General Bias/Variance Decomposition with Target Independent Variance of Error Functions Derived from the Exponential Family of Distributions

Jakob V. Hansen
Department of Computer Science
University of Aarhus, Ny Munkegade, Bldg. 540
DK-8000 Aarhus C, Denmark
E-mail: vogdrup@daimi.au.dk

Tom Heskes
Foundation for Neural Networks (SNN)
Department of Medical Physics and Biophysics
University of Nijmegen, Geert Grooteplein 21
NL 6525 EZ Nijmegen, The Netherlands
E-mail: tom@mbfys.kun.nl

Abstract

An important theoretical tool in machine learning is the bias/variance decomposition of the generalization error. It was introduced for the mean square error in [3]. The bias/variance decomposition includes the concept of the average predictor. The bias is the error of the average predictor, and the systematic part of the generalization error, while the variability around the average predictor is the variance. We present a large group of error functions with the same desirable properties as the bias/variance decomposition in [3]. The error functions are derived from the exponential family of distributions via the statistical divergence measure. We prove that this family of error functions contains all error functions decomposable in that manner.

We state the connection between the bias/variance decomposition and the ambiguity decomposition [7] and present a useful approximation of ambiguity that is quadratic in the ensemble coefficients.

1 Notation and problem domain

The problem domain of this paper is finding the functional relationship between output and input based on an example set of target-input pairs $S = \{(t_1, x_1), \ldots, (t_n, x_n)\}$. To make this a relevant problem it is assumed that the set is generated with noise from a function $r(\bar{x})$. We wish to find a predictor $f(\bar{w}; \bar{x})$ that is as close as possible to $r(\bar{x})$. The vector $\bar{w}$ refers to the parameters that describe the predictor, e.g. the weights in a neural network. Furthermore, we are interested in the situation where we have an ensemble of predictors characterized by a distribution $E$, which is independent of the noise distribution. The mean operator is denoted $\langle \cdot \rangle_E$. The set of predictors can be finite or infinite. We will generally look at only one input point, so for notational convenience we will omit dependency of functions on input; we also omit the parameters of the predictors. The inaccuracy or error of a predictor is measured with an error function $E(t, f)$.

2 Bias/variance decomposition

If $S$ is noisy it is not guaranteed that $r(\bar{x}_i) = t_i$ for all $i$. It is therefore not optimal to find a predictor with $f(\bar{x}_i) = t_i$ for all $i$. If $S$ is noise-free and we find a predictor with $f(\bar{x}_i) = t_i$ for all $i$, then the predictor can be different from the function $r(\bar{x})$ on all other points. In both cases, by using the principle of Occam’s Razor, the class of possible predictors should be restricted, e.g. by limiting the number of weights in neural networks. This raises an important question: Just how large a class of predictors should be used? If the class is too small, the predictors are too simple and cannot predict the target functions. On the other hand, if the class is too large, the predictors can become too complex and overfit. The two cases correspond to two different...
kinds of errors: Bias and Variance. To fully understand the difference between the errors we look at an ensemble of predictors. The mean of the predictors is the average predictor. The error of the average predictor expresses the systematic error of the predictors, i.e. the bias, while the mean of the error between the predictors and the average predictor expresses the stochastic error i.e. the variance.

Generally, both kinds of errors will be made by an ensemble of predictors. We would like to be able to split the mean of the generalization error into a bias and a variance term: Error = Bias + Variance. This is the bias/variance decomposition. It was introduced in [3] for the mean square error $\langle E_{mse}(t,f) \rangle_F = \frac{1}{2}(t-f)^2$. The average predictor is $\bar{f} = \langle f \rangle_F$. The mean generalization error is $\langle E_{mse}(t,f) \rangle_F$. We have

$$\langle E_{mse}(t,f) \rangle_F = E_{mse}(t,\bar{f}) + \langle E_{mse}(\bar{f},f) \rangle_F,$$

where $E_{mse}(t,\bar{f})$ is the bias and $\langle E_{mse}(\bar{f},f) \rangle_F$ is the variance.

The mean square error and the corresponding bias/variance decomposition have a number of crucial properties: The error is zero and minimal when the predictor equals the target. The bias only depends on the properties: The error is zero and minimal when the bias/variance decomposition have a number of crucial properties.

The above mentioned requirements can be formulated mathematically:

- R1: $\text{argmin}_f E(t,f) = t$.
- R2: $E(t,t) = 0$.
- R3: The bias/variance decomposition:

$$\langle E(t,f) \rangle_F = E(t,\bar{f}) + \langle E(\bar{f},f) \rangle_F,$$

where

$$\bar{f} = \text{argmin}_f \langle E(t,f) \rangle_F.$$

R1 and R2 can be derived from R3.

## 3 General bias/variance decomposition

Not all error functions obey the requirements R1-R3. E.g. the 0-1 loss error function is impossible to decompose as in R3 (see e.g. [2, 6]). A natural question is: What error functions obey the requirements? We prove that only the error functions corresponding to the deviance of one-parameter members of the exponential family of distributions obey R3.

The deviance error function for members of the exponential family is presented in section 4. A sketch of the proof that they are the only error functions obeying the requirements R1-R3 can be found in section 5. In section 6 some examples of deviance error functions are given. In section 7 the connection between the bias/variance decomposition and the ambiguity decomposition is made, and an approximation of ambiguity is presented, which is quadratic in the ensemble coefficient.

## 4 Deviance error function and the exponential family

It is well-known that the mean square error can be interpreted as the negative loglikelihood under a Gaussian noise model. The density of the Gaussian (Normal) distribution, with standard deviation equal to one, is given by

$$p(t|f) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(t-f)^2 \right].$$

The mean square error is connected to the density by

$$E(t, f) = -\log p(t|f) + \log p(t|t),$$

where the last term is added to ensure requirement R2. This is also called the deviance [8]. The Normal distribution is a special case of the one-parameter exponential family. The general form is [1]

$$p(t|f) = \exp[c(f)T(t) + d(f) + S(t)],$$

where $c$ is the canonical link function, the function $T$ is the sufficient statistic, and $d$ is a normalization term. The function $S$ plays no role in the following. The density $p(t|f)$ yields the deviance error function

$$E(t, f) = [c(t) - c(f)]T(t) + d(t) - d(f).$$

To ensure $\text{argmin}_f E(t,f) = t$ we also need the constraint

$$c'(y)T(y) + d'(y) = 0$$

(6)

The Normal distribution has $c(f) = f, T(t) = t$, and $d(f) = -\frac{1}{2} f^2$ for standard deviation equal to one. Furthermore the constraint (6) is obeyed.

The function $d$ is given by the constraint (6), so the deviance error function is completely determined by the canonical link function $c$ and the sufficient statistic $T$.

The error function in (5) upholds R1-R3. R1 follows from the constraint (6). Requirement R2 is obeyed because of the definition of the deviance error. For the error functions in (5) we have the corollary

$$\bar{f} = \text{argmin}_t \langle E(t,f) \rangle_F = c^{-1}(c(f))_F.$$
With this definition of the average predictor the decomposition (2) is easily verified. The variance is by definition independent of the target, and is given by
\[
\text{var}(f) = \langle E(\tilde{f}, f) \rangle_E = \langle d(\tilde{f}) - d(f) \rangle_E
\]  
(7)

5 Error functions with “nice” bias/variance decompositions

As mentioned, the error functions on the form in (5) are the only error functions obeying the requirements R1-R3. Here we sketch the proof.

We will show that \( \frac{\delta E(t, f)}{\delta \tilde{f}} = A_1(f)A_2(t) \). This suffices since it shows that the error functions must be of the form

\[
\text{canonical link} \times \text{ sufficient statistic}
\]

plus some terms depending on either \( t \) or \( f \). The crux of the proof is that slightly changing the distribution over \( f \), the change in the average predictor affects \( \frac{\delta E(t, f)}{\delta \tilde{f}} \) multiplicatively and independently of \( t \).

The derivative with regard to \( t \) of the bias/variance decomposition (2) is

\[
\frac{\partial (E(t, f))}{\partial t} = \frac{\partial E(t, \tilde{f})}{\partial t} + \frac{\partial}{\partial t} \int df \delta(f) E(t, f).
\]

for all distributions \( E \). Let \( \Pi \) denote the distributions with density \( \pi(f) \) and average predictor \( \tilde{f} \). Let \( \delta(f) \) be a function that integrates to zero over the domain of \( f \). Then \( \pi(f|e) = \pi(f) + e\delta(f) \) is a density with average predictor \( \tilde{f}(e) \). We have that

\[
\frac{\partial (E(t, f))}{\partial t} = \frac{\partial (E(t, \tilde{f}))}{\partial t} + \epsilon \int df \delta(f) E(t, f).
\]

Differentiation with regard to \( \epsilon \) yields

\[
\frac{\partial^2 E(t, \tilde{f}(e))}{\partial \epsilon \partial t} \frac{\partial \delta(f)}{\partial \epsilon} = \int df \delta(f) E(t, f).
\]

In the limit \( \epsilon \to 0 \), \( \frac{\partial^2 E(t, \tilde{f}(e))}{\partial \epsilon \partial \tilde{f}} \) becomes \( \frac{\partial^2 E(t, f)}{\partial \tilde{f} \partial f} \). The right hand side does not depend on \( \tilde{f} \), while \( \frac{\partial \delta(f)}{\partial \epsilon} \) does not depend on \( t \). Then it must be that indeed \( \frac{\partial^2 E(t, f)}{\partial \tilde{f} \partial f} = A_1(f)A_2(t) \).

6 Examples of error functions

We consider two special cases, linear sufficient statistic and linear canonical link, and show how they can be regarded as being “transposed”.

The common univariate distributions in the exponential family have linear sufficient statistic, but generally non-linear canonical link. The constraint (6) becomes \( d'(y) = -c'(y)y \), which is equivalent with \( d(y) = -c(y)y + C(y) \), where \( C \) is the anti-derivative of \( c \).

Table 1 gives an overview of some of the error functions with non-linear average predictors. Any constant factors are omitted.

<table>
<thead>
<tr>
<th>DIST.</th>
<th>ERROR FUNCTION</th>
<th>DOMAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( \frac{1}{2}(f-t)^2 )</td>
<td>(-\infty; \infty[)</td>
</tr>
<tr>
<td>Poisson</td>
<td>( [f-t] + t \log \frac{f}{t} )</td>
<td>([0; \infty[)</td>
</tr>
<tr>
<td>Binomial</td>
<td>( t \log \frac{f}{t} + (1-t) \log \frac{1-t}{1-f} )</td>
<td>([0; 1])</td>
</tr>
<tr>
<td>Gamma</td>
<td>( \left( \frac{1}{2} - 1 \right) + \log \frac{f}{t} )</td>
<td>([0; \infty[)</td>
</tr>
<tr>
<td>Inv. Gauss.</td>
<td>( (f-t)^2 / (f^2t) )</td>
<td>([0; \infty[)</td>
</tr>
</tbody>
</table>

Table 1. Err. func. with linear suf. statistics.

As mentioned above the defining functions are the sufficient statistic and the canonical link. By interchanging these two we find the transposed family of error functions. To ensure the constraint (6) the functions \( d \) in the transposed error functions are set to

\[
d_{\text{transpose}}(f) = -d(f) + c(f)T(f).
\]

The error functions in table 1 have linear sufficient statistics. The transposed error functions therefore have linear canonical links. The average predictor is given by \( \tilde{f} = c^{-1}(c(f))_E \), so the transposed error functions have linear average predictors. In table 2 is an overview of transposed error functions with linear average predictor. Constant factors are omitted.

<table>
<thead>
<tr>
<th>DIST.</th>
<th>ERROR FUNCTION</th>
<th>DOMAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( \frac{1}{2}(f-t)^2 )</td>
<td>(-\infty; \infty[)</td>
</tr>
<tr>
<td>Poisson</td>
<td>([f-t] + f \log \frac{f}{t} )</td>
<td>([0; \infty[)</td>
</tr>
<tr>
<td>Binomial</td>
<td>( f \log \frac{f}{t} + (1-f) \log \frac{1-f}{1-t} )</td>
<td>([0; 1])</td>
</tr>
<tr>
<td>Gamma</td>
<td>( \left( \frac{1}{2} - 1 \right) + \log \frac{f}{t} )</td>
<td>([0; \infty[)</td>
</tr>
<tr>
<td>Inv. Gauss.</td>
<td>( (f-t)^2 / (f^2t) )</td>
<td>([0; \infty[)</td>
</tr>
</tbody>
</table>

Table 2. Err. func. with linear can. links.

Note that the transposed error functions have the predictor and the target interchanged compared to their counter-
parts. Also note that the Normal error function and transposed counterpart both are the mean square error. The transposed Binomial error function is not very useful, since it is undefined for target equal to one or zero.

7 Applications

For a finite ensemble there is an alternative to the bias/variance decomposition, namely the ambiguity decomposition. It is mathematically equivalent to the bias/variance decomposition. It was introduced in [7] for the mean square error and linear average predictor. By letting the ensemble predictor be equivalent with the average predictor, the ambiguity decomposition can be generalized for the family of error functions in (4). Since the ensemble set is finite, the mean operator \( \langle c(\hat{f}) \rangle_E \) is defined by a summation \( \sum_{i} \alpha_i c(f_i) \), where \( \sum_i \alpha_i = 1, \alpha_i \geq 0 \). The decomposition is

\[
E(t, \hat{f}) = \langle E(t, f) \rangle_E - \langle E(\hat{f}, f) \rangle_E,
\]

where \( \langle E(\hat{f}, f) \rangle_E \) is the ambiguity. The decomposition proves that the error of the ensemble predictor \( \hat{f} \) always is less than or equal to the mean of the individual ensemble predictors, \( \langle E(t, f) \rangle_E \), for the error functions in (4). The target independent ambiguity has a number of applications. It can be used to estimate the generalization error from unlabeled data, optimize the ensemble coefficients, and find well-suited training examples [7, 5]. An overview of ensemble methods is presented in [4].

The ambiguity \( \langle E(\hat{f}, f) \rangle_E \) is generally non-linear in the coefficients \( \alpha_i \) because of the non-linearity of the error function and/or the average predictor. This makes coefficient optimizing a non-linear programming problem. To circumvent this the ambiguity can be approximated by

\[
\sum_i \alpha_i E(\hat{f}, f_i) \approx \frac{1}{4} \sum_{i,j} \alpha_i \alpha_j [E(f_i, f_j) + E(f_j, f_i)],
\]

which can be found from repeated application of (2) with \( t \) replaced by one of the predictors. The problem of optimizing the coefficients is hereby reduced to a quadratic optimizing problem. This approximation was introduced in [5].

8 Conclusion

We have proven that any error function with the bias/variance decomposition

\[
\langle E(t, f) \rangle_E = E(t, \hat{f}) + \langle E(\hat{f}, f) \rangle_E,
\]

where \( \hat{f} = \text{argmin}_t \langle E(t, f) \rangle_E \) must be of the form

\[
E(t, f) = [c(t) - c(f)]T(t) + d(t) - d(f)
\]

with the constraint \( c'(y)T'(y) + d'(y) = 0 \). These are the error functions corresponding to the deviance of densities in the exponential family of distributions.

We have stated that the bias/variance decomposition is equivalent with the ambiguity decomposition for a finite ensemble of predictors. The ambiguity decomposition is of great practical importance. Often the ambiguity is non-linear in the ensemble coefficients. We have presented an approximation that is quadratic in the ensemble coefficients.

In many applications the mean square error is used, implicitly making Gaussian noise assumption. In this paper we have shown that similar bias/variance decomposition and ensemble averaging techniques can be derived for any noise assumption corresponding to a one-parameter exponential density.

Acknowledgments

We would like to thank Brian Mayoh and Anders Krogh for their contributions. A special thanks to Jens Ledet Jensen for help on statistical subjects.

References