Complementary log-log and probit: activation functions implemented in artificial neural networks

Gecynalda Gomes and Teresa Bernarda Ludermir

May 3, 2009

Contents

1 Introduction
- Introduction

2 Complementary log-log and probit functions
- New activation functions

3 Experimental results
- Results

4 Conclusions
- Conclusions

5 Main references
- Main references

Introduction

- Artificial neural networks (ANN) may be used as an alternative method to binomial regression models for binary response modelling.
- The binomial regression model is a special case of an important family of statistical models, namely Generalized Linear Models (GLM) (Nelder and Wedderburn, 1972).
- Briefly outlined, a GLM is described by distinguishing three elements of the model: the random component, the systematic component and the link between the random and systematic components, known as the link function.

- The definition of the neural network architecture includes the selection of the number of nodes in each layer and the number and type of interconnections.
- The number of input nodes is one of the easiest parameters to select; the independent variables have been preprocessed because each independent variable is represented by its own input.
- The majority of current neural network models use the logit activation function, but the hyperbolic tangent and linear activation functions have also been used.

- However, a number of different types of functions have been proposed. Hartman *et al.* (1990) proposed *gaussian bars* as a activation function. *Rational transfer functions* were used by Leung and Haykin (1993) with very good results. Singh and Chandra (2003) proposed a class of sigmoidal functions that were shown to satisfy the requirements of the universal approximation theorem (UAT).

- The choice of transfer functions may strongly influence complexity and performance of neural networks.

- Our main goal is broaden the range of activation functions for neural network modelling. Here, the nonlinear functions implemented are the inverse of the complementary log-log and probit link functions.

Contents

1. Introduction
   - Introduction
2. Complementary log-log and probit functions
   - New activation functions
3. Experimental results
   - Results
4. Conclusions
   - Conclusions
5. Main references
   - Main references

New activation functions

- The aim of our work is to implement sigmoid functions commonly used in statistical regression models in the processing units of neural networks and evaluate the prediction performance of neural networks.

- The binomial distribution belongs to exponential family.

- The functions used are the inverse functions of the following link functions.

Type	η
logit	$\log[\pi/(1-\pi)]$
probit	$\Phi^{-1}(\pi)$
complementary log-log	$\log[-\log(1-\pi)]$

New activation functions

- We use multilayer perceptron (MLP) networks. The calculations made for the outputs $y_i(t) = \phi_i(\mathrm{w}_i^\top(t)\mathrm{x}(t)), \quad i = 1, \ldots, q$, such that w_i is the weight vector associated with the node i, $\mathrm{x}(t)$ is the attribute vector and q is the number of nodes in the hidden layer.

- The activation function ϕ is given by one of the following forms:

$$\phi_i(u_i(t)) = 1 - \{\exp[-\exp(u_i(t))]\}, \tag{1}$$

$$\phi_i(u_i(t)) = \Phi(u_i(t)) = 1/\sqrt{2\pi}\int_{-\infty}^{u_i(t)} e^{-u_i(t)^2/2}du_i(t), \tag{2}$$

- The derivatives form of the complementary log-log and probit are, respectively,

$$\phi_i'(u_i(t)) = -\exp(u_i(t)) \cdot \exp\{-\exp(u_i(t))\} \tag{3}$$

$$\phi_i'(u_i(t)) = \{\exp(-u_i(t)^2/2)\}/\sqrt{2\pi} \tag{4}$$

- The complementary log-log and probit functions are nonconstant, bounded and monotonically increasing. As funções complemento log-log e probit são não-constantes, limitadas e monotonicamente crescentes.
- Thus, those functions are sigmoidal functions with the requisite properties (UAT) for being an activation functions.

Contents

1. Introduction
   - Introduction
2. Complementary log-log and probit functions
   - New activation functions
3. Experimental results
   - Results
4. Conclusions
   - Conclusions
5. Main references
   - Main references

Main results

- The evaluation of the implementation of the new activation functions is based on the framework of a Monte Carlo experiment.
- At the end of the experiments, average and standard deviation were calculated for the mean square error (MSE) in the framework of a Monte Carlo experiment with 1,000 replications.
- To evaluate the functions implemented and evaluate their performance as universal approximators, we generate p input variables for the neural network from a uniform distribution after generating values for the response variable based on the function

$$y^* = \phi_k(\sum_{i=0}^{q} m_{ki}\phi_i(\sum_{j=0}^{p} w_{ij}x_j)),$$

- in which m_{0i} and w_{0i} denote, respectively, the weights of the connections between the bias and the output and between the bias and hidden nodes.
- In the generation of y^*, we use the inverse functions of the logit, complementary log-log and probit link functions as activation function, ϕ.
- The activation functions used in the generation are cited as “Reference LOGIT”, “Reference CLOGLOG” and “Reference PROBIT”.
- The simulated data were fitted with different activation functions: logit, hyperbolic tangent (hyptan), complementary log-log (cloglog) and probit.

- We conduct experiments for data generating processes varying sample sizes, $n = \{50, 100, 200\}$, number of input nodes, $p = \{2, 10, 25\}$, number of hidden nodes, $q = \{1, 2, 5\}$ and learning rate, $\nu = \{0.4, 0.6, 0.8\}$, for each function.
- These parameters were arbitrarily chosen. The training lengths ranging from 100 to 5,000 iterations until the network converges.
- For each data generating process, the data set was divided into two sets – 75% of the set for training and 25% for testing.
- Three different configurations were chosen to illustrate the results (CASE 1: $n = 50$, $p = 2$, $\nu = 0.4$, CASE 2: $n = 100$, $p = 10$, $\nu = 0.6$ and CASE 3: $n = 200$, $p = 25$, $\nu = 0.8$).

- Significance of the differences between the average MSE in the framework of a Monte Carlo experiment was tested using the Student's *t*-test for independent samples and a 5% significance level was adopted.
- In the Tables presents the *P*-values.
- For example, the cell "Cloglog-Logit" in reference CLOGLOG indicates comparison of the performance of the network with the complementary log-log activation function to the performance of the network with the logit activation function.
- The symbol "$<$" indicates that the average MSE of the complementary log-log function is smaller than the average MSE of the logit function. The absence of the symbols "$<$" and "$>$" implies that there is no difference between the average MSE of these functions.

- In the CASE 1, for the LOGIT reference with $q = 1$ there is no statistically significant difference (SSD) between the average MSE of the functions.
- For $q = 2$ and $q = 5$, there is a SSD between the average MSE of the functions in the majority of cases.
- For the CLOGLOG reference, there is a SSD between the average MSE of the functions in all cases when the activation function used is the complementary log-log.
- For the PROBIT reference, there is a SSD between the average MSE of the functions in the majority of cases when the activation function used is the probit.

Table: Results of the *P*-values of the differences between the average of the MSE of the MLP networks with different activation functions, 50 exemplars, input nodes $p = 2$, learnig rate $\nu = 0.4$ and number nodes of hidden layer $q = \{1, 2, 5\}$.

Reference LOGIT			
Comparation	$q = 1$	$q = 2$	$q = 5$
Logit-Hyptan	0.8773	$0.0000^{<}$	$0.0000^{<}$
Logit-Gauss	0.6112	$0.0000^{<}$	$0.0000^{<}$
Logit-Cloglog	0.6213	$0.0000^{<}$	$0.0000^{<}$
Logit-Probit	0.0592	$0.0000^{<}$	$0.0000^{<}$
Hyptan-Cloglog	0.7562	$0.0000^{>}$	$0.0002^{>}$
Hyptan-Probit	0.1049	$0.0000^{>}$	$0.0000^{>}$
Gauss-Cloglog	0.9585	$0.0000^{<}$	0.6656
Gauss-Probit	0.2056	$0.0000^{>}$	$0.0000^{>}$
Cloglog-Probit	0.1469	$0.0000^{>}$	$0.0000^{>}$

Results

Reference CLOGLOG			
Comparation	$q = 1$	$q = 2$	$q = 5$
Logit-Hyptan	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$
Logit-Gauss	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Logit-Cloglog	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Logit-Probit	$0.0000^{>}$	0.8462	0.7167
Hyptan-Cloglog	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Hyptan-Probit	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Gauss-Cloglog	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Gauss-Probit	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$
Cloglog-Probit	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$

Results

Reference PROBIT			
Comparation	$q = 1$	$q = 2$	$q = 5$
Logit-Hyptan	0.1225	$0.0000^{<}$	$0.0000^{<}$
Logit-Gauss	0.1843	$0.0000^{<}$	$0.0000^{>}$
Logit-Cloglog	0.9825	$0.0000^{<}$	$0.0000^{<}$
Logit-Probit	$0.0196^{>}$	$0.0000^{>}$	$0.0000^{>}$
Hyptan-Cloglog	0.1835	$0.0000^{>}$	$0.0000^{>}$
Hyptan-Probit	$0.0412^{>}$	$0.0000^{>}$	$0.0000^{>}$
Gauss-Cloglog	0.2449	$0.0000^{<}$	$0.0000^{<}$
Gauss-Probit	0.1574	$0.0000^{>}$	$0.0000^{>}$
Cloglog-Probit	$0.0450^{>}$	$0.0000^{>}$	$0.0000^{>}$

- In the CASE 2, for the LOGIT reference regardless of the number of hidden nodes, there is a SSD between the average MSE of the functions in all cases when the activation function used is logit.
- For the CLOGLOG and PROBIT references, there is a SSD between the average MSE of the functions in the majority of cases when the activation function used is the complementary log-log and probit.

Table: Results of the P-values of the differences between the average of the MSE of the MLP networks with different activation functions, 100 exemplars, input nodes $p = 10$, learnig rate $\nu = 0.6$ and number nodes of hidden layer $q = \{1, 2, 5\}$.

Reference LOGIT			
Comparation	$q = 1$	$q = 2$	$q = 5$
Logit-Hyptan	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$
Logit-Gauss	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$
Logit-Cloglog	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$
Logit-Probit	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$
Hyptan-Cloglog	$0.0009^{<}$	$0.0000^{<}$	$0.0000^{>}$
Hyptan-Probit	$0.0000^{>}$	$0.0000^{<}$	$0.0000^{>}$
Gauss-Cloglog	$0.0033^{<}$	$0.0000^{>}$	$0.0010^{<}$
Gauss-Probit	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Cloglog-Probit	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$

Results

Reference CLOGLOG			
Comparation	$q = 1$	$q = 2$	$q = 5$
Logit-Hyptan	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$
Logit-Gauss	0.4069	$0.0000^{>}$	$0.0000^{>}$
Logit-Cloglog	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Logit-Probit	$0.0000^{>}$	0.9961	$0.0000^{<}$
Hyptan-Cloglog	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Hyptan-Probit	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Gauss-Cloglog	0.3010	$0.0000^{>}$	$0.0000^{>}$
Gauss-Probit	0.3341	$0.0000^{<}$	$0.0000^{<}$
Cloglog-Probit	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$

Results

Reference PROBIT			
Comparation	$q = 1$	$q = 2$	$q = 5$
Logit-Hyptan	$0.0000^{<}$	$0.0000^{<}$	0.1233
Logit-Gauss	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$
Logit-Cloglog	$0.0000^{<}$	$0.0000^{<}$	$0.0000^{<}$
Logit-Probit	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Hyptan-Cloglog	$0.0000^{>}$	$0.0000^{>}$	0.1415
Hyptan-Probit	$0.0000^{>}$	$0.0000^{>}$	0.1228
Gauss-Cloglog	$0.0000^{<}$	$0.0000^{>}$	$0.0000^{<}$
Gauss-Probit	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Cloglog-Probit	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$

- In the CASE 3, for the LOGIT reference with $q = 1$, there is a SSD between the average MSE of the functions in the majority of cases, although the activation function used is the probit.
- In the MLP networks with $q = 2$ and $q = 5$, there is a SSD between the average MSE of the functions in all cases when the activation function used is the logit.
- For the CLOGLOG reference, there is a SSD between the average MSE of the functions in all cases when the activation function used in the MLP network is the complementary log-log.
- For the PROBIT reference, there is a SSD between the average MSE of the functions in the majority of cases, when the activation function used is the probit.

Table: Results of the *P*-values of the differences between the average of the MSE of the MLP networks with different activation functions, 200 exemplars, input nodes $p = 25$, learnig rate $\nu = 0.8$ and number nodes of hidden layer $q = \{1, 2, 5\}$.

Reference LOGIT			
Comparation	$q = 1$	$q = 2$	$q = 5$
Logit-Hyptan	0.3233	$0.0000^{<}$	$0.0000^{<}$
Logit-Gauss	0.7553	$0.0000^{<}$	$0.0000^{<}$
Logit-Cloglog	0.6394	$0.0000^{<}$	$0.0000^{<}$
Logit-Probit	$0.0000^{>}$	$0.0000^{<}$	$0.0441^{<}$
Hyptan-Cloglog	0.3230	$0.0000^{>}$	$0.0000^{>}$
Hyptan-Probit	0.3168	$0.0000^{>}$	$0.0000^{>}$
Gauss-Cloglog	0.8763	$0.0000^{<}$	$0.0000^{>}$
Gauss-Probit	$0.0000^{>}$	$0.0000^{>}$	$0.0026^{>}$
Cloglog-Probit	$0.0000^{>}$	$0.0000^{>}$	$0.0451^{<}$

Results

Reference CLOGLOG			
Comparation	$q = 1$	$q = 2$	$q = 5$
Logit-Hyptan	$0.0033^{<}$	$0.0000^{<}$	$0.0000^{<}$
Logit-Gauss	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{<}$
Logit-Cloglog	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Logit-Probit	$0.0001^{<}$	0.0819	$0.0010^{<}$
Hyptan-Cloglog	$0.0032^{>}$	$0.0000^{>}$	$0.0000^{>}$
Hyptan-Probit	$0.0033^{>}$	$0.0000^{>}$	$0.0000^{>}$
Gauss-Cloglog	$0.0000^{>}$	$0.0000^{>}$	$0.0000^{>}$
Gauss-Probit	$0.0000^{<}$	$0.0000^{<}$	0.0734
Cloglog-Probit	$0.0000^{<}$	$0.0000^{<}$	$0.0009^{<}$

Results

Reference PROBIT			
Comparation	$q = 1$	$q = 2$	$q = 5$
Logit-Hyptan	0.0000 $^{<}$	0.0000 $^{<}$	0.0000 $^{<}$
Logit-Gauss	0.0000 $^{<}$	0.0000 $^{<}$	0.0000 $^{<}$
Logit-Cloglog	0.0000 $^{<}$	0.0000 $^{<}$	0.0008 $^{<}$
Logit-Probit	0.0000 $^{>}$	0.0000 $^{>}$	0.0000 $^{<}$
Hyptan-Cloglog	0.0000 $^{>}$	0.0000 $^{>}$	0.0000 $^{>}$
Hyptan-Probit	0.0000 $^{>}$	0.0000 $^{>}$	0.0010 $^{>}$
Gauss-Cloglog	0.0000 $^{<}$	0.0139 $^{>}$	0.0012 $^{>}$
Gauss-Probit	0.0000 $^{>}$	0.0000 $^{>}$	0.4276
Cloglog-Probit	0.0000 $^{>}$	0.0000 $^{>}$	0.0000 $^{<}$

Contents

1 Introduction
- Introduction

2 Complementary log-log and probit functions
- New activation functions

3 Experimental results
- Results

4 Conclusions
- Conclusions

5 Main references
- Main references

Conclusions

- The Monte Carlo simulations were performed with 1,000 replications, at the end of the experiments, the average and standard deviation were calculated for the MSE.
- The simulated data were fitted with different known activation functions known – logit and hyperbolic tangent; and the new activation functions complementary log-log and probit.
- For the majority of the settings used, the mean values of the measures of error revealed statistically significant differences.

- The results reveal that the difference in the average MSE of the functions was lower and statistically significant when the reference function was equal to the activation function used in the MLP network.
- The complementary log-log and probit as activation functions generally presented a lower average MSE than the logit and hyperbolic tangent functions.
- Moreover, the new functions satisfy the requirements of the UAT for being an activation function.

Contents

Main references

- Nelder, J. A. and Wedderburn, W. M. Generalized linear models. Journal of The Royal Statistical Society, 3, 370–384, 1972.
- Hartman, E., Keeler, J. D. and Kowalski, J. M. Layered neural networks with gaussian hidden units as universal approximations. Neural Comput., 2(2), 210–215, 1990.
- Leung, H. and Haykin, S. Rational function neural network. Neural Computation, 5(6), 928–938, 1993.
- Singh, Y. and Chandra, P. A class +1 sigmoidal activation functions for FFANNs. Journal of Economic Dynamics and Control, 28(1), 183–187, October 2003.