Neural networks and logistic regression
Part I

Martin Schumacher[1,2], Reinhard Roßner[1] and Werner Vach[2]

[1]Institute of Medical Biometry and Informatics
University of Freiburg
Stefan-Meier-Str. 26, D-79104 Freiburg im Breisgau
Germany

[2]Center for Data Analysis and Model Building
University of Freiburg
Albertstr. 25-28, D-79104 Freiburg im Breisgau
Germany

Abstract:

Feed-forward neural networks have recently been applied in situations where an analysis based on the logistic regression model would have been a standard statistical approach; direct comparisons of results, however, are seldomly attempted. We therefore present a comparative investigation of both logistic regression models and feed-forward neural networks including some extensions. The theoretical features and properties are reviewed and illustrated in two examples, also discussing practical problems with their application. In Part II of the paper some further important aspects of approximation, overfitting and model selection are investigated in more detail both analytically and by means of simulation studies.

Key Words: Feed-forward neural networks; Logistic regression model; Newton Raphson algorithm; Survival data; Overfitting; Model selection.

1 Introduction

During the last four to five years the successful application of neural networks to practical problems has been demonstrated in numerous papers. In the medical field, these applications range from diagnosis of myocardial infarction (Baxt 1991, Pazos et al. 1992) over classification of EEG signals (Reddy et al. 1992) and PET scans (Kippenhan et al. 1992) to the prediction of mechanisms of action in cancer drug development (Weinstein et al. 1992). Many or even most of these applications are such that an analysis based on a logistic regression model (Cox 1970, Cox & Snell 1989, Hosmer & Lemeshow 1989) or other techniques of discriminant analysis (McLachlan 1992) would have been a standard approach for a statistical analysis. As a typical example, Somoza & Somoza (1993) study the dependence of the admission decision in a psychiatric emergency room on several patient characteristics and diagnostic features. They come to the conclusion that "a neural network can be trained to make clinical decisions that are in substantial agreement with those of experienced clinicians" (Somoza & Somoza 1993).

Despite of the fact that only "successful" applications are reported whereas failures are neither published nor sold (Kree 1994) it is interesting to note that direct comparisons of neural networks to statistical analyses based on the logistic regression model or discriminant analysis are seldomly attempted; we have discovered only very few exceptions (Gallinari et al. 1991, Weinstein et al. 1992, Kippenhan et al. 1993). On the other hand, recent review articles and monographs on statistical aspects of neural networks (Geman et al. 1992, White 1992, Ripley 1993, de Jongh & de Wet 1993, Cheng & Titterington 1994, Cherkassky & Mulier 1994, Murtagh 1994, Michie et al. 1994, Ripley 1994) concentrate on comparisons with modern statistical methods like classification and regression trees (Breiman et al. 1984), generalized additive models (Hastie & Tibshirani 1990), multivariate adaptive regression splines (Friedman 1991) and other nonparametric regression techniques. The logistic regression model and its equivalence to a perceptron with a logistic activation function representing the most simple neural network is usually only briefly mentioned. Thus we feel that a thorough comparative investigation of logistic regression and neural networks still deserves attention.

The plan of the paper is as follows: After a brief review of logistic regression we introduce the basic features of feed-forward neural networks, essentially the perceptron and some extensions, and the back-propagation algorithm. Concentrating on the Kullback-Leibler distance (Kullback & Leibler, 1951) rather than the Euclidean distance as an energy function we show conceptual similarities and discrepancies between both approaches. Then we discuss practical problems occurring in the application of neural networks to given data. The points made are illustrated in two examples, the first using well-known data from the literature (Finney 1947) whereas the second represents a current and typical study in the medical field (Sauerbrei et al. 1994). A discussion finishes Part I of the paper. In Part II (Vach et al. 1994) some further important aspects are investigated in more detail, both analytically and by means of simulation studies. The aspects investigated there include the problem how well regression functions of logistic models can be approximated by neural networks, the role of weight decay and a comparison with model selection procedures.

2 Brief review of logistic regression

In the sequel we consider the situation that we observe a binary outcome variable y and a vector $x = (x_1, \ldots, x_I)$ of regressor variables or covariates for each of N individuals. The logistic regression model (Cox 1970, Cox & Snell 1989, Hosmer & Lemeshow 1989) relates y to x in assuming that

$$Pr(y=1|x) = \Lambda(\beta_0 + \sum_{i=1}^{I} \beta_i x_i) \tag{2.1}$$

where $\Lambda(u) = 1/(1+\exp(-u))$ denotes the logistic function. The unknown regression coefficients β_i which have to be estimated from the data are directly interpretable as log-odds ratios or, in terms of $\exp(\beta_i)$, as odds ratios. The logistic model as a non-linear regression model is a special case of a generalized linear model (McCullagh & Nelder 1989), i.e.

$$\begin{aligned} E(y|x) &= \pi(x,\beta) \\ Var(y|x) &= \pi(x,\beta)(1-\pi(x,\beta)) \end{aligned} \tag{2.2}$$

using the logit link

$$\begin{aligned} \pi(x,\beta) &= \Lambda(\beta_0 + \sum_{i=1}^{I} \beta_i x_i) \\ &= \Lambda(\beta^T x) \end{aligned} \tag{2.3}$$

setting $\beta = (\beta_0, \beta_1, \ldots, \beta_I)$ and adding $x_0 \equiv 1$ to the vector x. Estimation is usually based on the maximum likelihood principle, i.e. on maximizing the log-likelihood function,

$$L(\beta) = \sum_{n=1}^{N} \left[y^{(n)} \log \pi(x^{(n)}, \beta) + (1-y^{(n)}) \log(1-\pi(x^{(n)}, \beta)) \right] , \tag{2.4}$$

where $(x^{(n)}, y^{(n)})$ denote the observed data of the n-th individual, $n = 1, \ldots, N$. Maximizing $L(\beta)$ is equivalent to the minimization of the Kullback-Leibler distance (Kullback & Leibler 1951), which is with the convention that $0 \cdot \log 0 = 0$,

$$\sum_{n=1}^{N} \left[y^{(n)} \log \frac{y^{(n)}}{\pi(x^{(n)}, \beta)} + (1-y^{(n)}) \log \frac{1-y^{(n)}}{1-\pi(x^{(n)}, \beta)} \right] . \tag{2.5}$$

(2.5) can also be written as

$$\sum_{n=1}^{N} -\log(1 - |y^{(n)} - \pi(x^{(n)}, \beta)|)$$

which makes the interpretation of a "distance" between y and $\pi(x, \beta)$ more obvious.

The maximum likelihood estimate $\widehat{\beta}$ can be obtained by solving the likelihood equations ($i = 0, 1, \ldots, I$) that in case of a logistic regression model read

$$\frac{\partial L(\beta)}{\partial \beta_i} = \sum_{n=1}^{N} x_i^{(n)}(y^{(n)} - \pi(x^{(n)}, \beta)) = 0 \ . \tag{2.6}$$

The second derivatives of the log-likelihood function (2.4) are obtained as

$$\frac{\partial^2 L(\beta)}{\partial \beta_i \partial \beta_k} = -\sum_{n=1}^{N} x_i^{(n)} x_k^{(n)} \pi(x^{(n)}, \beta)(1 - \pi(x^{(n)}, \beta)) \ . \tag{2.7}$$

Denoting by $\partial L(\beta)$ the vector of first derivatives and by $\partial^2 L(\beta)$ the matrix of second derivatives of $L(\beta)$, the Newton-Raphson method (Thisted 1988, p. 206) which is commonly used for finding the maximum likelihood estimate $\widehat{\beta}$ can be written as

$$\widehat{\beta}^{(\nu+1)} = \widehat{\beta}^{(\nu)} - \partial^2 L(\widehat{\beta}^{(\nu)})^{-1} \partial L(\widehat{\beta}^{(\nu)}) \tag{2.8}$$

for $\nu = 0, 1, \ldots$ and $\widehat{\beta}^{(0)}$ being an appropriate starting value. It should be noted that in a logistic regression model the Newton-Raphson method and the method of scoring - also known as the Gauss-Newton method (Thisted 1988) - where $\partial^2 L(\beta)$ is replaced by its expected value coincide (cf. Schumacher 1980).

For an individual with covariate vector $x = (x_1, \ldots, x_I)$ the probability $Pr(y = 1|x)$ can then be predicted by

$$\widehat{\pi} = \pi(x, \widehat{\beta}) = \Lambda(\widehat{\beta}_0 + \sum_{i=1}^{I} \widehat{\beta}_i x_i) \ . \tag{2.9}$$

These terms are often simply referred to as "predictions" for given x.

3 Feed-forward neural networks

3.1 The logistic perceptron and back-propagation

The simplest feed-forward neural network is the so-called logistic perceptron displayed in figure 1. This perceptron consists of $I + 1$ input units where x_0

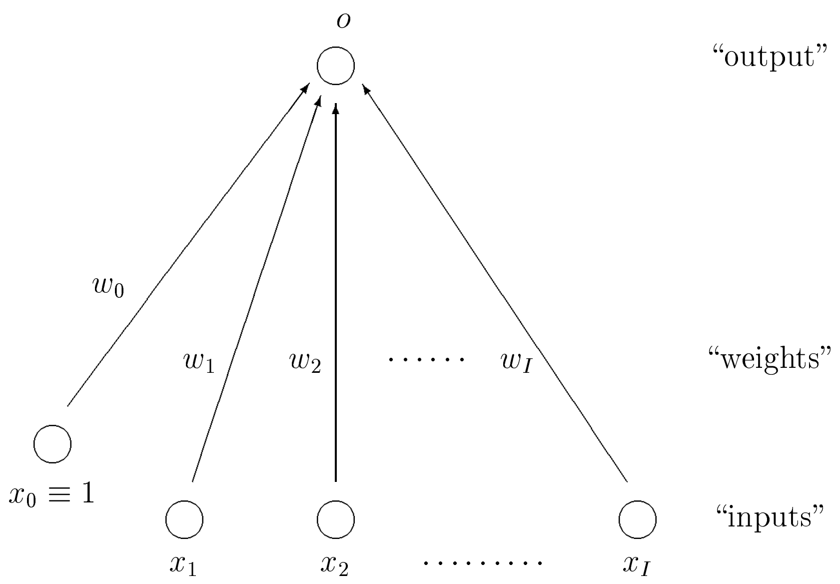

Figure 1: The logistic perceptron

has constant input equal to 1 and one output unit. The input values x_i are weighted with weights w_i $(i = 0, 1, \ldots, I)$ and the sum of weighted inputs is then usually transformed with the logistic function. That is, the output value o can be written as a function f of input values x and weights w,

$$\begin{aligned} o &= f(x, w) \\ &= \Lambda(w_0 + \sum_{i=1}^{I} w_i x_i) \ , \end{aligned} \tag{3.1}$$

where $x = (x_0, x_1, \ldots, x_I)$ denotes the vector of inputs and $w = (w_0, w_1, \ldots, w_I)$ the vector of weights, respectively.

"Learning" in this framework means that based on observations $(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})$ of N individuals weights $\widehat{w}$ are to be determined such that the so-called energy or learning function

$$E(w) = \sum_{n=1}^{N} (y^{(n)} - f(x^{(n)}, w))^2 \tag{3.2}$$

is minimized. This is simply the least squares criterion and for finding the weights $\widehat{w}$ minimizing $E(w)$ we obtain the following estimating equations

$$\frac{\partial E(w)}{\partial w_i} = \sum_{n=1}^{N} 2f(x^{(n)}, w)(1 - f(x^{(n)}, w))x_i^{(n)}(y^{(n)} - f(x^{(n)}, w)) = 0 \tag{3.3}$$

For solving (3.3) the so-called "back-propagation"-method is proposed (Rumelhart et al., 1986). This method, which should more precisely be referred to as least-squares back-propagation (LS-BP) is defined by

$$\widehat{w}^{(\nu+1)} = \widehat{w}^{(\nu)} - \eta \partial E(\widehat{w}^{(\nu)}) \tag{3.4}$$

for $\nu = 0, 1, \ldots$ and $\widehat{w}^{(0)}$ being a suitable starting value. The positive constant η is called the "learning rate" and (3.4) is easily recognized as the well-known method of steepest descent (Thisted 1988, p. 207) with constant step lengths. The nonlinear model underlying the logistic perceptron is identical to the logistic regression model (2.1), i.e. $f(x, w) = \pi(x, \beta)$ and the weights w are equal to the regression coefficients β and can therefore be interpreted in the same way.

As an alternative energy function the Kullback-Leibler distance (2.5) has been proposed (Barron & Barron 1988, Hinton 1989, Fogel 1991, Spackman 1992, Kree 1994) which reads in terms of the logistic perceptron

$$E^*(w) = \sum_{n=1}^{N} \left[y^{(n)} \log \frac{y^{(n)}}{f(x^{(n)}, w)} + (1 - y^{(n)}) \log \frac{1 - y^{(n)}}{1 - f(x^{(n)}, w)} \right] \tag{3.5}$$

Due to its connection to the maximum likelihood principle the back-propagation method using $E^*(w)$ is called maximum-likelihood back-propagation (ML-BP) and can be written as

$$\widehat{w}^{(\nu+1)} = \widehat{w}^{(\nu)} - \eta \partial E^*(\widehat{w}^{(\nu)}) \tag{3.6}$$

for $\nu = 0, 1, \ldots$ and $\widehat{w}^{(0)}$ being a appropriate starting value and $\eta > 0$ the learning rate. The vector of first derivatives is again denoted by $\partial E^*(w)$ with components

$$\frac{\partial E^*(w)}{\partial w_i} = \sum_{n=1}^{N} x_i^{(n)}(y^{(n)} - f(x^{(n)}, w)) \tag{3.7}$$

This leads to estimating equations identical to the likelihood equations (2.6), the maximum likelihood back-propagation method, again the method known as steepest descent, is different to the iteration procedure of the Newton Raphson method in giving all differences $(y - f(x, w))$ equal weights. Having obtained weights $\widehat{w}$ that either minimize (3.5) or (3.2), a "prediction" for an individual with covariate vector $x = (x_1, \ldots, x_I)$ is then, similar as in the logistic regression model, given by $\widehat{o} = f(x, \widehat{w})$.

When estimating w via LS-BP the difference between the observed or target value y and its predicted value $f(x, w)$ is weighted by its variance $f(x, w)(1 - f(x, w))$ and not by the inverse of that variance. Statistical theory (Fahrmeir & Kaufmann 1985, Chamberlain 1987) tells us, that solving the equation (2.6) is optimal, if we are interested in estimating β_i or w_i, respectively, with highest accuracy. Hence weighting the deviance between $y^{(n)}$ and $f^n(x^{(n)}, w)$ as done in (3.3) may be considered questionable from a statistical point of view. However, one can argue, that the use of (3.3) might be suitable for classification, i.e. for the decision whether $f(x, w)$ is less or greater than 0.5. In that case it might be sensible to give deviations $(y - f(x, w))$ for values of $f(x, w)$ close to 0.5 more weight than others in the estimation process.

3.2 Extension of the logistic perceptron to K output units

An obvious way to extend the logistic perceptron is to consider $K > 1$ output units. This neural network which is displayed in figure 2 is characterized by weights w_{ik} $(i = 1, \ldots, I; k = 1, \ldots, K)$ leading to output values o_k which are given by

$$\begin{aligned} o_k &= f(x, w_k) \\ &= \Lambda(w_{0k} + \sum_{i=1}^{N} w_{ik} x_i) \end{aligned} \tag{3.8}$$

for $k = 1, \ldots, K$. In statistical terms this is equivalent to a polychotomous logistic regression model (Bock 1975, Engel 1988) with a categorical outcome

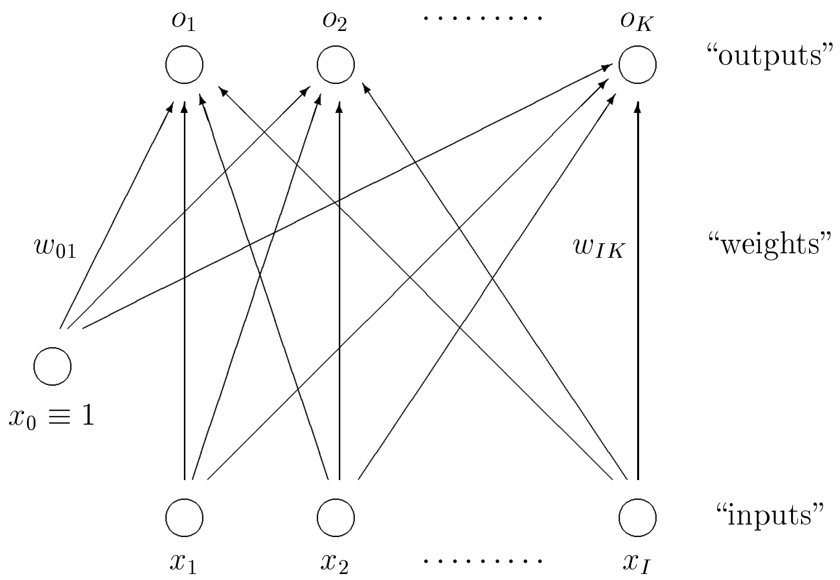

Figure 2: The logistic perceptron with K output units

variable y taking values from $\{1, 2, \ldots, K\}$. Such a model is defined by

$$Pr(y = k|x) = f(x, w_k) / \sum_{k'=1}^{K} f(x, w_{k'}) \tag{3.9}$$

for $k = 1, \ldots, K$. Thus the weights can again be directly interpreted as regression coefficients and ML-BP leads to the maximum likelihood estimates of these coefficients. It should be noted that (3.9) does not impose any ordering on $\{1, 2, \ldots, K\}$. If this is intended this can be reached by some constraints for the weights w_k $(k = 1, \ldots, K)$. The model (3.9) then becomes the well-known proportional odds model for ordered responses (McCullagh 1980, Anderson 1984).

3.3 Networks with one hidden layer

A further extension of the logistic perceptron consists in introducing an additional layer of units, usually called the "hidden layer", between input and output units. The units in this hidden layer are also called "hidden units"; we first consider the case of one hidden unit and one output unit which is schematically displayed in figure 3. All input units are connected with the hidden unit having weights $w_0, w_1, \ldots, w_I$; the hidden unit is then connected with the output unit having weight W_1. An additional connection ("skip layer") with weight W_0 is also drawn from the constant input x_0 and the output unit.

The output o is then obtained as a function of input values given by

$$\begin{aligned} o &= f(x, w) \\ &= \Lambda(W_0 + W_1 \cdot \Lambda(\sum_{i=0}^{I} w_i x_i)) \end{aligned} \tag{3.10}$$

where w summarizes $w_0, w_1, \ldots, w_I$ and W_0, W_1. This network can be seen as a rescaled version of a logistic regression model and, as shown in part II of this paper, the logistic regression model can be approximated by (3.10). The function $u \longmapsto \Lambda(W_0 + W_1 \cdot \Lambda(u))$ is always monotone, thus an essential extension of the logistic perceptron cannot be reached by adding only one hidden unit. ML-BP is a straightforward generalization of (3.6) minimizing the Kullback-Leibler distance, explicit formulae can be found in Spackman (1992) or Ripley (1993, 1994). Instead of one unit in the hidden layer we can also consider $J > 1$ units in this layer. The output is then obtained as a function of input values given by

$$\begin{aligned} o &= f(x, w) \\ &= \Lambda(W_0 + \sum_{j=1}^{J} W_j \cdot \Lambda(\sum_{i=0}^{I} w_{ij} x_i)) \end{aligned} \tag{3.11}$$

Of course, (3.11) can be combined with (3.8) to allow for more than one output unit.

3.4 Extensions for grouped survival data

In some recent articles the application of feed-forward neural nets to survival data has been discussed (Ravdin et al. 1992, Liestøl et al. 1994, Kappen &

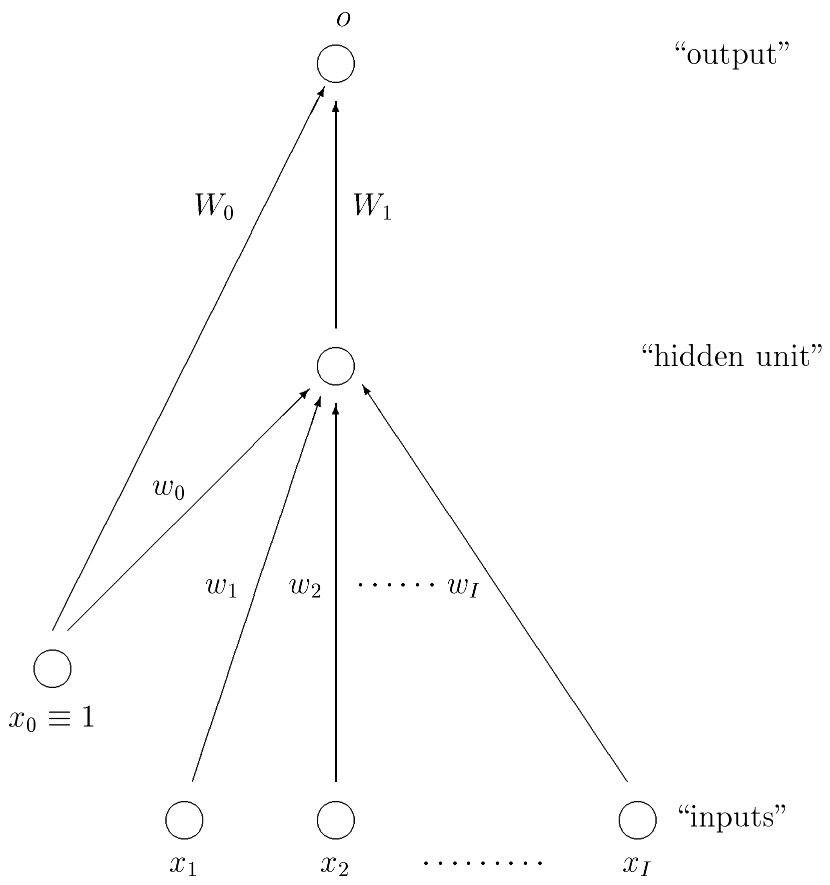

Figure 3: Feed-forward neural network with one hidden unit

Neijt 1993). It is fairly straightforward (Liestøl et al. 1994) to consider a 'grouped' version of the Cox regression model (Cox 1972) in the framework of this paper. Denoting by T the survival time random variable and by I_k the time interval $t_{k-1} \leq t < t_k, k = 1, \ldots, K$, where $0 < t_0 < t_1 < \ldots < t_K < \infty$ the model can be specified through the conditional probabilities

$$Pr(T \in I_k | T \geq t_{k-1}, x) = \Lambda(\beta_{0k} + \sum_{i=1}^{I} \beta_{ik} x_i) \tag{3.12}$$

for $k = 1, \ldots, K$. (3.12) is the original proposal by Cox (1972), other models for grouped survival data can be obtained using other link functions (Prentice

& Gloeckler 1978; Doksum & Gasko 1990). The neural network corresponding to (3.12) is the logistic perceptron (3.8) with K output units displayed in figure 2, the output o_k at the k-th output unit corresponding to the conditional probability of dying in the k-th time interval $I_k, k = 1, \ldots, K$. Data for the n-th individual consists of a vector $x^{(n)}$ of regressor variables (inputs) and a vector $(y_1^{(n)}, \ldots, y_{K_n}^{(n)})$, where $y_k^{(n)}$ is an indicator for individual n dying in the interval I_k and $K_n \leq K$ is the number of intervals in which individual n is observed. Thus $y_1^{(n)}, \ldots, y_{K_n-1}^{(n)}$ are all zero and $y_{K_n}^{(n)}$ is equal to 1 if this individual died in I_{K_n} and equal to 0 if it is censored. This situation implies that the network has a randomly varying number of output nodes according to those time intervals where an individual is "at risk". This is a standard problem of survival analysis (cf. Kalbfleisch & Prentice 1980) and can easily be accomplished by using a slight modification of the energy function E^* from (3.5) or, equivalently, minus the likelihood function in (2.4).

In particular, setting

$$o_k = f(x, w) = \Lambda(\beta_{0k} + \sum_{i=1}^{I} \beta_{ik} x_i)$$

where w summarizes $\beta_{01}, \ldots, \beta_{0K}, \beta_{11}, \ldots, \beta_{IK}$, we obtain

$$E^*(w) = \sum_{n=1}^{N} \sum_{k=1}^{K_n} -\log(1 - |y_k^{(n)} - f(x^{(n)}, w)|) \tag{3.13}$$

This could also be written as the sum over all K output units by introducing an indicator for individual n being at risk at the k-th time interval I_k, $k = 1, \ldots, K$, as it is usually done in survival analysis (Andersen et al. 1993). The proportional hazards assumption can then be implemented in (3.12) through the constraints

$$\beta_{ik} = \beta_i \text{ for } i = 1, \ldots, I \text{ and } k = 1, \ldots, K$$

whereas the parameters β_{0k} are allowed to differ for $k = 1, \ldots, K$.

There have been some other proposals for analysing grouped survival data by means of a feed-forward neural net. Ravdin et al. (1992) and Ravdin & Clark (1992) use a network with only one output unit but using the number k of the time interval as additional input. Moreover, they consider the unconditional survival probability of dying before t_k rather than the conditional one as output. Their underlying model then reads

$$Pr(T < t_k|x) = \Lambda(\beta_0 + \sum_{i=1}^{I} \beta_i x_i + \beta_{I+1} \cdot k) \tag{3.14}$$

for $k = 1, \ldots, K$. This parametrization ensures monotonicity of the survival probabilities but also implies a rather stringent and unusual shape of the survival distribution, since in the case that no covariates are considered (3.14) reduces to

$$\begin{aligned} Pr(T < t_k) &= \Lambda(\beta_0 + \beta_{I+1} \cdot k) \\ &= \frac{1}{1 + \exp(-(\beta_0 + \beta_{I+1} \cdot k))} \end{aligned}$$

for $k = 1, \ldots, K$. Obviously, the survival probabilities do not depend on the length of time intervals I_k $(k = 1, \ldots, K)$, which a is rather strange and undesirable feature. If an additional hidden layer is included (3.14) becomes a straightforward extension of (3.11) having still all the features summarized above.

Other applications of neural nets to survival data are discussed by Liestøl et al. (1994), Kappen & Neijt (1993) and Faraggi & Simon (1995).

4 Practical problems with neural networks

4.1 Learning rate and local minima

Inherent in the back-propagation algorithm (3.4) or (3.6) is the choice of the learning rate η. In the next section we will show that in practical applications some “fine-tuning” is necessary. This is different from using algorithms as the Newton-Raphson or the Gauss-Newton method where the “learning rate” is automatically implemented and changed according to the local behaviour of the energy or learning function. A constant learning rate implies several difficulties. Having a small learning rate the chance of finding minima or escaping from a local minimum is very poor despite of very many iterations. With a large learning rate there is a fair chance of escaping from local minima and going into the direction of the global minimum but the algorithm is very unstable at that global minimum. Of course, this is only relevant in neural networks with a hidden layer where the energy function cannot generally be expected to be unimodal. In those situations some careful alteration of the learning rate in a certain range and/or a repetition of the whole minimization process using different starting values is required to avoid the occurrence of local minima.

4.2 Network selection and weight decay

Application of neural networks to given data requires the specification of the network to be used. In statistical terms this is known as model selection; this term could also be used when concentrating on the non-linear regression model implicitly defined by the neural net. The aspect of model selection is concerned with the choice of variables to be entered into the model or - in terms of neural networks - to be used as input nodes. This can in principle be done in a similar way as in the different strategies for variable selection (cf. Miller 1990) by adding or removing input nodes from the network and comparing the values of the learning function obtained under different settings.

Another important question is how many units in the hidden layer are needed or whether a hidden layer is needed at all. Practically, this can also be investigated by adding or removing units from the hidden layer. For both aspects the differences with respect to complexity of the networks, e.g. number of parameters (regression coefficients or weights), have to be taken into account. For the Kullback-Leibler distance as the energy or learning function some proposals for such a "complexity regularization" (Barron 1991) or "optimal network selection" (Fogel 1991) have been made which are based on variants of Akaike's information criterion (Akaike 1969). In practice, one has to add the number of parameters to the Kullback-Leibler distance for the comparison of networks with either different number of input nodes or different number of units in the hidden layer. For a more theoretical discussion of this problem see also Amari (1993).

Once a network is selected there is still the danger of substantial overfitting of the data, especially in networks with a hidden layer. There are several proposals to avoid such an overfitting; the most popular one seems to be introduction of "weight decay" (Ripley 1993). More precisely, this corresponds to adding a term

$$\lambda \cdot \left(\sum_{j=1}^{J} W_j^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ij}^2 \right) \tag{4.1}$$

to the energy function $E(w)$ or $E^*(w)$ in the neural network specified by (3.11). Of course, other specifications of the weight decay are possible; see Weigend et al. (1991) and Ripley (1993) for a discussion of proposals and their relation to ridge regression and Bayesian statistical methods. Within the framework of this paper we have only used the standard form of weight decay (4.1) which still requires the choice of the decay parameter λ. In the sequel we will see that also for this choice some "fine-tuning" is necessary.

4.3 Software aspects

We do not intend to give an overview on existing software usable for neural nets. Therefore we did not examine the wide area of available tools, we simply picked up packages for our purpose. Special software for neural nets gives much flexibility and an user-friendly interface allowing all specifications of the network, involved calculations and necessary refinements to be done in an interactive mode. Usually, the back-propagation algorithm or one of its variants is used and the network aspects are placed into the foreground. Modifications necessary from a statistical point of view may be rather difficult to implement in these systems coming from experts in informatics.

The second type relies on existing statistical packages like S-Plus (Statistical Sciences Inc. 1991) or SAS (SAS Institute Inc. 1989, 1994) concentrating on the non-linear regression models implicitly defined by a neural net. Here, general minimization algorithms as the Newton-Raphson or Gauss-Newton method are implemented. These algorithms perform very efficiently; on the other hand, possibilities for building and handling neural nets are rather poor.

We used 'SNNS' (Stuttgart Neural Network Simulator, Version 3.2, Zech et al. 1993) as a representative of the first type of software products with the following modifications

a) logistic activation function without bias, that means modeling the intercept explicitly,

b) the standard least-squares learning function was used for LS-BP; a user-written learning function, based on the Kullback-Leibler distance was added for ML-BP.

As a representative of the second type we used Ripley's nnet (Ripley 1993) available in the Statlib library. This S-Plus based software was used in those situations where the network specifications contained at least one unit in the hidden layer. All calculations have been done on a Sun SPARC station 10/40 under Solaris 2.3.

5 Examples

5.1 Deep inspiration and vaso-constriction of the fingers

This well-known example was first presented by Finney (1947, 1971) where also a table of the original data can be found (Finney 1971, p. 184). The data consist of $N = 39$ binary responses denoting the presence ($y = 1$) or absence ($y = 0$) of vaso-constriction in the skin of the fingers after inspiration of a volume V of air at an average inspiration rate R. The data have been extensively used for illustrative purposes; see Pregibon (1982), Morgenthaler (1992) and Fahrmeir & Tutz (1994), among others.

In a first step we start analysing these data by considering the covariate $x_1 = \log R$ only, i.e. we consider the model

$$Pr(y = 1|x) = \Lambda(\beta_0 + \beta_1 x_1) \tag{5.1}$$

Results of this logistic regression analysis (SAS 6.09, PROC LOGISTIC, SAS Institute Inc. 1989) are given in the upper part of table 1. They show a marginally significant effect of the logarithm of the inspiration rate on the probability for the presence of a vaso-constriction in the skin of the fingers. A perceptron (3.1) with two input units ($x_0 \equiv 1, x_1 = \log R$) tranfers this regression model to a neural net. Using ML-BP we know that maximizing the log-likelihood function (2.4) is equivalent to minimizing the Kullback-Leibler distrance (3.5). Thus we should reach identical results if no numerical problems occur. Setting the learning rate to $\eta = 0.001$ for the last iterations ML-BP (SNNS) delivers the weights $\widehat{w}_0 = -0.469$ and $\widehat{w}_1 = 1.335$ with a

Kullback-Leibler distance of $E^* = 24.554$. These figures are in sufficient agreement with the parameter estimates and minus the log-likelihood, $-L = 24.543$; thus a smaller learning rate is not needed. It cannot be expected that LS-BP will produce a similar result since for a data set with $N = 39$ observations the minimization of the least-squares energy function $E(w)$ in (3.2) is not the same as the minimization of the Kullback-Leibler distance. The weights calculated by LS-BP are $\widehat{w}_0 = -0.386$ and $\widehat{w}_1 = 1.172$ which are at least not very far away from the values obtained by the maximum-likelihood method. The two regression curves displayed in figure 4 show clearly that the two methods do not lead to the same predicted response probabilities.

Variable	Estimated Regression Coefficient	Standard Error	P-Value
Intercept	-0.470	0.439	-
$\log R$	1.336	0.665	0.045
		(L=-24.543)	
Intercept	-2.924	1.288	-
$\log R$	4.631	1.789	0.010
$\log V$	5.221	1.858	0.005
		(L=-14.632)	

Table 1: Results of the logistic regression analyses for the vaso-constriction data (model (5.1) upper part; model (5.2) lower part)

An adequate analysis of this data requires the use of the second covariate; thus we consider the logistic regression model

$$Pr(y = 1|x) = \Lambda(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \tag{5.2}$$

with $x_1 = \log R$ and $x_2 = \log V$, respectively. The results of this analysis are presented in the lower part of table 1; both covariates show a significant influence on the probability of a vaso-constriction in the skin of the fingers.

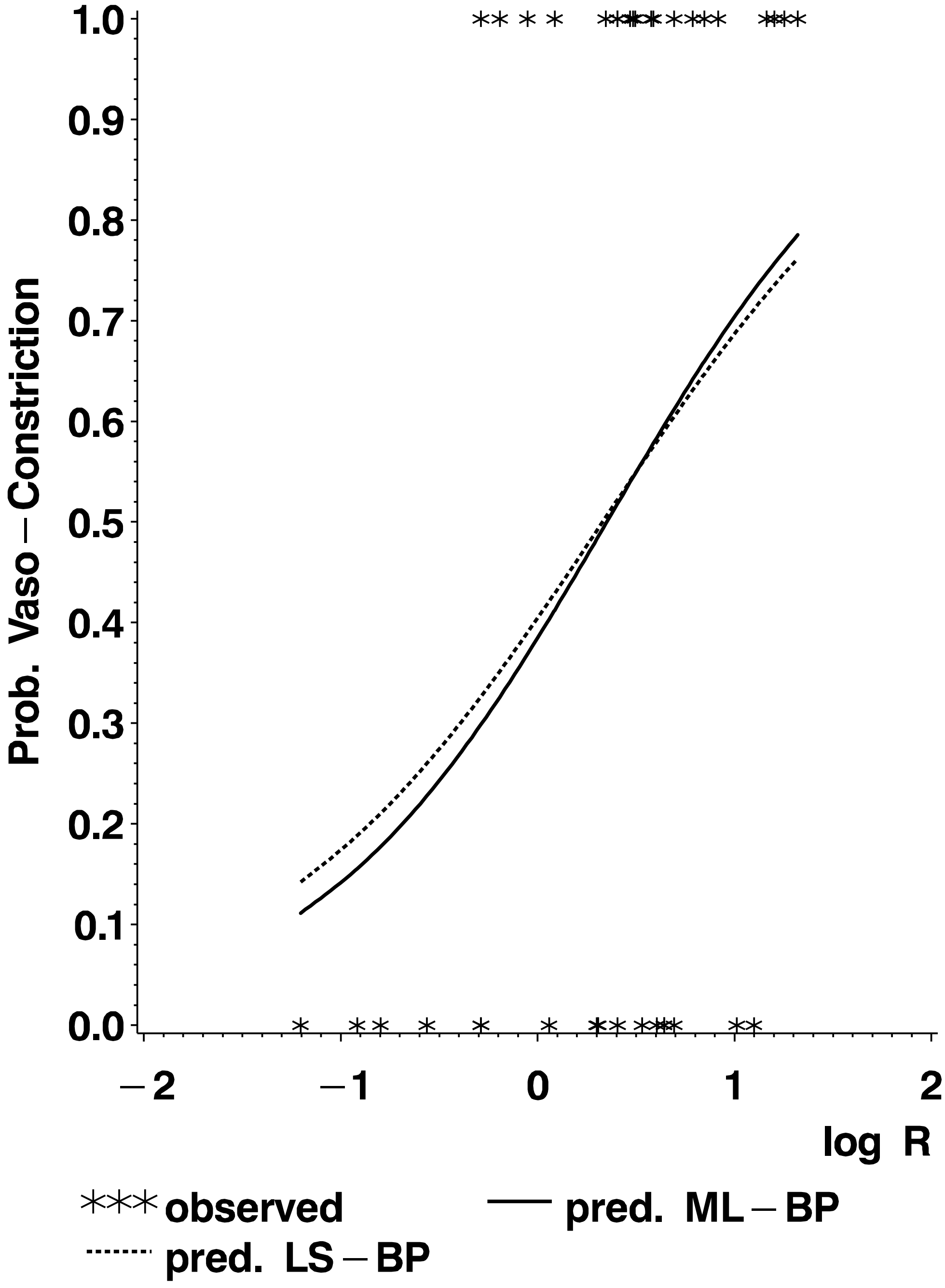

Figure 4: Estimated response probabilities obtained by LS-BP (—) and ML-BP (- - - -) and observed responses (*) for the vaso-constriction data based on model (5.1)

The scatter plot displayed in figure 5 shows that this data represent a rather extreme situation; only two observations prevent the groups defined by $y = 0$ and $y = 1$, respectively, to be separated nearly by a straight line.

A perceptron (3.1) with three input units ($x_0 \equiv 1, x_1 = \log R, x_2 = \log V$) should give similar results, if ML-BP is used. This was nearly achieved ($\widehat{w}_0 = -2.938, \widehat{w}_1 = 4.650, \widehat{w}_2 = 5.240$) although the back-propagation algorithm required several changes of the learning rate from an initial value of $\eta = 0.2$ down to $\eta = 0.000001$ involving approximately 20000 iterations to get a value of the Kullback-Leibler distance of $E^* = 14.73$ comparable to that of minus the log-likelihood ($L = -14.632$). The "learning process" is graphically displayed in figure 6 for both ML-BP and LS-BP. The latter algorithm continues to reduce the quadratic error. The absolute values of the weights $\widehat{w}_j$ increase continuously leading to a very steep regression function at the border between the observations with $y = 0$ and $y = 1$. The two observations with $y = 1$ sitting among those with $y = 0$ (marked with a circle in figure 5) add an amount of maximally 2.0 to the least squares energy function, whereas such a misclassification under ML-BP would increase the Kullback-Leibler distance to infinity. The weights obtained after about 1500 iterations are $\widehat{w}_0 = -7.706, \widehat{w}_1 = 10.265$ and $\widehat{w}_2 = 11.942$ leading to a least-squares energy function of $E = 3.7$.

Adding a hidden layer with one unit to the net as given by (3.10) the absolute values of the weights are again drastically increased leading to a Kullback-Leibler distance of $E^* = 8.763$. This was achieved by minimizing the Kullback-Leibler distance with a Gauss-Newton type algorithm - which is essentially maximum likelihood estimation - by nnet (Ripley 1993) within 50 iterations. ML-BP in SNNS needed $3 \cdot 10^6$ iterations to reach a comparable value of the Kullback-Leibler distance. With this network all predictions for observations with $y = 0$ are set to $\widehat{o} = 0.134$ and to $\widehat{o} = 1$ for observations with $y = 1$, with the exception of those two marked observations (see figure 5) where the predictions are also set to $\widehat{o} = 0.134$ (see also figure 7). For comparing the Kullback-Leibler distances a value of $+2$ has to be added to $E^* = 8.763$ since two more parameters are involved than in the logistic perceptron. With $J = 3$ units in the hidden layer it was possible (in 3 out of 10 trials) to predict all observations correctly, i.e. $\widehat{o} = 0$ for observations with $y = 0$ and $\widehat{o} = 1$ for observations with $y = 1$. "Maximum likelihood estimation" via nnet was again performed within 50 iterations whereas ML-BP implemented in SNNS needed about a million iterations. In the other

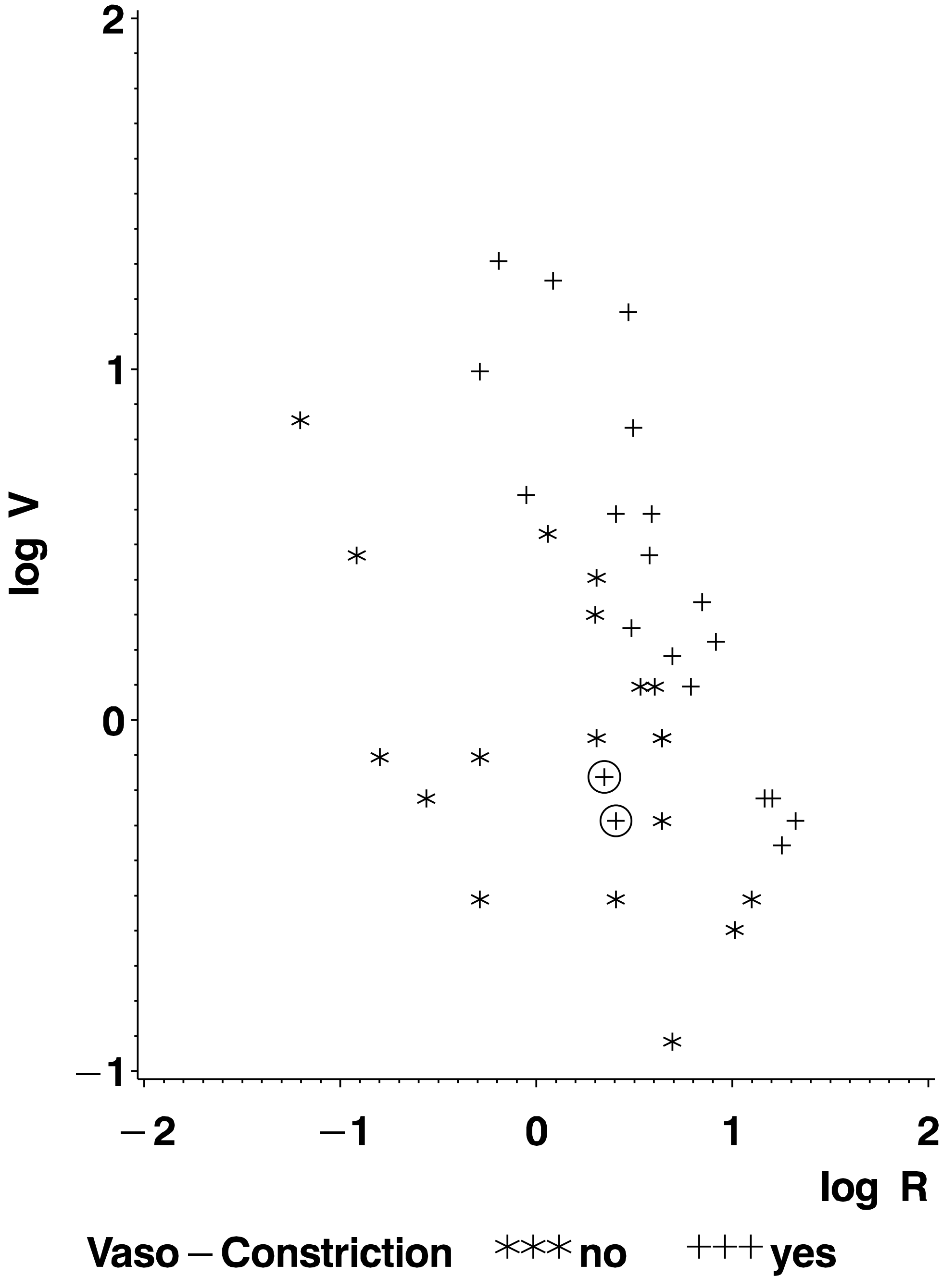

Figure 5: Scatter plot of vaso-constriction data with observed responses ($y = 0$: '$*$'; $y = 1$: '$+$'); two influential observations are marked with '$\oplus$'

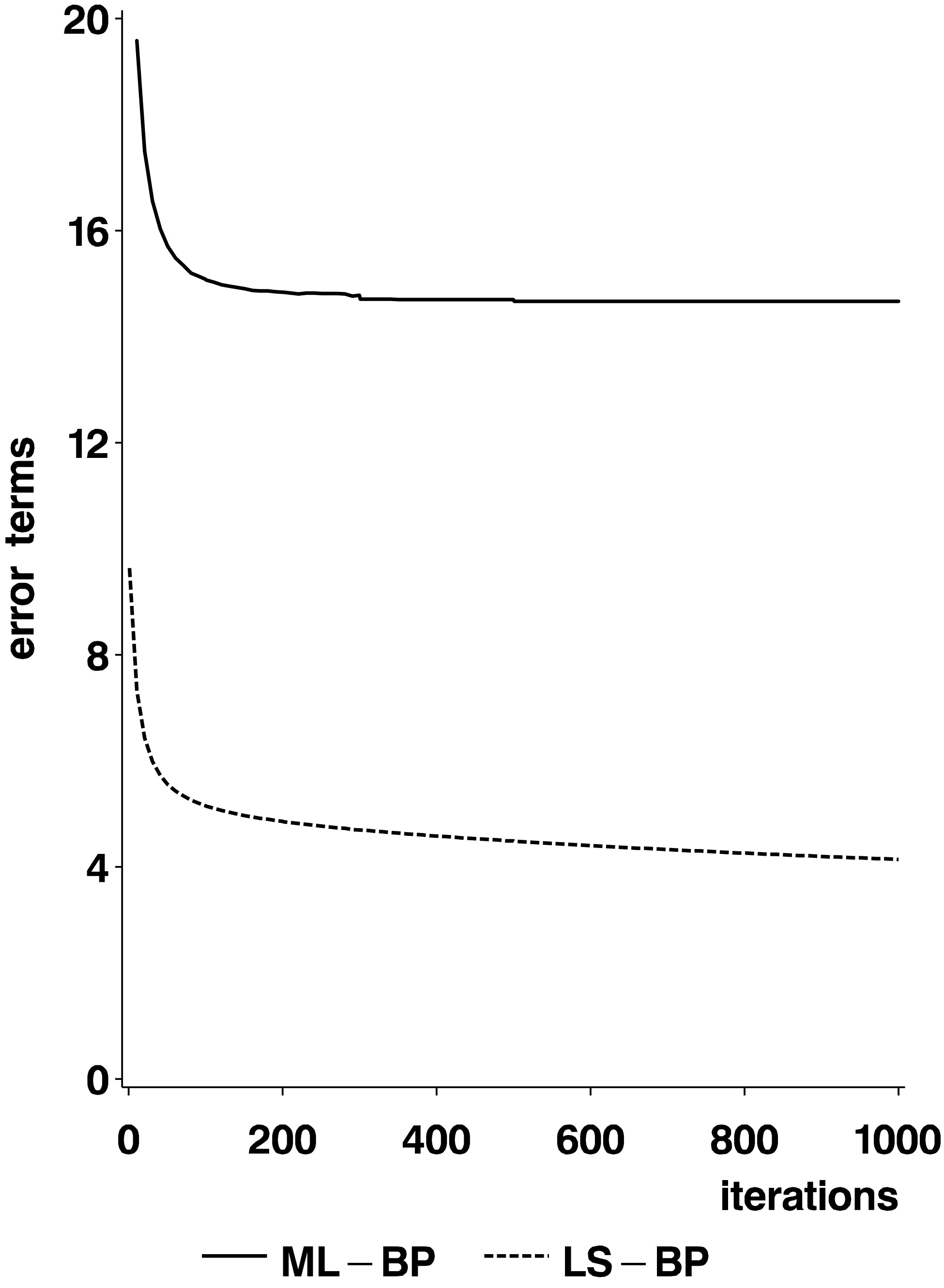

Figure 6: Learning process of the logistic perceptron based on model (5.2) for LS-BP (—) and ML-BP (- - - -)

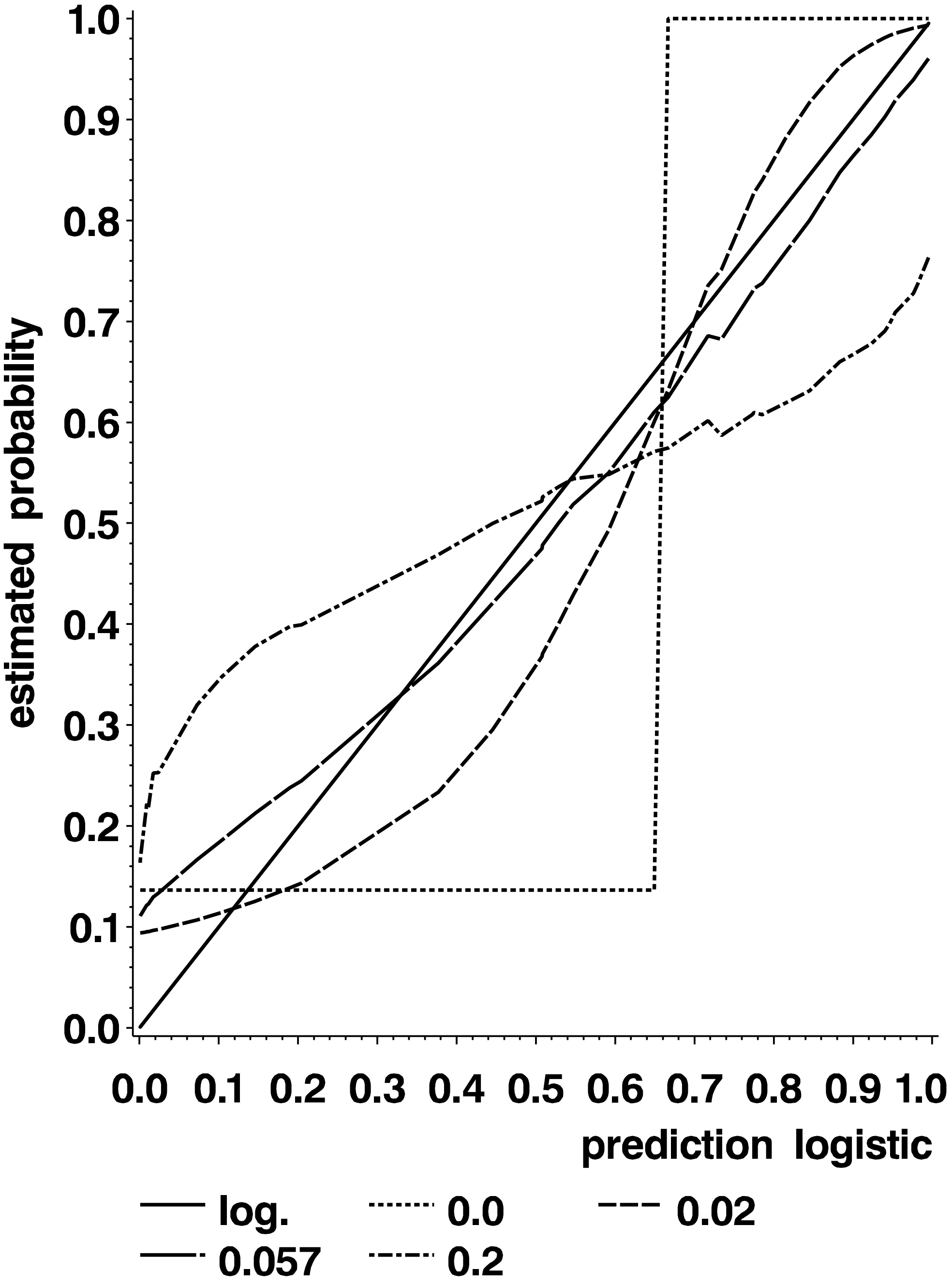

Figure 7: Estimated response probabilities for the vaso-constriction data from a feed- forward neural net with one hidden unit and weight decay for different choices of the decay parameter λ using the predictions from the logistic regression model (5.2) as baseline.

7 (out of 10) trials both algorithms stopped in a local minimum leading to qualitatively similar results as obtained with only one unit in the hidden layer.

A tool for preventing such overfitting is weight decay as outlined in section 4. So we added a term proportional to the sum of squared weights to the Kullback-Leibler distance penalizing large weights. For the neural net with one unit in the hidden layer we tried several choices for the decay parameter λ. Figure 7 shows the influence on the predicted response probabilities using the results from the logistic regression analysis based on (5.2) as reference. The choice of $\lambda = 0$ corresponds to the case that no weight decay is used which has been described above. The choice of $\lambda = 0.02$ for the decay parameter is still associated with some tendency to overfitting. A value of $\lambda = 0.057$ leads to the same Kullback-Leibler distance as obtained in the perceptron with no hidden layer and gives similar predictions. A choice of $\lambda = 0.2$ leads to a value of the Kullback-Leibler distance of $E^* = 19.28$ indicating a poor fit of the model which can also be seen from figure 7.

5.2 Sonographic measurements and diagnosis in breast cancer

This study (Sauerbrei, Madjar & Prömpeler, 1994) has been performed to examine the role of noninvasive sonographic measurements (colour Doppler flow signals) for differentiation of benign and malignant breast tumors. Within approximately two years, $N = 458$ women have been enrolled in the study and several sonographic measurements and patient characteristics $(x_1, \ldots, x_I)$ have been recorded. Histological verification yielded 325 benign tumors $(y = 0)$ and 133 malignant tumors $(y = 1)$ as final diagnosis. A preliminary logistic regression analysis with six covariates identified three as marginally significant. However, strong collinearity lead to grossly inflated variances and standard errors. These well-known difficulties added to the numerical deficiencies of the back-propagation algorithm did not admit to get similar results with ML-BP in a logistic perceptron. Even the first and second digits of the weights were not the same as those of the estimated regression coefficients in the logistic regression model.

We therefore decided to consider only three covariates, age, number of arteries in the tumor (AT) and number of arteries in the contralateral breast

(AC), respectively. The results of this reduced logistic regression model, which was also obtained by a backward elimination procedure, are presented in table 2. It is interesting to note that the log-likelihood is nearly identical to the one obtained in the model where all six covariates have been included and that problems with collinearity are no longer present.

Variable	Estimated Regression Coefficient	Standard Error	P-Value
Intercept	-8.179	0.924	-
Age (x_1)	0.070	0.017	0.0001
$\log AT + 1$ (x_2)	5.187	0.575	0.0001
$\log AC + 1$ (x_3)	-1.074	0.437	0.014
		(L=-79.996)	

Table 2: Results of a logistic regression analysis for the breast tumor data

A logistic perceptron with four input units ($x_0 \equiv 1, x_1 =$age$, x_2 = \log AT + 1, x_3 = \log AC + 1$) and ML-BP yielded similar results after carefully altering the learning rate within about 30000 iterations. The Kullback-Leibler distance was $E^* = 80.003$ and the weights obtained were $\widehat{w}_0 = -8.108, \widehat{w}_1 = 0.069, \widehat{w}_2 = 5.162, \widehat{w}_3 = -1.081$. In this well-behaved situation with three covariates the sample size of $N = 458$ was large enough to insure consistency also of the weights obtained with LS-BP; with weights $\widehat{w}_0 = -8.247, \widehat{w}_1 = 0.071, \widehat{w}_2 = 5.180, \widehat{w}_3 = -0.792$ and a least squares energy function $E = 23.543$ at least qualitatively similar results were achieved with LS-BP as compared to ML-BP and maximum likelihood estimation. The results of the logistic regression are compared to a classification developed by Sauerbrei et al. (1994) based on the CART-method (Breiman et al. 1984) and feed-forward neural networks with a hidden layer and an increasing number J of hidden units. For comparison, we use the sensitivity; specificity and percentage of correct classifications knowing that these quantities are too optimistic since they are estimated for the same data. The results are summarized in table 3; for the classification rules we always use a cutpoint of 0.2 for the

Method	Sensitivity	Specificity	Percentage of correct classifications
Logistic regression	95.5%	90.2%	91.7%
CART	97.9%	89.5%	91.9%
NN(1)	95.5%	92.0%	93.0%
NN(2)	97.0%	92.0%	93.4%
NN(3)	96.2%	92.3%	93.4%
NN(4)	94.7%	93.5%	93.9%
NN(6)	97.7%	94.8%	95.6%
NN(8)	97.7%	95.4%	96.1%
NN(10)	98.5%	98.5%	98.5%
NN(15)	99.2%	98.2%	98.5%
NN(20)	99.2%	99.1%	99.1%
NN(40)	99.2%	99.7%	99.6%

Table 3: Results of classification rules based on logistic regression, CART and feed-forward neural networks with J hidden units ($NN(J)$) for the breast tumor data

predictions. As can be seen the diagnostic ability of the classification rules based on logistic regression and the CART-method is nearly the same. On the other hand, an increasing number J of hidden units leads to a perfect classification of benign and malignant tumors; this is a further indication that there is inherently a substantial overfitting of the observed data.

6 Discussion

In this paper we have compared the logistic regression model with feed-forward neural networks. As we have seen the logistic perceptron as the most simple neural net is equivalent to a logistic regression model and the weights in the neural net can be interpreted as regression coefficients in the corresponding logistic model. The method of estimation, however, is different as long as the least squares criterion is used as energy function. From a statistical point of view it is desirable to use the Kullback-Leibler distance or entropy criterion as energy function. In this case the estimating equations are identical to the likelihood equations and thus lead to the same estimates with established properties.

In neural networks, the back-propagation method is used as a standard algorithm for minimizing the energy function or, equivalently, for solving the estimating equations. As shown in our paper, this method, known as "steepest descent", has certain drawbacks (see also Cheng & Titterington 1994). Its use is only understandable when having the neural net as a simplified model of human brain in mind. If one is prepared to accept that a neural net is nothing than a special non-linear regression model, back-propagation has no longer to be seen as an essential ingredient and other methods as the Newton-Raphson and Gauss-Newton algorithm should be preferred. This also avoids the considerations on the choice of the learning rate inherent in the back-propagation method.

There are some obvious extensions of the logistic perceptron allowing for more than one output unit and/or having units in a hidden layer. As shown above, these feed-forward neural nets have also non-linear regression models as their natural counterparts and can in principle be handled in the same way as the logistic perceptron. There is one difference in so far, that the weights in the neural net cannot be generally interpreted in the simple way as in the logistic perceptron. Some of these extensions proposed for grouped survival data deserve a special remark. There is some danger that naive applications of neural nets to survival data do ignore the developments concerning the statistical methods in survival analysis over the past forty years; Liestøl et al. (1994) is a 'sign-posting' reference for future developments in this area.

Neural nets with more than one unit in the hidden layer which correspond to a large number of parameters in the non-linear regression models are very flexible and therefore carry the danger of substantial overfitting. As shown in the two examples presented this can at least partially be compensated by the introduction of weight decay. This, however, implies the proper choice of the decay parameter. These problems and a comparison of neural nets to logistic regression models involving quadratic or interaction terms are investigated in more detail in Part II of this paper (Vach et al. 1994).

We have concentrated on a comparison of neural nets to logistic regression models and not, as many others, on a comparison to modern non-parametric regression methods. There have been two reasons for that: The first is that in many applications, at least in the medical field, an analysis based on the logistic regression model would have been a standard statistical approach. Secondly, we felt that for many of these modern non-parametric techniques there is still some deficit with respect to a full understanding of their statistical properties. In our second example we have seen that standard logistic regression techniques can yield a classification rule with both high sensitivity and specificity for the diagnosis of breast cancer; the results from a CART analysis are very similar. The diagnostic accuracy can only slightly be improved by a neural net when the number of units in the hidden layer is substantially increased. The number of additional parameters in the corresponding non-linear regression model implies the risk of substantial overfitting of the observed data which has to be taken into account. In such a case, single reporting that a neural net has been successfully applied to decision making in the diagnosis of breast cancer (Wu et al. 1993) is not interesting without a comparison to the results obtained by standard statistical techniques. Thus it is not sufficient to mention the need for such comparisons (Tu & Guerriere 1993); they have to be done and their results explicitly given.

Finally, there is the question when and under which circumstances neural nets should be used or preferred to statistical techniques. There is a tremendous discussion about that point ranging from statements

> 'The main advantage of neurocomputing over statistics $(\cdots)$ seems to be neurocomputing's ability to draw from many more sources of inspiration $(\cdots)$ for its methods than are exploited by workers in statistics' (Hecht-Nielsen 1990, p. 120)

and

> 'Neural networks are enticing, statistics are not.' (White 1992, p. 81)

to

> 'Many types of artificial neural networks are merely re-inventions of well-known statistical methods implemented via inefficient algorithms' (SAS-Institute 1994).

Our paper is not intended to give a definite answer to this question which might be somewhere between these extreme statements. We hope, however, that it will, jointly with others, help to clarify the role of neural networks for future developments.

Acknowledgements

We would like to thank our colleagues Rudolf Lerche, Ludger Rüschendorf, Willi Sauerbrei and Jürgen Schulte Mönting for fruitful discussions during a series of seminars on "Statistical aspects of neural networks". We are also greatful to Dr. Helmut Madjar for the permission to use his data for illustrative purposes.

References

Akaike, H. (1969): *Fitting autoregressive models for prediction.* Annals of the Institute of Statistical Mathematics 21, 243-247.

Amari, S.J. (1993): *Mathematical methods in neurocomputing.* In: Barndorff-Nielsen, O.E. et al. (eds.), *Chaos and networks-statistical and probabilistic aspects.* London: Chapman and Hall.

Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N. (1993): *Statistical methods based on counting processes*, Springer Series in Statistics, New York: Springer-Verlag.

Anderson, J.A. (1984): *Regression and ordered categorical variables (with discussion).* Journal of the Royal Statistical Society B 46, 1-30.

Barron, A.R. (1991): *Complexity regularization with application to artificial neural networks.* In: Roussas, G. (ed.), *Nonparametric functional estimation and related topics.* Amsterdam: Kluwer Academic Publishers.

Barron, A.R., Barron, R.L. (1988): *Statistical learning networks: A unifying view.* In: Wegman, E. (ed.), *Computing science and statistics: Proceedings 20th Symposium Interface*: 192-203. Washington: American Statistical Association.

Baxt, W. G. (1991): *Use of an artificial neural network for the diagnosis of myocardial infarction.* Annals of Internal Medicine 115, 843.848

Bock, D.R. (1975): *Multivariate statistical methods in behavioral research*, New York: McGraw-Hill.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984): *Classification and regression trees*, Monterey: Wadsworth and Brooks/Cole.

Chamberlain, G. (1987): *Asymptotic efficiency in estimation with conditional moment restrictions.* Journal of Econometrics 34, 305-324.

Cheng, B., Titterington, D.M. (1994): *Neural networks: A review from a statistical perspective (with discussion).* Statistical Science 9, 2-54.

Cherkassky, V., Mulier, F. (1994): *Statistical and neural network techniques for nonparametric regression.* In: Cheeseman, P. & Oldford, R.W., *Selecting models from data: AI and statistics IV*: 383-392. New York: Springer.

Cox, D.R. (1970): *Analysis of binary data*, London: Methuen.

Cox, D.R. (1972): *Regression models and life tables (with discussion).* Journal of the Royal Statistical Society B 34, 187-220.

Cox, D.R., Snell, E.J. (1989): *Analysis of binary data, 2nd edition*, London: Chapman and Hall.

de Jongh, P.J., de Wet, T. (1993): *An introduction to neural networks.* South African Statistical Journal 27, 103-128.

Doksum, K.A., Gasko, M. (1990): *On a correspondence between models in binary regression analysis and in survival analysis.* International Statistical Review 58, 243-252.

Engel, J. (1988): *Polytomous logistic regression.* Statistica Neerlandica 42, 233-252.

Fahrmeir, L., Kaufmann, H. (1985): *Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models.* Annals of Statistics 13, 342-368.

Fahrmeir, L., Tutz, G. (1994): *Multivariate statistical modelling based on generalized linear models*, New York: Springer.

Faraggi, D. & Simon, R. (1995): *A neural network for survival data.* Statistics in Medicine 14, 73-82.

Finney, D.J. (1947): *The estimation from individual records of the relationship between dose and quantal response.* Biometrika 34, 320-334.

Finney, D.J. (1971): *Probit analysis*, Third edition, Cambridge: Cambridge University Press.

Fogel, D.B. (1991): *An information criterion for optimal neural network selection.* IEEE Transactions On Neural Networks 2 no. 5, September 1991, 490-497.

Friedman, J.H. (1991): *Multivariate adaptive regression splines (with discussion).* Annals of Statistics 19, 1-141.

Gallinari, P., Thiria, S., Badran, F., Fogelman-Soulie, F. (1991): *On the relations between discriminant analysis and multilayer perceptrons.* Neural Networks 4, 349-360.

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

Hastie, T.J., Tibshirani, R.J. (1990): *Generalized additive models*, London: Chapman and Hall.

Hecht-Nielsen, R. (1990): *Neurocomputing*, Reading, MA: Addison-Wesley.

Hertz, J.A., Krogh, A., Palmer, R.G. (1991): *Introduction to the theory of neural computation*, Redwood City, U.S.A.: Addison Wesley Publishing Company.

Hinton, G.E. (1989): *Connectionist learning procedures.* Artificial Intelligence 40, 185-234.

Hosmer, D.W., Lemeshow, S. (1989): *Applied logistic regression*, New York: John Wiley.

Kalbfleisch, J.D., Prentice, R. (1980): *The statistical analysis of failure time data*, New York: Wiley.

Kappen, H.J., Neijt, J.P. (1993): *Neural network analysis to predict treatment outcome.* Annals of Oncology 4, Suppl, 31-34.

Kippenhan, J.S., Barker, W.W., Pascal, S., Nagel, J., Duara, R. (1992): *Evaluation of a neural-network classifier for PET scans of normal and Alzheimer disease subjects.* Journal of Nuclear Medicine 33, 1459-1467.

Kree, R. (1994): *Neural networks: Architectures, learning and performance.* In: Bock, H.H., Lenski, W., Richter, M. (eds.), *Information systems and data analysis.* Berlin-Heidelberg: Springer.

Kullback, S., Leibler, R.A. (1951): *On information and sufficiency.* Annals of Mathematical Statistics 22, 79-86.

Liestøl, K., Andersen, P.K., Andersen, U. (1994): *Survival analysis and neural nets.* Statistics in Medicine 13, 1189-1200.

McCullagh, P. (1980): *Regression models for ordinal data (with discussion).* Journal of the Royal Statistical Society B 42, 109-142.

McCullagh, P., Nelder, J.A. (1989): *Generalized linear models*, Second edition. New York: Chapman and Hall.

McLachlan, G.J. (1992): *Discriminant analysis and statistical pattern recognition*, New York: Wiley.

Michie, D., Spiegelhalter, D.J., Taylor, C.C. (1994): *Machine learning, neural and statistical classification*, New York: Ellis Horwood.

Miller, A.J. (1990): *Subset selection in regression*, London: Chapman and Hall.

Morgenthaler. S. (1992): *Least-absolute-deviations fits for generalized linear models.* Biometrika 79, 767-754.

Murtagh, F. (1994): *Neural networks and related 'massively parallel' methods for statistics: A short overview.* International Statistical Review 62, 275-288.

Pazos, A., Maojo, V., Martin, F., Ezquerra, N. (1992): *A neural network approach to assess myocardial infarction.* In: Lun et al. (eds.), *Medinfo*: 92. 659-663: Amsterdam.Elsevier

Pregibon, D. (1982): *Resistant fits for some commonly used logistic models with medical applications.* Biometrics 38, 485-498.

Prentice, R.L., Gloeckler, L.A. (1978): *Regression analysis of grouped survival data with application to breast cancer data.* Biometrics 34, 57-67.

Ravdin, P.M., Clark, G.M. (1992): *A practical application of neural network analysis for predicting outcome of individual breast cancer patients.* Breast Cancer Research and Treatment 22, 285-293.

Ravdin, P.M., Clark, G.M., Hilsenbeck, S.G., Owens, M.A., Vendely, P., Pandian, M.R. (1992): *A demonstration that breast cancer recurrence can be predicted by neural network analysis.* Breast Cancer Research and Treatment 21, 47-53.

Reddy D.C., Korrai, D.R. (1992): *Neural networks for classification of EEG signals.* In: Lun et al. (eds.), *Medinfo*: 92. 653-658: Amsterdam.Elsevier

Ripley, B.D. (1993): *Software for neural networks*, Technical Report: University of Oxford.

Ripley, B.D. (1993): *Statistical aspects of neural networks.* In: Barndorff-Nielsen, O.E. et al. (eds.), *Chaos and networks - statistical and probabilistic aspects.* London: Chapman and Hall.

Ripley, B.D. (1994): *Neural networks and flexible regression and determination.* In: Mardia, K.V. (ed.), *Statistics and images.* Abingdon: Carfax.

Ripley, B.D. (1994): *Neural networks and related methods for classification (with discussion).* Journal of the Royal Statistical Society B 56, 409-456.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986): *Learning representations by back-propagation errors.* Nature 323, 533-536.

SAS-Institute Inc. (1989): *SAS/STAT user's guide*, Version 6, Fourth Edition. Cary N.C.: SAS-Institute Inc.

SAS-Institute Inc. (1994): *Information on neural network macros.*

Sauerbrei, W., Madjar, H., Prömpeler, H.J. (1994): *Development of diagnostic rules based on color Doppler flow signals for differentiation of benign and malignant breast tumors*: Preprint.

Schumacher, M. (1980): *Point estimation in quantal response models.* Biometrical Journal 22, 315-334.

Somoza, E., Somoza, J.R. (1993): *A neural-network approach to predicting admission in a psychiatric emergency room.* Medical Decision Making 13, 273-280.

Spackman, K.A. (1992): *Maximum likelihood training of connectionist models: Comparison with least squares back-propagation and logistic regres-*

sion: Proceedings of the 15th Annual Symposium on Computer Applications in Medical Care 1991.

Statistical Sciences Inc. (1991): *S-Plus data analysis software, Version 3.x*, Seattle: Statistical Sciences Inc..

Thisted, R.A. (1988): *Elements of statistical computing: Numerical computation*, New York - London: Chapman and Hall.

Tu, J.V., Guerriere, M.R.J. (1993): *Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery.* Computers in Biomedical Research 26, 220-229.

Vach, W., Roßner, R., Schumacher, M. (1994): *Neural networks and logistic regression, Part II*, this volume.

Weigend, A., Rumelhart, D., Huberman, B. (1991): *Generalization by weight elimination with application to forecasting*, In: Advances in neural information processing III: Morgan Kaufman.

Weinstein, J.N., Kohn, K.W., Greuer, M.R., Viswanadhan, V.N., Rubinstein, L.V. et al. (1992): *Neural computing in cancer drug development: Predicting mechanism of action.* Science 258, 447-451.

White, H. (1992): *Artificial neural networks: Approximation and learning theory*, Oxford: Basil Blackwell.

Wu, Y., Giger, M.L., Doi, K., Vyborny, C.J., Schmidt, R.A., Metz, C.E. (1993): *Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer.* Radiology 187, 81-87.

Zell, A., Mache, N., Hübner, R., Mamier, G., Vogt, M. et al. (1993): *Stuttgart Neural Network Simulator; User Manual.* Version 3.2, Report no. 3/93. Institute for Parallel and Distributed High Performance Systems, University of Stuttgart.