To appear: Computational Statistics and Data Analysis

INTERPRETING NEURAL-NETWORK RESULTS: A SIMULATION STUDY

Orna Intrator
Center for Gerontology and Health Care Research
Brown University
Orna_Intrator@Brown.Edu

Nathan Intrator
Computer Science Department
Tel-Aviv University
nin@tau.ac.il

February 2001

Abstract

Artificial Neural Networks (ANN) seem very promising for regression and classification, especially for large covariate spaces. Yet, their usefulness for medical and social research is limited because they present only prediction results and do not present features of the underlying process relating the inputs to the output. ANNs approximate a non-linear function by a composition of low dimensional ridge functions, and therefore appear to be less sensitive to the dimensionality of the covariate space. However, due to non uniqueness of a global minimum and the existence of (possibly) many local minima, the model revealed by the network is non-stable. We introduce a method that demonstrates the effects of inputs on output of ANNs by using novel robustification techniques. Simulated data from known models are used to demonstrate the interpretability results of the ANNs. Graphical tools are used for studying the interpretation results, and for detecting interactions between covariates. The effects of different regularization methods on the robustness of the interpretation are discussed; in particular we note that ANNs must include skip layer connections. An application to an ANN model predicting 5-year mortality following breast cancer diagnosis is presented. We conclude that neural networks estimated with sufficient regularization can be reliably interpreted using the method presented in this paper.

KEYWORDS: Logistic Regression; Nonlinear Models; Data mining tools; Interaction Effects; Splitlevel Plots.
RUNNING TITLE: Interpreting Neural Networks.

1 Introduction

Interpretability of statistical models, or the understanding of the way inputs relate to an output in a model, is a desirable property in applied research. For health outcome data, interpretation of model results becomes acutely important, as the intent of such studies is to gain knowledge about the underlying mechanisms. Interpretation is also used to validate findings as results that are inconsistent or contradictory to common understanding of issues involved may indicate problems with data or models. Commonly used models, such as the logistic regression model, are interpretable, but often do not provide adequate prediction, thus making their interpretation questionable. Statistical aspects of ANNs, such as approximation and convergence properties, have been discussed, and compared with properties of more "classical" methods (Barron and Barron, 1988; Geman et al., 1992; Ripley, 1993). ANNs have proven to produce good prediction results in classification and regression problems (e.g. Ripley, 1996). This has motivated the use of ANN on data that relates to health outcomes such as death or diagnosis. One such example is the use of ANN for the diagnosis of Acute Coronary Occlusion (Baxt, 1990). In such studies, the dependent variable of interest is a class label, and the set of possible explanatory predictor variables – the inputs to the ANN – may be binary or continuous.

Neural networks become useful in high dimensional regression by looking for low dimensional decompositions or projections (Barron, 1991). Feed-forward neural networks with simple architecture (one or two hidden layers) can approximate any L^2 function and its derivatives with any desired accuracy (Cybenko, 1989; Hornik et al., 1990; Hornik et al., 1993). These two properties of ANN make them natural candidates for modeling multivariate data.

The large flexibility provided by neural network models results in prediction with a relatively small bias, but a large variance. Careful methods for variance control (Barron, 1991; Breiman, 1996; Raviv and Intrator, 1996; Breiman, 1998; Intrator, 2000) can lead to a smaller prediction error and are required to robustify the prediction. While artificial neural networks have been extensively studied and used in classification and regression problems, their interpretability still remains vague. The aim of this paper is to present a method for interpreting ANN models.

Interpretability of common statistical models is usually done through an understanding of the effect of the independent variables on the prediction of the model. One approach to interpretation of ANN models is through the study of the effect of each input individually on each neuron in the network. We argue that a method for interpretation must combine the effect of the input on all units in the network. It should also allow for combining effects of different network architectures. Since substantial interest usually focuses on the effect of covariates on prediction, it is natural to study the derivative of the prediction p with respect to each predictor. For binary response outcomes it is natural to study the derivative of the log-odds ($\log \frac{p}{1-p}$) with respect to each input.

We calculate the derivative of the log odds of the ANN prediction with respect to each of the explanatory variables (inputs) while taking various measures for achieving robust results. The method allows to determine which variables have a linear effect, no effect, or nonlinear effect on the predictors. Graphical tools useful for identifying interactions, and for examination of the prediction results are presented. Using simulated data we demonstrate the importance of using different regularization methods on robustification and interpretation. Finally, we present an application of the method to study 5-year mortality from breast cancer.

A preliminary version of this paper appeared in (Intrator and Intrator, 1997).

2 Methods

2.1 Regularization of neural networks

The use of derivatives of the prediction with respect to the input data, sometimes called sensitivity analysis, is not new (Deif, 1986; Davis, 1989). Since a neural network model is parametric (with possibly a large parameter space), a discussion of the derivatives of the function is meaningful (Hornik et al., 1990, 1993). However, there are several factors which degrade the reliability of the interpretation that need to be addressed. First, the solution to a fixed ANN architecture and learning rule is not unique. In other words, for any given training set and any given model (architecture, i.e. the number of hidden units), the weight matrix is not uniquely determined. This means that ANN models are not identifiable. Second, gradient descent, which is often used for finding the estimates, may get stuck at local minima. This means that based on the random sequence in which the inputs are presented to the network and based on the initial values of the input parameters different solutions may be found. Third, there is the problem of optimal network architecture selection (number of hidden layers, number of hidden units, weight constraints, etc.) This problem can be addressed to some degree by cross validatory choice of architecture (Breiman, 1996; Breiman, 1998), or by averaging the predictors of several network with different architecture (Wolpert, 1992).

The non-identifiability of neural network solutions caused by the (possible) non uniqueness of a global minima, and the existence of (possibly) many local minima, leads to a large prediction variance. The large variance of each single network in the ensemble can be tempered with a regularization such as weight decay (Krogh and Hertz, 1992; Ripley, 1996, provide a review). Weight decay regularization imposes a constraint on the minimization of the squared prediction error of the form:

$$E = \sum_p |t_p - y_p|^2 + \gamma \sum_{i,j} w_{i,j}^2, \tag{1}$$

where t_p is the target (observation) and y_p the output (prediction) for the p'th example pattern. $w_{i,j}$ are the weights and γ is a parameter that controls the amount of weight decay regularization. There is some compelling empirical evidence for the importance of weight decay as a single network stabilizer (Breiman, 1996; Ripley, 1996; Breiman, 1998).

The success of ensemble averaging of neural networks is due to the fact that neural networks, in general, find many local minima; even with the same training set, different local minima are found when starting from different random initial conditions. These different local minima lead to somewhat independent predictors, and thus, the averaging can reduce the variance. (Hansen and Salamon, 1990; Wolpert, 1992; Perrone and Cooper, 1993; Breiman, 1996; Raviv and Intrator, 1996; Breiman, 1998; Intrator, 2000)

When a large set of independent networks is needed, but only little data is available, data reuse methods can be helpful. Re-sampling (with return) from the training data leads to partially independent training sets, and hence to improved ensemble results (Breiman, 1996; Breiman, 1998) . Smoothed bootstrap, which simulates the true noise in the data (Efron and Tibshirani, 1993), is potentially more useful since larger sets of independent training samples can be generated.

Noise added to the input during training can also be viewed as a regularizing parameter that controls, in conjunction with ensemble averaging, the capacity and the smoothness of the estimator (Raviv and Intrator, 1996; Bishop, 1995). Adding noise results in different estimators pushed to to different local minima, thus producing a more independent set of estimators. Best performance is then achieved by averaging over the estimators. For this regularization, the level of the noise may be larger than the 'true' level which can be indirectly estimated.

2.2 Interpretability of single hidden-layer Neural Networks

The most common feed-forward neural network for classification has the following form:

$$p = \sigma(\sum_{i=1}^{l} \lambda_i \sigma(x \cdot w_i)),$$

where l is the number of hidden units, σ is the sigmoidal function given by $\sigma(x) = 1/(1+exp(-x))$, x are the inputs (covariates) and w are the weights (parameter) attached to each neuron. The design of the input includes an intercept term (often called "bias" in Neural Network lingo) so that $x \cdot w_i \overset{\text{def}}{=} \sum_k x_k w_{ik} + w_{i0}$.

In terms of log odds, the common feed-forward network can be written as

$$\log(p/(1-p)) = \sum_{i=1}^{l} \lambda_i \sigma(x \cdot w_i).$$

This is a nonlinear model for the effect of the inputs on the log odds, as each projection $x \cdot w_i$ has a nonlinear effect on the output mediated through the sigmoidal function.

In a manner similar to the interpretation of logistic regression, we study the effect of an infinitesimal change in variable x_j on the logit transform of the probability. Since $p(x)$ is a smooth function of x, it is meaningful to examine the derivative

$$\frac{\partial}{\partial x_j} \log(p(x)/(1-p(x))) = \sum_{i=1}^{l} \lambda_i \sigma'(x \cdot w_i) w_{ij}.$$

In logistic regression, the effect of each covariate x_j on the log odds is given by the individual weights w_j since the odds are expressed as a linear combination of the inputs. The effect of each covariate x_j for the neural network model is given by what we term a *generalized weight*:

$$\tilde{w_j}(x) = \sum_{i=1}^{l} \lambda_i \sigma'(x \cdot w_i) w_{ij}, \tag{2}$$

Thus, in neural network modeling, the generalized weights have the same interpretation as weights have in logistic regression, i.e., the contribution to the log odds. However, unlike logistic regression, this contribution is local at each specific point x. The dependence of the generalized weights on the specific covariate levels attests to the nonlinearity of the ANN model, even with respect to the log-odds. For example, it is possible that the same variable may have a positive effect for some of the observations and a negative effect for others and its average effect may be close to zero. The distribution of the generalized weights, over all the data shows whether a certain variable has an overall strong effect, and determines if the effect is linear or not. A small variance of the distribution suggests that the effect is linear. A large variance suggests that the effect is nonlinear as it varies over the observation space. In contrast, in logistic regression the respective distribution is concentrated at one value.

A generalization of the common feed-forward neural network is one with skip layer connections. In this case the inputs are also directly connected with the outputs, so that the model is

$$p = \sigma(\sum_{i=1}^{l} \lambda_i \sigma(x \cdot w_i) + x \cdot \beta),$$

where the additional term permits the estimation of a simple logistic regression model. The definition of the generalized weights can easily be extended to include this model.

A robust generalized weight (RGW) for each input is an average of the generalized weights obtained from predictions of an ensemble of estimates. Each prediction is based on model parameter estimates. Model parameter estimates are obtained using some regularization methods as discussed in Section 2.1 with an alternating sequence of inputs.

Two types of plots are used to summarize the results. Scatter plots of the RGW of each variable with respect to its values provide a mean for examining the possibility of nonlinearity, although not necessarily detecting its form. A smoothed plot of the average effects at the neighborhood of each input level can indicate the nonlinear form. Split-level plots (available in Splus) are used to detect interactions. They present the RGWs of one variable on the y-axis with respect to either its levels or levels of another variable on the x-axis. The RGWs are averaged within quintiles of a variable other than the one presented on the x-axis in order to indicate interactions. In the figures presented in this paper, 5 lines are plotted, each corresponding to a quintile of information of the extra variable. For example, to test the possible interaction effects between variables X_1 and X_2 we may examine any or all of 4 plots: (1) the smoothed plots of X_1 and the RGW of X_1 at quintiles of X_2; (2) as in (1) but the RGW of X_2; (3) the smoothed plots of X_2 and the RGW of X_1 at quintiles of X_1; (4) as in (3) but the RGW of X_2.

2.3 Simulation studies

We simulated binomial data using the logit link function to assess the quality of interpretation. Of particular interest to us was the sensitivity of the interpretation to the regularization methods.

For two independent continuous covariates x_1 and x_2 that are uniformly distributed between 0 and 1 we simulated the following models:

1. A deterministic model: $I\{x_2 > 0\}$, where I is the indicator function;
2. $\text{logit}(p) = ax_1 + bx_2$ with $a = 1,\ b = 2$;
3. $\text{logit}(p) = ax_1x_2$ with $a = 1$;
4. $\text{logit}(p) = x_1^2 + x_2$.

Each simulation contained 800 data points and used ensembles of six hidden units single layer nets. Ripley's S-Plus 'nnet' implementation of a feed-forward network was used (Ripley, 1996) together with our implementation of the RGWs The minimization criterion was mean squared error with weight decay. We tested weight decay parameter values γ between 5e-5 and 0.5. Noise values added to the inputs were normally distributed with zero mean and standard deviation up to 20% of the standard deviation of the input. We used the *skip layer connections* option of Ripley's code (namely a model that includes logistic regression). Robustification of the generalized weights was based on network ensembles of 5 to 11 networks.

3 Results

In all figures, the left hand panel is a scatter plot of each individual observation's generalized weight at its observed data point. This is presented twice: the top figure is for x_1 and the bottom figure is for x_2. These plots present a rough picture of the RGWs and suggest nonlinearity as discussed above. A more detailed examination of the results are the quintile split-level plots of the RGWs (right panels) which are averages of the RGWs of one variable within quintiles of values of another variable, plotted at different levels of each of the variables.

Model 1: $I(x_2 > 0)$.

[Figure 1 about here]

This trivial model demonstrates a fundamental problem with model interpretation when the true link function is not finitely approximated well by a mixture of sigmoidal functions (Figure 1). We used a step link function which corresponds to an infinite slope of the sigmoidal. The anticipated RGW is null everywhere except at zero where it is infinite. The stronger the weight decay, the smoother the RGW, with a tamer effect at zero, and a slower decay to zero elsewhere. Weight decay reduced the level of the RGW at zero from those in the order of hundreds at $\gamma < 1e - 5$ to less than 10 at $\gamma = 0.1$.

Model 2: $\text{logit}(p) = ax_1 + bx_2$ where $a = 1, \ b = 2$.

[Figure 2 about here]

In this model we expect the interpretation to be a constant function fixed at 1 for the derivative of the logit with respect to x_1, and a constant function fixed at 2 for the derivative with respect to x_2. In the scatter plots in Figure 2 we see that the estimates are scattered around their true values. The split-level plots of RGWs of each variable versus the other should be parallel at values specified by the levels, while those of each variable with respect to its values should be constant at all values of the levels. The results are from 15 averages of networks with 7 units, noise injection at 0.03, weight decay at 0.02, with skip layer connection and 9 cross-validation sets. Both the network and linear logistic regression models had cross validated average prediction error of 7.62%. The ROC of the network is 0.72 and that of a logistic regression model is 0.78. The fact that the network, as an approximator, did as well as the true linear logistic regression model is encouraging.

[Figure 3 about here]

Figure 3 presents the results of a neural network model with no skip layer connection (the most common feed-forward architecture). The scale of the generalized weights is around 0.1 (10% of the true model). The variability of the generalized weights (in the range of 0.3) may incorrectly indicate a nonlinear model. We see that without the skip layer connections, the neural network architecture is unable to correctly approximate a simple logistic regression model.

Model 3: $\text{logit}(p) = ax_1x_2$ where $a = 1$

[Figure 4 about here]

In this model we expect to see parallel split-level plots of the RGWs of x_1 by x_1 (since the derivative is ax_2), and parallel split-level plots for the RGWs of x_2 by x_2 (since the derivative is ax_1). The split-level plots of the RGW of x_2 versus x_1 are expected to be a single increasing line, with no difference between the quintile split-level plots. Likewise, the split-level plots of the RGW of x_1 versus x_2.

Figure 4 depicts interpretation result using moderate regularization: weight decay ($\gamma = 0.05$), no noise injection and no averaging. We first note that the scale of the result is between -10 and 10, and the slope in the lower panels is around 5, way beyond the model parameter $a = 1$. We see that the split-level plots of generalized weight of x_1 by x_1 (and those of x_2) are not always parallel, and are not evenly spaced. They exhibit heavy shrinkage at the data points with large absolute values.

[Figure 5 about here]

The regularization needed to produce robust plots involves a large weight decay ($\gamma = 0.5$) a high level of noise (at least 0.3 SD), and averaging. Figure 5 presents the dependence on the level of noise. The effect of noise injection (with ensemble averaging) on robustifying the results is clearly demonstrated.

[Figure 6 about here]

Figure 6 presents the robustified generalized weights with satisfactory levels of regularization. As expected, both variables exhibit a nonlinear effect in the left panels. The split-level plots aid in detecting that the nonlinearity is due to an interaction between the two variables. The interaction is apparent from the middle panel which displays the (almost) single straight line corresponding to x_1 and its RGW, at the 5 quintile split-levels of x_2, and from the parallel horizontal lines of the RGWs of x_1 versus x_2, at the various quintiles of x_1. The boundaries display some deviations from the expected values.

Model 4: $\text{logit}(p) = x_1^2 + x_2$

[Figure 7 about here]

For this model we restricted the inputs to the range [-1.5,1.5] due to the quadratic effect of x_1. We expect to detect the quadratic effect of x_1 when examining the RGW of x_1 versus its levels, which should be linearly increasing from -3 at $x_1 = -1.5$ to 3 at $x_1 = 1.5$. We also expect to see a single straight line of the RGW of x_1 versus its levels, at all quintiles of x_2. We expected to see the RGW of x_1 versus x_2 as parallel lines of twice the average levels of the quintiles of x_1. The RGW of x_2 should always be 1.

We used a 7 hidden unit neural network architecture with noise injection (Raviv and Intrator, 1996), and averaged over an ensemble of 15 networks. We observed little sensitivity to network size from 5 to 15 hidden units. The results for x_1^2 are quite good in the range [-1,1], but for $x_1 < -1$, the results are misleading. Possibly, the weight decay parameter should have been smaller. On the other hand weight decay for x_2 should be increased. The interpretation of x_2 is somewhat weaker, with estimates in the range of 0.8 to 1.2. For comparison, the simple linear logistic regression model was $\text{logit}(p) = 0.70 - 0.02x_1 + 0.83x_2$, with insignificant parameter estimate for x_1, and a strongly significant effect for x_2. The linear logistic model nullified the effect of x_1. The estimate of x_2 is also slightly biased downwards.

The prediction of the ensemble of networks has a 26.4% 9-fold cross-validated error rate with ROC value of 0.73. The linear logistic regression gave a 9-fold cross validation error of 32.1% with ROC value of 0.69. The cross validated results indicate that the neural network model had better prediction. A similar test performed on a new test data set gave a 26.7% error rate for neural network with ROC=0.75, and a 30.3% error rate for the linear logistic regression model with ROC=0.69, showing that the cross validated estimates were not biased.

4 Five-Year Breast Cancer Mortality

The data for this example come from six breast cancer studies conducted by the Eastern Cooperative Oncology Group and was kindly provided by Robert Grey. This data has been analyzed using survival analysis methods in Gray (1992), Kooperberg et al. (1992), and Intrator and Kooperberg (1995). All patients had disease involvement in their axillary lymph nodes at diagnosis indicating some likelihood that the cancer had spread through the lymphatic system to other parts of the body; however, none of the patients had evidence of disease at the time of entry into the study, which was following surgical removal of the primary tumor and axillary metasteses. In this paper we examine 5 year mortality from breast cancer.

Two thousand, four hundred and four breast cancer patients were available from the six studies, and were followed for upto 12 years; of them 467 patients were censored before 5 years, and were omitted from analyses because their survival status at the end of the first 5 years could not be determined. This left 970 patients who survived. and 967 patients who died within 5 years following diagnosis (mortality rate of 49.9%).

There were six covariates: (1) estrogen receptor status (ER 0 is 'negative', 1 is 'positive') with 63.0% of the patients with positive ER; (2) the number of positive axillary lymph nodes at diagnosis, which varied between 1-35 with an average of 5.6 (std=5.6), and a median of 4 nodes; (3) size of the primary tumor, which varied between 3-100mm, with an average of 31.9mm (std=16.7) and median 30mm; (4) age at study admission, between 22-81 with an average 52 (std=12); (5) menopausal status, with 51.5% of patients post-menopausal; (6) body mass index (BMI: defined as weight/height^2), with a mean 26.2kg/m^2 (std=5.1) and median 25kg/m^2. Although the empirical distribution of the number of nodes is highly skewed to the right, we did not transform it, and allowed the neural network model to adjust itself.

We modeled the data using 6-inputs single-output neural networks with 6-7 hidden units (i.e. 36-42 weights and 6 linear weights), weight decay parameter of either 1e-5 or 1e-4. For robustification we injected noise which was normally distributed with 0 mean and 0.5 variance, and used ensemble averaging over 8 networks. These parameters led to the best cross-validated ROC values. We computed an estimate of the error of each individual's RGW, of each covariate, by computing the standard deviation of the 8 results from each of the networks (for each patient's RGWs). Cross-validated prediction results showed that the ROC for the neural network models ranged between 0.649-0.673, and for the linear logistic regression model between 0.624-0.632. This results suggests that the neural networks (a) outperformed the linear logistic regression model by between 2.5-7.5%; and (b) that the parameters used for the ANN models were adequate, so the interpretation results would be valid.

Table 1 presents estimators from a linear logistic regression model and averages of RGW over all patients with their sample standard deviation, and the estimation error for each covariate. The neural network results for all the continuous variables are very similar to the coefficients of the logistic regression models for these variables. This suggests that the linear logistic regression model is appropriate as a first approximation. However, the standard deviations of the RGWs, especially for nodes, were quite large, suggesting nonlinear and/or interaction effects. The RGWs for the binary variables were much smaller on average, suggesting that the ANN model was different from the linear logistic model, certainly with respect to the effects of these covariates.

Table 1 about here.

We concentrate on the effect of the number of nodes and its interactions. One way to examine b

	Beta	SE†	Signif	Mean	Std	SE ††
Estrogen	-.780	.390	**	-.106	.041	.013
Nodes	.115	.001	**	.127	.050	.013
Size	.015	.003	**	.011	.006	.002
Age	-.019	.007	*	-.004	.004	.002
BMI	.023	.010		.017	.007	.003
Menopause	.705	.165	**	.076	.031	.016

† Standard error of parameter estimate from logistic regression.

†† Average over all observations of standard deviation of RGW computed over 8 networks.

Table 1: Parameter estimates from linear logistic regression model and from average RGWs of neural network model.

non-linearity of predictors is to plot the smoothed generalized additive model (GAM) functions of a covariate (Hastie and Tibshirani, 1990). GAM provides a semi-parametric smoothed estimate of an additive generalized linear model that allows no interaction effects. For binary response the smoothed estimates are of the log-odds model. Figure 8 presents the GAM of the effect of the number of nodes with 95% standard error bands, and the RGW of nodes by the number of nodes, overlaid with a smoothed graph. The RGWs are the derivative (with respect to nodes) of the effect function, therefore, in shape, they are not completely comparable to the GAM model, which is the specific functional form. Likewise, the scale of the estimates is not comparable. According to the GAM model, the effect of nodes increases linearily from -0.8 for 1 node to approximately 1.2 for 17 nodes, and then levels off. The derivative of this function would therefore decrease and have an inflection point at 17 nodes. The RGWs appears to increase from 0.14 at a single node to approximately 0.17 at 5 nodes, and then decrease monotonically to a nill effect with no clear inflection point at 17 nodes. The average RGW is 0.127, quite similar to the estimated linear logistic model effect (0.115), suggesting that the linear approximation may be appropriate.

[Figure 8 about here]

Studying interaction effects in logistic regression is only done by testing the inclusion of specific interactions. While we are assured that the ANN model incorporates interactions, as necessary, we would like to be able to detect and interpret them. To do so, we examined the split-level plots of nodes and size (an interaction that was reported by Intrator and Kooperberg, 1995), and also whether there is a possible interaction between age and nodes, since those two variables displayed the most non-linearity. Figure 9 displays the split-level plots of nodes and size, and nodes and age. We present only those plots with nodes on the x-axis since the effect of number of nodes is, by far, the largest. It appears that for less nodes (up to about 15) the RGWs of nodes are smaller for larger tumor sizes, and possibly likewise for the RGWs of size. This suggests that the log odds might be modeled as c nodes$^3 + d$ size (nodes $-$ 15)$-$, where $(x)- = \{0 \; x > 0x$ otherwise . The split-level plots by age suggest that for upto 15 nodes, the RGWs of nodes change from convex for younger ages to concave for older ages.

[Figure 9 about here]

5 Discussion

Neural networks have been considered “black boxes” and “data mining tools”. Their acceptability as valid methods for medical and social research requires that beyond providing better predictions they provide meaningful results that can be understood by clinicians, policy makers, intervention planners, academicians and lay persons. This paper achieves this end by presenting a method for interpreting results of neural network models. The results are meaningful, and as seen on a real-world application, for the most part, agree with other results, and provide additional insights into the underlying processes that generate the data.

Using simulated data we demonstrated that the method provides an appropriate model description. The simulation studies demonstrated that the popular feed forward architecture which does not include skip layer connections is limited and cannot adequately represent or extend a linear model. An important contribution of this method is its ability to directly identify multiplicative interactions. Since neural networks provide estimation for general approximations, there is no need to specifically model interactions. What one needs instead is a graphical method to examine and detect them. Such a method is provided in this paper: a graphical examination of the split-level plots of the generalized weights averaged over quintiles of the data by the values of the inputs.

We stress that the interpretation results rely heavily on appropriate use of neural network regularization and that the usage of skip layer architecture is essential. Furthermore, weight decay and noise injection along with ensemble averaging, should be applied. These “tweaking” parameters are also important in order to obtain better (cross validated) prediction results. Thus, cross validated prediction should direct a better choice of these regularization parameters, which would lead to valid interpretation. When these methods are not appropriately used, one may easily arrive at false model interpretation.

An example of the effect on interpretation of the nonlinearity of the RGWs concerns their shrinkage at larger values of the inputs. Weight decay (Equation 1) results in shrunken parameter estimates, i.e., weights w_{ij} smaller than the true model parameters. These smaller weights propagate to the RGWs (Equation 2) in a non-linear way, emphasized at higher levels of the RGWs. This is seen in the simulations as a stronger reduction of effect size in the tails, where the absolute value of the RGWs are highest.

The analysis of the breast cancer data presents the complexity of the interpretation of the neural network model. Using the interpretability methods developed in this paper we were able to see that the ANN model identified a model that was different in some respects from both the linear logistic regression model and the semiparametric generalized additive model. Both the ANN and GAM models indicated that the effects of menopausal status and estrogen receptor status were not as strong as those detected by the linear logistic regression model. The ANN indicated a nonlinear effect of the number of nodes that was somewhat different from that identified by GAM. Since the prediction of the ANN model was significantly better than that of the linear logistic regression model, these findings ring out a word of caution to the ready acceptability of the linear logistic regression results. Formerly, Intrator and Kooperberg (1995) had shown in their survival trees analysis of the data that the effects of estrogen receptor status and menopause were not strong. Moreover, survival trees identified that the root split was at 4-6 nodes, and that the interaction with size was discernible for small number of nodes. Both results correspond well with the results of the ANN model. However, the large effects of menopause and ER status indicated in the linear logistic regression model are also observed in the hazard regression analysis (Intrator and Kooperberg, 1995). This suggests that there may be more than a unique model for these data. It is possible that there is an unobserved or unrecorded parameter in the women that predisposes some women towards one model, and others towards the second model.

The interpretation method presented here produces unbiased estimates of the underlying model parameters. An initial exploration of the validity of the RGW estimates was presented in the application to the breast cancer data by means of the estimates of the dispersion of the RGW's between ANN models. It is imperative that we begin to examine inferential methods for testing hypotheses regarding the effect of inputs.

References

Barron, A. R., Complexity regularization with application to artificial neural networks, in: Roussas, G., editor, *Nonparametric Functional Estimation and Related Topics*, (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991) 561–576.

Barron, A. R. and Barron, R. L., Statistical learning networks: A unifying view, in: Wegman, E., editor, *Computing Science and Statistics: Proc. 20th Symp. Interface*, (American Statistical Association, Washington, DC, 1988) 192–203.

Baxt, W. G., Use of an artificial neural network for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion. *Neural Computation*, 2(4) (1990) 480–489.

Bishop, C. M., Training with Noise is Equivalent to Tikhonov Regularization. *Neural Computation*, 7(1) (1995) 108–116.

Breiman, L., Bagging predictors. *Machine Learning*, 24 (1996) 123–140.

Breiman, L., Arcing classifiers. *The Annals of Statistics*, 26(3), (1998) 801–849.

Cybenko, G., Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2 (1989) 303–314.

Davis, G. W., Sensitivity analysis in neural net solutions. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(5) (1989) 1078–1082.

Deif, A. S., *Sensitivity Analysis in Linear Systems.* (Springer Verlag, Berlin-Heidelberg-New York, 1986).

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

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

Gray, R. J., Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87 (1992) 942–951.

Hansen, L. K. and Salamon, P., Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intellignce*, 12(10) (1990) 993–1001.

Hastie, T. and Tibshirani, R., *Generalized Additive Models* (Chapman and Hall, London, 1990).

Hornik, K., Stinchcombe, M., and White, H., Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3 (1990) 551–560.

Hornik, K., Stinchcombe, M., White, H., and Auer, P., Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Mimeo. (Discussion paper 93-15, Department of Economics, UCSD, 1993).

Intrator, N., Robust prediction in many parameter models: Specific control of variance and bias, in: Kay J. W. and Titterington D. M., editors, *Statistics and Neural Networks: Advances at the Interface* (Oxford University Press, 2000) 97–128.

Intrator, O. and Intrator, N., Robust interpretation of neural-network models, in: *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, (Florida, 1997) 255–264.

Intrator, O. and Kooperberg, C. L., Trees and splines in survival analysis. *Statistical Methods in Medical Research*, 4(3) (1995) 237–261.

Krogh, A. and Hertz, J. A., A simple weight decay can improve generalization, in: Moody, J., Hanson, S., and Lippmann, R., editors, *Advances in Neural Information Processing Systems*, Vol 4 (Morgan Kaufmann, San Mateo, CA, 1992) 950–957.

Perrone, M. P. and Cooper, L. N., When networks disagree: Ensemble method for neural networks, in: Mammone, R. J., editor, *Neural Networks for Speech and Image processing.* (Chapman-Hall, 1993).

Raviv, Y. and Intrator, N., Bootstrapping with noise: An effective regularization technique. *Connection Science, Special issue on Combining Estimators*, 8 (1996) 356–372.

Ripley, B. D., Statistical aspects of neural networks, in: Barndorff-Nielsen, O., Jensen, J., and Kendall, W., editors, *Networks and Chaos – Statistical and Probabilistic Aspects.* (Chapman and Hall, 1993).

Ripley, B. D., *Pattern Recognition and Neural Networks.* (Oxford Press, 1996).

Wolpert, D. H., Stacked generalization. *Neural Networks*, 5 (1992) 241–259.

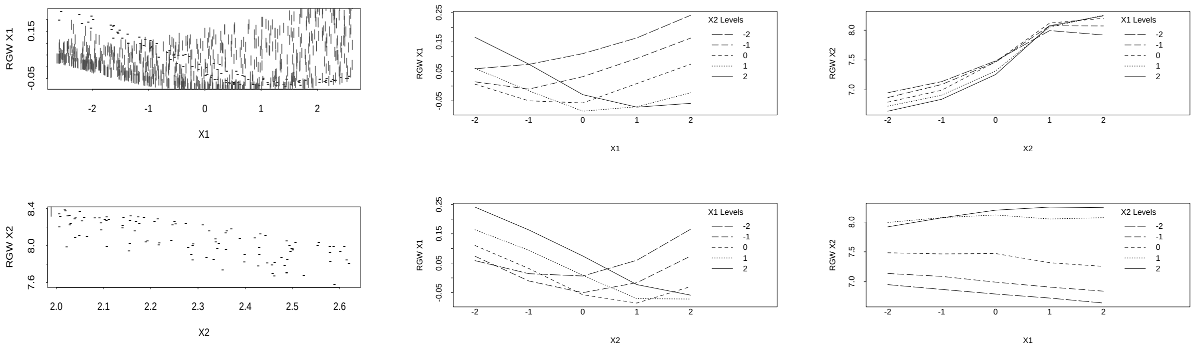

Figure 1: Interpretation of model (1). Strong weight decay tames the effect at zero, and does not decay rapidly to zero elsewhere. Notice that there is no effect of x_1.

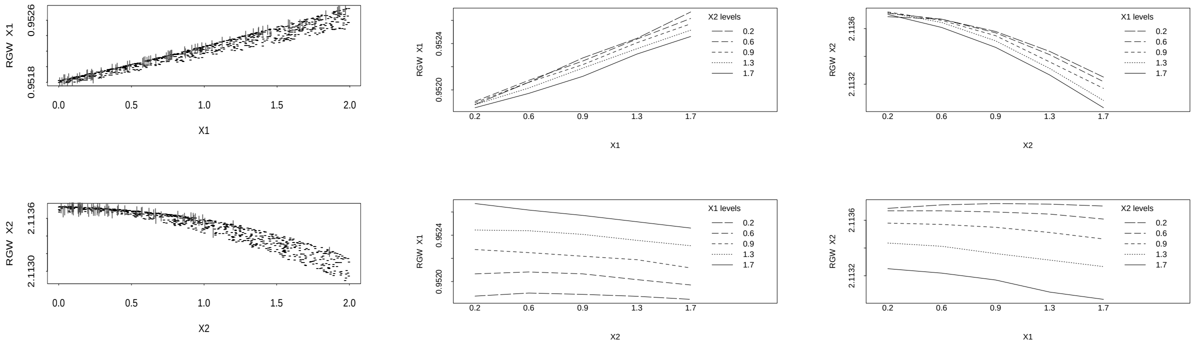

Figure 2: Interpretation of Model (2); Simple linear model gives a correct effects of the covariates when using skip layer connection architecture.

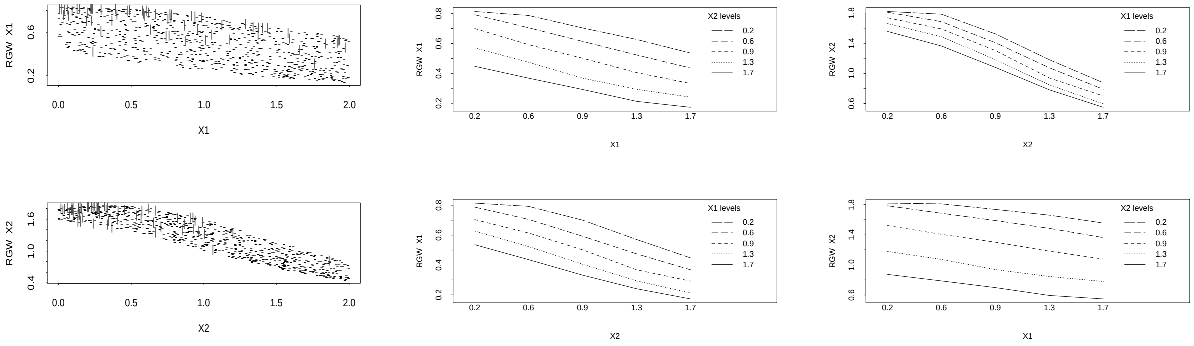

Figure 3: Interpretation of Model (2); Network with no skip layer connections. The effects are scaled wrong (around 0.1 and not around 1.0). The variability of the effect is increased to levels which might incorrectly indicate a nonlinear model.

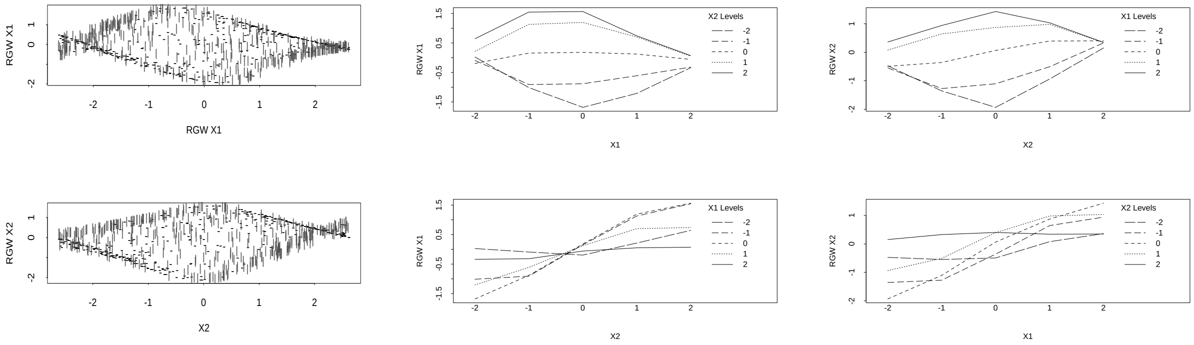

Figure 4: Model (3): Interpretation of interaction. Little robustification: small weight decay, no averaging of networks and no noise.

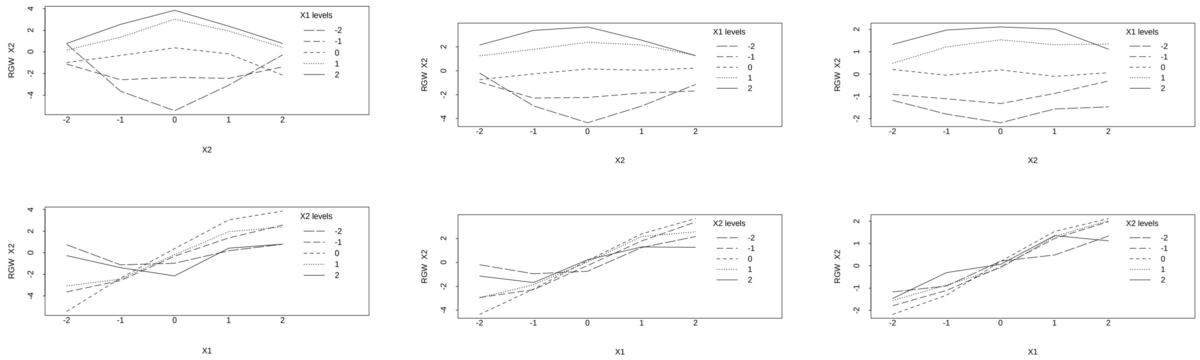

Figure 5: Model (3): Interpretation of interaction. Left: zero input noise, middle: input noise at 10% of the input SD, right: input noise at 20% of the input SD.

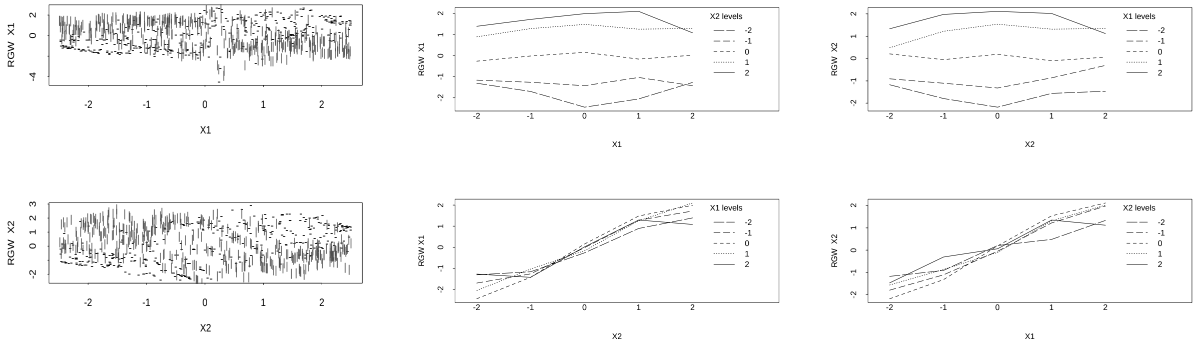

Figure 6: Model (3): Sufficient regularization for optimal prediction via weight decay, noise injection and ensemble averaging leads to more reliable interpretation. (Compare with Figure 4.)

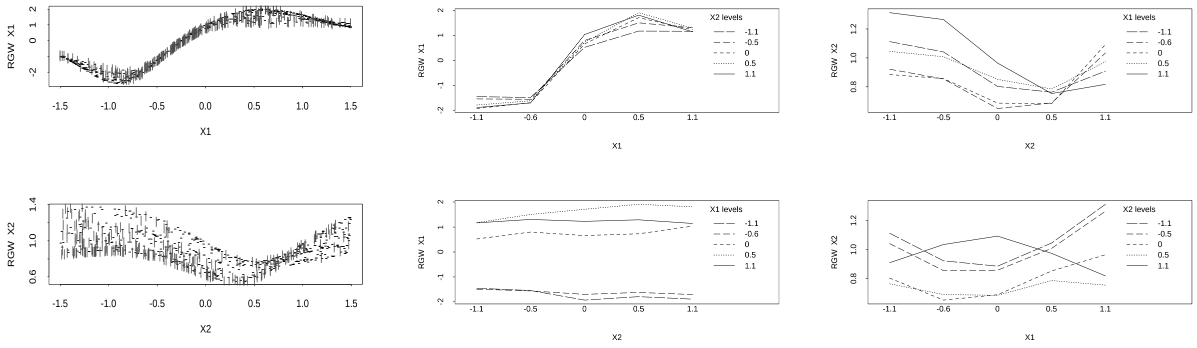

Figure 7: Model (4): Interpretation of $x_1^2 + x_2$. The nonlinear part of the model, x_1^2 appears to be more reliably interpreted than the linear part x_2.

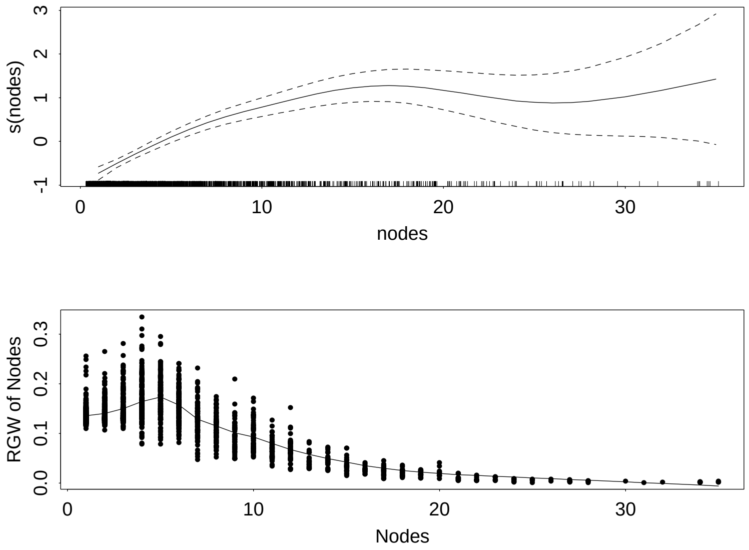

Figure 8: Effect of number of nodes on log-odds 5-year mortality. Generalized additive model with 95% confidence bands (top panel), where logit(p)=s(snodes)+f(other covariates). ANN model RGW, where $\frac{\partial \mathrm{logit(p)}}{\partial \mathrm{nodes{=}RGW\ of\ Nodes}}$, overlaid with smoothed curve (bottom panel).

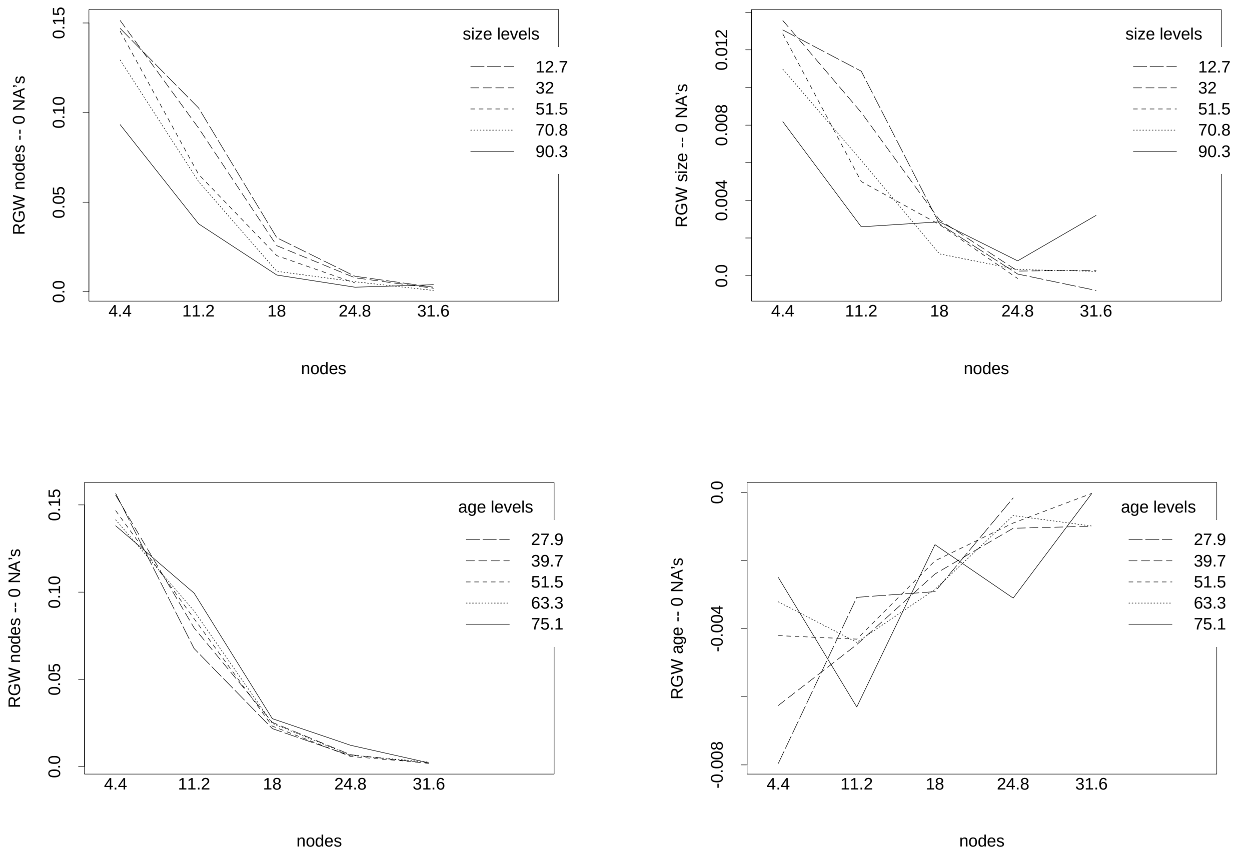

Figure 9: Split-level plots exploring interaction effect of nodes with size and age.