A detailed derivation of the relationship between generalization error and ambiguity in regression ensembles

Gabriele Zenobi
Artificial Intelligence Group
Computer Science Department
Trinity College Dublin
Gabriele.Zenobi@cs.tcd.ie

Abstract

In this technical report we will show the complete sequence of steps for the derivation of the equation of the Ensemble-Error $E = \overline{E} - \overline{A}$, introduced in the paper by Krogh and Vedelsby [1], that describes the error E of an ensemble of networks related to the average error $\overline{E}$ of the single network and the Ambiguity $\overline{A}$ of the ensemble, which in turn is a measure of the "disagreement" among the networks.

1. Introduction

The reason why an ensemble of many different predictors has been widely utilized is because it has been seen that their combination of predictions is better than the one of a single predictor.
A fundamental theoretic work, which confirms in a formal context for ensembles of networks the previous assumption, has been made by Krogh and Vedelsby [1], who have introduced the definition of *Ambiguity* of an ensemble as a measure of the "disagreement" among the networks of the ensemble, and have derived an elegant equation which describes the error of the ensemble in terms of the single network error and of the ambiguity so defined.
All the necessary steps to introduce and derive the equation, together with the motivation of the new definitions and the discussion of some of the results, are given in the paper by Krogh and Vedelsby [1], but some of the algebraic steps are missing in it.
The aim of this report is simply to offer a mathematical complement, in terms of algebraic derivation, to the theory and the equations shown by Krogh and Vedelsby [1], so that only the basic definitions will be reprised, leaving any further consideration about motivation and results to the paper mentioned above. For this reason the notation used, in terms of definition of the problem and variable naming, will be consistent with the theoretic work by Krogh and Vedelsby [1].

2. The Ensemble-Error equation (single input)

From this point up to the end of this report we will assume that our task is to learn a function $f : \Re^n \to \Re$ for which we have available a sample of p samples $(x^\mu, y^\mu) = (x^\mu, f(x^\mu))$ where $x^\mu \in \Re^n$, $\mu = 1,...,p$. Furthermore these examples are assumed to be collected randomly from the distribution *p(x)*.
We suppose the ensemble to be made of *N* neural networks, so that the single output of the network α $(\alpha = 1,...,N)$ on the input vector *x* will be denoted as $V^\alpha(x)$.
Then, by definition, the <u>*Weighted Ensemble Prediction*</u> will be indicated by:

$$(1) \qquad \overline{V}(x) = \sum_\alpha w_\alpha \cdot V^\alpha(x)$$

Where the weights are chosen in order to hold the normalization property: $\sum_\alpha w_\alpha = 1$

The quadratic error of, respectively, the single network α of the ensemble and the ensemble itself, are defined as follows:

$$(2) \qquad \varepsilon^\alpha(x) = [f(x) - V^\alpha(x)]^2$$

$$(3) \qquad e(x) = [f(x) - \overline{V}(x)]^2$$

We then define the <u>*Ambiguity*</u> of the network α as the quadratic difference of the prediction of the network α and the ensemble prediction:

$$(4) \qquad a^\alpha(x) = [V^\alpha(x) - \overline{V}(x)]^2$$

The <u>*Ensemble Ambiguity*</u> on the input *x* is finally defined as the averaged ambiguity of the single networks:

$$(5) \qquad \overline{a}(x) = \sum_\alpha w_\alpha \cdot a^\alpha(x) = \sum_\alpha w_\alpha \cdot [V^\alpha(x) - \overline{V}(x)]^2$$

Given these definitions we may now proceed from the *(5)* adding and subtracting *f(x)*:

$$\overline{a}(x) = \sum_\alpha w_\alpha \cdot [V^\alpha(x) - \overline{V}(x) + f(x) - f(x)]^2 =$$
$$= \sum_\alpha w_\alpha \cdot \{[f(x) - \overline{V}(x)] - [f(x) - V^\alpha(x)]\}^2$$

And then calculating the square:

$$\overline{a}(x) = \sum_\alpha w_\alpha \cdot \left\{ [f(x) - V^\alpha(x)]^2 + [f(x) - \overline{V}(x)]^2 - 2[f(x) - \overline{V}(x)][f(x) - V^\alpha(x)] \right\}$$

The easiest way to reach our goals is now to separate the summation for the first term from the remaining two:

$$\overline{a}(x) = \sum_\alpha w_\alpha \cdot [f(x) - V^\alpha(x)]^2 + \sum_\alpha w_\alpha \left\{ [f(x) - \overline{V}(x)] \cdot [f(x) - \overline{V}(x) - 2f(x) + 2V^\alpha(x)] \right\}$$

Using the definition *(2)* for the first term, and the fact that $[\, f(x) - \bar{V}(x)]$ is independent of α so that we can bring it outside the summation:

$$\bar{a}(x) = \sum_{\alpha} w_{\alpha} \cdot \varepsilon^{\alpha}(x) + [\, f(x) - \bar{V}(x)] \cdot \sum_{\alpha} w_{\alpha} \{\, f(x) - \bar{V}(x) - 2 f(x) + 2V^{\alpha}(x) \,\}$$

Separating the second summation for each of the four terms, and using the fact that *f(x)* and $\bar{V}(x)$ are independent of α :

$$\bar{a}(x) = \sum_{\alpha} w_{\alpha} \cdot \varepsilon^{\alpha}(x) + \\ + [\, f(x) - \bar{V}(x)] \cdot \{\, f(x) \cdot \sum_{\alpha} w_{\alpha} - \bar{V}(x) \cdot \sum_{\alpha} w_{\alpha} - 2 f(x) \cdot \sum_{\alpha} w_{\alpha} + 2 \cdot \sum_{\alpha} w_{\alpha} V^{\alpha}(x) \,\}$$

Using the property of normalization of the coefficients, together with the definition *(1)* for the last of the four terms in the second summation:

$$\bar{a}(x) = \sum_{\alpha} w_{\alpha} \cdot \varepsilon^{\alpha}(x) + [\, f(x) - \bar{V}(x)] \cdot \{\, f(x) - \bar{V}(x) - 2 f(x) + 2\bar{V}(x) \,\}$$

Using simple algebraic properties:

$$\bar{a}(x) = \sum_{\alpha} w_{\alpha} \cdot \varepsilon^{\alpha}(x) - [\, f(x) - \bar{V}(x)]^2$$

Using the definition *(3)*:

$$\bar{a}(x) = \sum_{\alpha} w_{\alpha} \cdot \varepsilon^{\alpha}(x) - e(x)$$

If we now define the *Weighted Average Error* of the individual errors as:

$$(6) \qquad \bar{\varepsilon}(x) = \sum_{\alpha} w_{\alpha} \cdot \varepsilon^{\alpha}(x)$$

We obtain the formula of the error for the ensemble, on a <u>given input *x*</u>, that is what we were looking for:

$$i) \qquad e(x) = \bar{\varepsilon}(x) - \bar{a}(x)$$

3. The Ensemble-Error equation (over the input distribution)

A global version of the *i)*, i.e. a formula independent of the single input *x*, can be obtained by averaging over the input distribution.

To this purpose we will denote any global variable with the capital letter. We need then to define the variables as:

$$(7) \qquad E^{\alpha} = \int \varepsilon^{\alpha}(x)\, p(x)\, dx$$ - network α (global) Generaliza tion Error

$$(8) \qquad A^{\alpha} = \int a^{\alpha}(x)\, p(x)\, dx$$ - network α (global) Ambiguity

$$(9) \qquad E = \int e(x)\, p(x)\, dx$$ - (global) Ensemble Generaliza tion Error

And, following the same criteria that brought us to the definitions *(5)* and *(6)*:

$$(10) \qquad \overline{A} = \sum_{\alpha} w_{\alpha} \cdot A^{\alpha} = \sum_{\alpha} w_{\alpha} \cdot \int a^{\alpha}(x)\, p(x)\, dx = \int \sum_{\alpha} w_{\alpha} \cdot a^{\alpha}(x)\, p(x)\, dx = \int \overline{a}(x)\, p(x)\, dx$$

- Weighted Average of the Individual (global) Ambiguitie s

$$(11) \qquad \overline{E} = \sum_{\alpha} w_{\alpha} \cdot E^{\alpha} = \sum_{\alpha} w_{\alpha} \cdot \int \varepsilon^{\alpha}(x)\, p(x)\, dx = \int \sum_{\alpha} w_{\alpha} \cdot \varepsilon^{\alpha}(x)\, p(x)\, dx = \int \overline{\varepsilon}(x)\, p(x)\, dx$$

- Weighted Average of the individual (global) Generalization Errors

Then, simply integrating the equation *i)* on the domain in *x* we obtain the equation:

$$\int e(x)\, p(x)\, dx = \int \overline{\varepsilon}(x)\, p(x)\, dx - \int \overline{a}(x)\, p(x)\, dx$$

$$ii) \qquad E = \overline{E} - \overline{A}$$

This last equation gives a linear relation between the ensemble generalization error and the weighted averages of the individual generalization errors and the individual ambiguities. It is a simple and elegant equation that allow us to "separate" the generalization error of an ensemble into a term that depends only on the *single individual* errors and a term that depends only on the *correlation* between the networks of the ensemble (the Ambiguity is in fact a measure of the correlation). This last term can be estimated entirely from unlabeled data, with no further knowledge of the real function to be approximated.
The equation *ii)* will be the base for further studies on ensemble of networks.

Reference:

[1] Krogh, A., Vedelsby, J., Neural Network Ensembles, Cross Validation and Active Learning, in Advances in Neural Information Processing Systems 7, G. Tesauro, D. S. Touretsky, T. K. Leen, eds., pp231-238, MIT Press, Cambridge MA, 1995.