

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

Friedman, J.H. (1996) On Bias, Variance, 0/1-loss, and the Curse of Dimensionality, Dept. of Statistics, Stanford University, Technical Report

Kohavi, R. and Wolpert, D.H. (1996) Bias Plus Variance Decomposition for Zero-One Loss Functions, ftp starry.stanford.edu/pub/ronnyk/biasVar.ps

Tibshirani (1996) Bias, Variance and Prediction Error for Classification Rules, Dept. of Statistics, University of Toronto, Technical Report