Accepted for publication in Neural Computation.

Bias/variance decompositions for likelihood-based estimators

Tom Heskes
Foundation for Neural Networks, University of Nijmegen
Geert Grooteplein 21, 6525 EZ Nijmegen, The Netherlands

Abstract

The bias/variance decomposition of mean-squared error is well understood and relatively straightforward. In this note a similar simple decomposition is derived, valid for any kind of error measure that, when using the appropriate probability model, can be derived from a Kullback-Leibler divergence or loglikelihood.

Introduction

Finding bias/variance decompositions for all kinds of error measures or loss functions is an active area of research. The decomposition for mean-squared error is well-known and easily derived [see e.g. (Geman, Bienenstock & Doursat 1992)]. Recently, several suggestions have been made for other loss functions such as zero-one loss [see (Breiman 1996, Dietterich & Bakiri 1995, Friedman 1996, James & Hastie 1997, Kohavi & Wolpert 1996, Tibshirani 1996, Wolpert 1997) and references therein]. The generalization of the decomposition for mean-squared error to a decomposition for zero-one loss depends on one's definition of desirable properties for the bias and the variance term. In this note, we will follow the requirements and definitions stated in (James & Hastie 1997). Applying these definitions to the Kullback-Leibler divergence, we will arrive at a simple generalization of the decomposition for mean-squared error.

Theory

Let Y be a random variable, which may be either discrete or continuous. We will proceed as if Y is continuous, where the discrete case follows immediately if one replaces integrals by summations and "probability densities" by "probability distributions". $q(y)$ is defined as the target probability density function that $Y = y$; $\hat{p}(y)$ is an estimator of this density. For example, $\hat{p}(y)$ may correspond to a probability statement derived from the output of a neural network (see the examples below). We have a (possibly infinite) ensemble of such estimators. Expectation with respect to this ensemble is indicated by the operator E. We use the Kullback-Leibler divergence

$$K(q, \hat{p}) \equiv \int dy\, q(y) \log \left[\frac{q(y)}{\hat{p}(y)} \right] \tag{1}$$

to measure the distance between densities $\hat{p}(y)$ and $q(y)$. The goal is to find a decomposition of the error $EK(q, \hat{p})$ in a bias and variance term.

In the usual setting, the ensemble consists of models that are obtained through application of a learning algorithm on different training sets, generated from the same problem domain. In a decomposition of the average error of these models, the bias is supposed to measure how closely the learning algorithm's average guess matches the target and the variance how much the learning algorithm's guess "bounces around" for different training sets (Kohavi & Wolpert 1996). Modifications on a learning algorithm tend to have an opposite effect on the bias and the variance: an increase in the number of degrees of freedom usually leads to a smaller bias and a higher variance.

Note that, contrary to most other papers on bias/variance decompositions, we do not write the loss function as a direct measure of the distance between a model's output and a target. Instead we first translate the model's output to a probability statement and then define the loss function as the Kullback-Leibler divergence between this probability statement and a target probability (see below for the straightforward generalization to the case where the target probability $q(y)$ is unknown and only a realization $Y = t$ is provided). A similar approach is pursued in (Wolpert 1997).

As suggested in (James & Hastie 1997), we start the decomposition by defining the variance as the smallest average distance, in this case the smallest average Kullback-Leibler divergence, between an estimator $\hat{p}(y)$ and some "average model" $\bar{p}(y)$. The asymmetry of the Kullback-Leibler divergence forces us to be more precise. We keep the densities $\hat{p}(y)$ in the role of estimators and define the average model as the target density that leads to the smallest possible Kullback-Leibler divergence between the target and the estimators:

$$\text{variance} = \min_{a: \int dy\, a(y) = 1} EK(a, \hat{p}) = EK(\bar{p}, \hat{p}) \, . \tag{2}$$

Introducing a Langrange multiplier for the constraint $\int dy\, a(y) = 1$ and taking the functional derivative to $a(y)$, we easily obtain for the average model

$$\bar{p}(y) = \frac{1}{Z} \exp[E \log \hat{p}(y)] \, , \tag{3}$$

with Z a normalization constant independent of y. In other words, the average model $\bar{p}(y)$ is a (normalized) geometric mean of the densities $\hat{p}(y)$, rather than a arithmetic mean, as for example proposed in (Hall 1987, Wolpert 1997). In the literature on combining experts' probability statements, (3) is called a logarithmic opinion pool [see e.g. (Bordley 1982, Genest & Zidek 1986, Heskes 1998) or (Jacobs 1995) for the similar but somewhat more involved supra Bayesian techniques]. A disadvantage of the logarithmic opinion pool is that if any of the experts assigns probability zero to a particular outcome, the average model assigns probability zero, independent of what the other experts claim. This property of the logarithmic opinion pool, however, is perfectly consistent with a Bayesian point of view[1] and is only a drawback if the densities $\hat{p}(y)$ are not carefully estimated (Bordley 1982).

[1] Whenever a Bayesian assigns probability zero or one, all further discussion is closed: no amount of new information can ever change his mind (Bordley 1982).

The bias is defined as the distance $K(q, \bar{p})$ between the average model and the target distribution. Substituting (3) into (1) we obtain

$$\text{bias } = K(q, \bar{p}) = EK(q, \hat{p}) + \log Z \; .$$

Using (3), the second term on the righthand side can be transformed into

$$\begin{aligned} \log Z &= \log \left[\frac{\exp[E \log \hat{p}(y)]}{\bar{p}(y)} \right] \quad \forall_{y:\bar{p}(y)>0} \\ &= E \left[\int dy \, \bar{p}(y) \log \left(\frac{\hat{p}(y)}{\bar{p}(y)} \right) \right] = -EK(\bar{p}, \hat{p}) = - \text{ variance} \; , \end{aligned}$$

with the variance defined in (2). Rearrangement of terms then gives the desired decomposition:

$$\text{error } = EK(q, \hat{p}) = K(q, \bar{p}) + EK(\bar{p}, \hat{p}) = \text{ bias } + \text{ variance} \; . \tag{4}$$

Other bias/variance decompositions often include a term measuring the intrinsic noise, which is a lower bound on the error that can be obtained by any learning algorithm. A learning algorithm which reproduces the probability distribution $q(y)$ has Kullback-Leibler divergence equal to zero. This explains why there is no intrinsic noise term in (4).

Equation (4) gives a decomposition for the Kullback-Leibler divergence between probability densities[2]. Now suppose that we do not know the complete target distribution $q(y)$, but only have a particular observation $Y = t$. In that case, it is more appropriate to consider the loglikelihoods $\log \hat{p}(t)$ for which we, following the same lines, obtain

$$-E \log \hat{p}(t) = - \log \bar{p}(t) + EK(\bar{p}, \hat{p}) \; . \tag{5}$$

The first term on the righthand side is the error of the average model, the second term the variance of the models in the ensemble. For a further decomposition of the error of the average model into an intrinsic noise term and a bias term, we have to integrate again over the probability density generating the targets[3]. This then yields

$$\begin{aligned} \text{error} = -E \left[\int dt \, q(t) \log \hat{p}(t) \right] &= - \int dt \, q(t) \log q(t) + K(q, \bar{p}) + EK(\bar{p}, \hat{p}) \\ &= \text{ intrinsic noise } + \text{ bias } + \text{ variance} \; . \end{aligned} \tag{6}$$

The decompositions (4) and (6) only differ in their definition of the error function. With the error definition in (6), the intrinsic noise term is equal to the Shannon entropy of the density $q(y)$. In the following examples, we will illustrate the decomposition (5) for a single observation t.

[2] If we define the error between the estimated probability $\hat{p}(y)$ and target $q(y)$ "the other way around", i.e., $K(q, \hat{p}) \equiv \int dy \, \hat{p}(y) \log[\hat{p}(y)/q(y)]$, we obtain exactly the same decomposition, but with as the average model the *linear* opinion pool $\bar{p}(y) = E\hat{p}(y)$ [see e.g. (Genest & Zidek 1986, Jacobs 1995) for a discussion of linear opinion pools]. This error measure, however, is much less in use since it cannot be transformed into a loglikelihood for a (finite set of) observation(s) instead of a target probability.

[3] This can be easily illustrated on the mean-squared error. Suppose (see the first example below) that the average model predicts $\bar{m}$ when the target is t. Without knowing the distribution from which the targets t are drawn, it is impossible to decompose the error $(\bar{m} - t)^2$ into a separate noise and bias term.

Examples

The mean-squared error is a special case of the Kullback-Leibler divergence if we interpret model outputs $\hat{m}$ as estimates of the mean of a normal distribution with some fixed variance σ^2:

$$\hat{p}(y) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left[\frac{-(y-\hat{m})^2}{2\sigma^2}\right] .$$

The logarithmic opinion pool (3) yields as the average model $\bar{p}(y)$ a Gaussian with the same standard deviation and mean $\bar{m} = E\hat{m}$, as expected. The decomposition (5) is, up to an irrelevant proportionality constant, equivalent to the usual one as e.g. in (Geman et al. 1992):

$$E\left[(\hat{m}-t)^2\right] = (\bar{m}-t)^2 + E\left[(\hat{m}-\bar{m})^2\right] .$$

As a generalization, we consider the case where we have estimates $\hat{m}$ and $\hat{\sigma}^2$ for both the mean and the variance (see e.g. (Bishop & Qazaz 1997, Williams 1996) and references therein). The average model $\bar{p}(y)$ is still a Gaussian with mean $\bar{m}$ and variance $\bar{\sigma}^2$ obeying

$$\frac{1}{\bar{\sigma}^2} = E\left(\frac{1}{\hat{\sigma}^2}\right) \quad \text{and} \quad \frac{\bar{m}}{\bar{\sigma}^2} = E\left(\frac{\hat{m}}{\hat{\sigma}^2}\right) ,$$

i.e., the logarithmic opinion pool (3) leads to an averaging of reciprocal variances and a weighted averaging of the estimated means. The decomposition (5) yields

$$E\left[\frac{(\hat{m}-t)^2}{\hat{\sigma}^2} + \log\hat{\sigma}^2\right] = \left[\frac{(\bar{m}-t)^2}{\bar{\sigma}^2} + \log\bar{\sigma}^2\right] + E\left[\frac{(\hat{m}-\bar{m})^2}{\hat{\sigma}^2} + \log\left(\frac{\hat{\sigma}^2}{\bar{\sigma}^2}\right)\right] .$$

The first term between brackets on the righthand side is the error of the average model, the second term measures the variance of the different estimators.

A new decomposition is obtained for the cross-entropy or logarithmic scoring function that can be used for classification purposes. We consider the binary case with Y a binary random variable, e.g., $Y \in \{0,1\}$. In the shorthand notation $\hat{p} \equiv \hat{p}(1)$, the logarithmic opinion pool (3) yields

$$\log\left(\frac{\bar{p}}{1-\bar{p}}\right) = E\left[\log\left(\frac{\hat{p}}{1-\hat{p}}\right)\right] ,$$

i.e., the average model can be found by averaging the logits (log-odds) of the estimated probabilities. Given an observed target t, the decomposition (5) can be written

$$E\left[t\log\hat{p} + (1-t)\log(1-\hat{p})\right] = \\ t\log\bar{p} + (1-t)\log(1-\bar{p}) - E\left[\bar{p}\log\left(\frac{\bar{p}}{\hat{p}}\right) + (1-\bar{p})\log\left(\frac{1-\bar{p}}{1-\hat{p}}\right)\right] . \quad (7)$$

This decomposition can be contrasted with the one proposed in (Wolpert 1997), which for the binary case in our notation reads

$$E\left[t\log\hat{p} + (1-t)\log(1-\hat{p})\right] = \\ t\log\bar{p} + (1-t)\log(1-\bar{p}) - E\left[t\log\left(\frac{\bar{p}}{\hat{p}}\right) + (1-t)\log\left(\frac{1-\bar{p}}{1-\hat{p}}\right)\right] , \quad (8)$$

where the average model is the linear opinion pool[4], i.e., $\bar{p} = E\hat{p}$. The main disadvantage of the decomposition (8) is that the variance term still depends directly on the target t. Whenever the expectation E is defined by averaging over models optimized on training sets generated from the target distribution, the variance term in (7) also depends on the target distribution (see (Wolpert 1997) for a full exposition of this point). However, keeping the operation E fixed, e.g. by keeping the distribution over training sets the same, the variance in (7) is independent of the (distribution of the) target t, whereas the variance in (8) does depend on the target t.

Most recent papers on bias/variance decompositions focus on zero-one loss for classification tasks. Given the target class label t, the loss is 0 if the model's estimate $\hat{y}$ equals t and 1 otherwise. As we will see, we can try to interpret zero-one loss as a limit case of a loglikelihood-type error. Suppose that we transfer the classification $\hat{y}$ into a probability statement which assigns probability 1 to class $\hat{y}$ and probability $\epsilon \ll 1$ to all other class labels $y \neq \hat{y}$:

$$\hat{p}_\epsilon(y) = \begin{cases} 1 & \text{if } y = \hat{y}, \\ \epsilon \ll 1 & \text{if } y \neq \hat{y}. \end{cases}$$

In principle we should normalize this distribution, but it is easy to show that for small ϵ, this normalization constant can be set to one. We call $f(y)$ the fraction of models that assigns the class label y, i.e., $f(y) = E\delta_{\hat{y},y}$. Application of (3) then yields in leading order of ϵ

$$\bar{p}_\epsilon(y) = \epsilon^{\max_{y'} f(y') - f(y)} \,, \tag{9}$$

i.e., in the limit $\epsilon \to 0$, the average model is nothing but the majority vote $\bar{y} = \underset{y}{\operatorname{argmax}} f(y)$. Decomposition (5) is still valid:

$$-E \log \hat{p}_\epsilon(t) = -\log \bar{p}_\epsilon(t) + EK(\bar{p}_\epsilon, \hat{p}_\epsilon) \,.$$

If we divide by $-\log \epsilon$ and take the limit $\epsilon \to 0$ on both sides, we arrive at the decomposition

$$1 - f(t) = [f(\bar{y}) - f(t)] + [1 - f(\bar{y})] \,. \tag{10}$$

Considering the way in which we have arrived at this decomposition for zero-one loss, we are tempted to call the second term between brackets the variance. However, in taking the limit $\epsilon \to 0$, we have lost the interpretation of the first term as the error of the average model. The crux is that the average model (9) in leading order of ϵ still depends on the classification frequencies $f(y)$ which, for that reason, also appear in (10). The average model for $\epsilon = 0$, on the other hand, only depends on the majority vote and is independent of the exact frequencies $f(y)$. For a further decomposition of the first term into a bias term and an inherent noise term, we have to sum over the distribution $q(t)$ that generated the class labels t. Most authors define the inherent noise term to be the error of the Bayes classifier and ascribe the remaining term to

[4] Equation (8) is, of course, true for any definition of the average model $\bar{p}$. Only by averaging the logits, one can make the variance independent of the target t and arrive at (7).

the bias. The exact decomposition seems to be somewhat arbitrary, since in practice one is only interested in changes in the bias and variance terms rather than in their absolute values. Our definition of variance is equivalent to those given in (Tibshirani 1996, James & Hastie 1997).

Discussion

We slightly reformulate what in (James & Hastie 1997) are called obvious requirements for a bias/variance decomposition. These requirements are similar in spirit to the desiderata stated in (Wolpert 1997).

1. The decomposition for the mean-squared error is a special case.

2. The variance does not depend on the target distribution directly. Furthermore it is nonnegative and zero iff all estimators are equivalent.

3. The bias only depends on the target distribution and the "average model", which is defined as the model minimizing the variance.

The main result of this note is that, for any likelihood-based estimator, it is indeed possible to find a decomposition fulfilling these requirements. To see that this is nontrivial, we will sketch how many decompositions are derived (see e.g. (Dietterich & Bakiri 1995, Kohavi & Wolpert 1996, Wolpert 1997)). For convenience, we will stick to the probabilistic notation. One starts by translating the models $\hat{p}(y)$ into some average model $\bar{p}(y)$ and defines the bias to be the error between this model and the target, minus the lowest error that can be obtained by any learning algorithm. In our notation we have

$$\text{bias} \ = K(q, \bar{p}) - K(q, q) \,.$$

The variance is defined as the part of the error that cannot be attributed to the noise and the bias:

$$\text{variance} \ = EK(q, \hat{p}) - K(q, \bar{p}) \,.$$

In principle, there is no need for this variance to fulfill the second requirement. In fact, this is where previously proposed bias/variance decompositions of Kullback-Leibler divergence (see e.g. (Hall 1987, Wolpert 1997)) have to give in. However, we have shown that for *any* likelihood-based estimator

1. there is an average model $\bar{p}(y)$ such that the variance no longer directly depends on the target density $q(y)$;

2. this variance is the average error to this average model;

3. this average model is the model that yields the lowest variance.

The mean-squared error, for which these nice properties have been known for long, appears to be nothing but a special case.

Only in some limit case, zero-one loss can be interpreted as a kind of Kullback-Leibler divergence. The resulting decomposition still obeys the first and second requirement, but the limiting operation "destroys" the third requirement: the bias is no longer just a function of the average model. None of the bias/variance decompositions for zero-one loss suggested in the literature (see (Breiman 1996, Dietterich & Bakiri 1995, Friedman 1996, Kohavi & Wolpert 1996, Tibshirani 1996, Wolpert 1997) and (James & Hastie 1997) for a discussion of most of them) satisfies all three requirements[5]. Most of them either define the bias and take for granted that the variance depends on the distribution of targets (the approach sketched in the beginning of this discussion), or start by defining the variance and are left with the difficult task to interpret the bias. The natural decomposition for likelihood-based estimators obtained in this note, may be seen as an argument in favor of the latter approach.

Acknowledgments

I would like to thank Bert Kappen and the anonymous referees for useful comments on earlier drafts, and Padhraic Smyth for pointing me to the bias/variance decomposition in (Hall 1987). Special thanks to Wim Wiegerinck for showing me the connection between Kullback-Leibler divergence and zero-one loss. This research was supported by the Technology Foundation STW, applied science division of NWO and the technology program of the Ministry of Economic Affairs.

References

Bishop, C. & Qazaz, C. (1997), Regression with input-dependent noise: a Bayesian treatment, *in* M. Mozer, M. Jordan & T. Petsche, eds, 'Advances in Neural Information Processing Systems 9', MIT Press, Cambridge, pp. 347–353.

Bordley, R. (1982), 'A multiplicative formula for aggregating probability assessments', *Management Science* **28**, 1137–1148.

Breiman, L. (1996), Bias, variance, and arcing classifiers, Technical report, University of California, Berkeley.
*http://www.stat.berkely.edu/users/breiman

Dietterich, T. & Bakiri, G. (1995), 'Solving multiclass learning problems via error-correcting output codes', *Journal of Artificial Intelligence Research* **2**, 263–286.

Friedman, J. (1996), On bias, variance, 0/1-loss, and the curse of dimensionality, Technical report, Department of Statistics, Stanford University.
*http://stat.stanford.edu/people/faculty/friedman.html

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

[5] It is possible to simply define the bias and the variance term such that they fulfill all three requirements, but then they do not add up to give the loss function (see (James & Hastie 1997)).

Genest, C. & Zidek, J. (1986), 'Combining probability distributions: a critique and an annotated bibliography', *Statistical Science* **1**, 114–148.

Hall, P. (1987), 'On Kullback-Leibler loss and density estimation', *Annals of Statistics* **15**, 1491–1519.

Heskes, T. (1998), Selecting weighting factors in logarithmic opinion pools, *in* 'Advances in Neural Information Processing Systems 10', MIT Press, Cambridge.

Jacobs, R. (1995), 'Methods for combining experts' probability assessments', *Neural Computation* **7**, 867–888.

James, G. & Hastie, T. (1997), Generalizations of the bias/variance decomposition for prediction error, Technical report, Department of Statistics, Stanford University.
*http://playfair.Stanford.EDU/ gareth/

Kohavi, R. & Wolpert, D. (1996), Bias plus variance decomposition for zero-one loss functions, *in* L. Saitta, ed., 'Proceedings of the 13th International Conference on Machine Learning', Morgan Kaufmann, San Mateo, CA.

Tibshirani, R. (1996), Bias, variance and prediction error for classification rules, Technical report, University of Toronto.
*http://utstat.toronto.edu/tibs/research.html

Williams, P. (1996), 'Using neural networks to model conditional multivariate densities', *Neural Computation* **8**, 843–854.

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