ON THE ERROR CRITERIA IN NEURAL NETWORKS AS A TOOL FOR HUMAN CLASSIFICATION MODELLING

Louis ten Bosch[1], *Roel Smits*[2]

[12] Institute of Perception Research/IPO, Eindhoven, NL
[1] Currently at Lernout & Hauspie Speech Products N.V., Ieper, Belgium, and
Institute of Phonetic Sciences/IFOTT, University of Amsterdam, NL
e-mail: louistb@fon.let.uva.nl
[2] Currently at UCL, Phonetics Dept., London, UK.

ABSTRACT

Multi Layer Perceptrons (MLPs) can be applied as a tool to model human classification behaviour. In the present theoretical study we attempt to interpret MLPs within the framework of mathematical psychological models for human classification behaviour, more specifically the General Recognition Theory and the Generalized Context Model. Next, four error criteria are discussed that can be used in training and test of the MLPs, in relation to two types of data representation: in terms of individual deterministic responses or in terms of probabilistic responses. All error measures considered are additive, i.e. can be written as a sum across individual stimuli.
It will be shown that some of these error measures have very different properties given a training set, and that the interpretation of the MLP as a means to provide knowledge about the underlying human decision process depends on the complexity of the MLP-topology.

1. Prerequisites

In mathematical psychology, models have been developed to study the process of (human) classification and identification in detail. Broadly speaking, these models can be distinguished into three types: (1) the General Recognition Theory (GRT), by Ashby and others; (2) the Generalized Context Model (GCM), by Nosofsky and colleagues, and the Fuzzy Logical Model of Perception (FLMP), by Massaro. We here briefly discuss the GRT and the GCM, since these models are close to an interpretation of the Multi Layer Perceptron as a model for human classification behaviour. In this paper, we will study the modelling of the process of classifying continuous stimuli in terms of a finite set of labels.
In the GRT (cf. Ashby & Townsend, 1986), it is assumed that the incoming stimulus can be represented as a point in a (high-dimensional) perceptual feature space S. Due to sensorial noise, the representation is blurred and as a consequence the stimulus yields a probability density function on S. The feature space S is assumed to be partitioned into regions within which all stimuli are assigned to one classification label. This assignment can be probabilistic or deterministic. The between-class boundaries may be linear or be more complex, depending on the structure of the classes to be distinguished.
In the GCM (cf. Nosofsky, 1986), each class is represented by a number of so-called exemplars. The assigment of a label to an incoming stimulus is performed after evaluation of the similarities between the stimulus and all exemplars of each class, and by selection of the class with the highest overall similarity with the stimulus.
While the GRT focusses on the boundaries between classes, the GCM emphasizes the exemplars of the classes. This difference is important for model claims with respect to human learning processes. In practice, it is not easy to discriminate between class boundary models and exemplar-based models on the basis of their performance (cf. Maddox & Ashby, 1993; McKinley & Nosofsky, 1995). We add to say that in the literature, the focus is often on the training process itself as a model for the human learning process. This learning aspect itself is entirely discarded in the present approach.
We finally observe that the entire process of stimulus classification can be described in three basic steps:
(1) **Representation**: the choice of the representation features of the stimulus, including the number of features, scaling, compression, etc. This yields a feature vector X for each stimulus.
(2) **Probabilistic choice model**: a mapping from the stimulus feature vector X of a vector of class probabilities $P(C_i|X)$ where $C_1, C_2, \ldots$ denote the classes to be distinguished.
(3) **Deterministic choice model**: the selection of a class, based on the probabilities $P(C_i|X)$. This step may be trivial (if the winner class takes all), but may also be complicated (for example if there is no clear winner class).
The results in this paper are more extensively described in ten Bosch & Smits (1996).

2. Multi Layer Perceptrons and human classification

Given the context of the described mathematical psychological models, it is interesting to study the behaviour of the MultiLayer Perceptron (MLP) in this aspect. The MLPs considered here have the sigmoid $1/(1+\exp(-x))$ as acti-

vation function for all hidden and output units. The MLP with M input units and N output units performs a non-linear mapping from an M-dimensional input space I to an N-dimensional output space O. If the N output units are properly scaled (normalized), these units can be interpreted to estimate the class probabilities $P(C_i|X)$, where $X \in I$ is the feature vector presented at the input side of the MLP (Richard & Lippmann, 1991). The shape of the class boundaries is known if the third step (the deterministic choice model) is represented by the winner-takes-all (WTA) strategy. Assuming WTA and no hidden layer, each class boundary (the class-to-complement boundary) is locally linear (such as in figure 1). Moreover, each class is convex. (Consequently, the classes such as in figure 1 cannot be trained by a 2-Layer Perceptron.) The class-to-class boundary (boundary between two *neighbouring* classes) is always linear. Such an MLP can fully be interpreted by the GCM by appropriate choice of exemplars and the euclidean metric. The MLP-results can also be interpreted with the GRT-paradigm, if one accepts *locally* linear instead of linear class boundaries.

The 3-Layer Perceptron shows non-linear class boundaries; classes may not be convex anymore. The class-to-class boundaries have a shape of the form shown in figure 2. The mathematical expression of such boundaries is

$$\frac{\alpha_1}{1+\exp(-L_1(X))} + \ldots + \frac{\alpha_H}{1+\exp(-L_H(X))} + \beta = 0$$

in which α_i and β and the L_i are trained by the MLP. In figure 2 possible class-to-class boundaries in two dimensions are shown. They indicate the 'rounding off' of locally linear surfaces. The class-to-complement boundaries have locally the shape of the class-to-class boundaries, so they may be quite complex indeed. These complex class boundaries cannot be modelled by the GCM with only a small number of exemplars per class, and also the interpretation in terms of the GRT is only possible if one accepts highly complex class shapes. This is an undesirable situation.

There is a way out to save the MLP interpretation in terms of GCM and GRT. With respect to the GCM, a solution is to relax the constraint of an exemplar being a specific point in the input space in two ways: by allowing an exemplar to be outside the input space I, and by allowing a set of exemplars to be an infinite subset of a class. An exemplar can be outside I, for example if one desires to determine the class 'tall' versus 'small' in a group of persons ordered with respect to length. The exemplar of the group 'tall' is a direction rather than a specific tall person, i.e. the class of 'tall' persons is unbounded. For every MLP such unbounded classes exist. Unbounded classes occur generally in classification.

The second relaxation is the infinite set of exemplars. Infinite exemplar sets conceptually exist for 3-Layer Perceptrons and can be constructed as follows. The mapping of a 3-Layer Perceptron is represented including the hidden space H: $I \rightarrow H \rightarrow O$. The WTA-strategy on $H \rightarrow O$ yields classes in H with locally linear class boundaries and linear

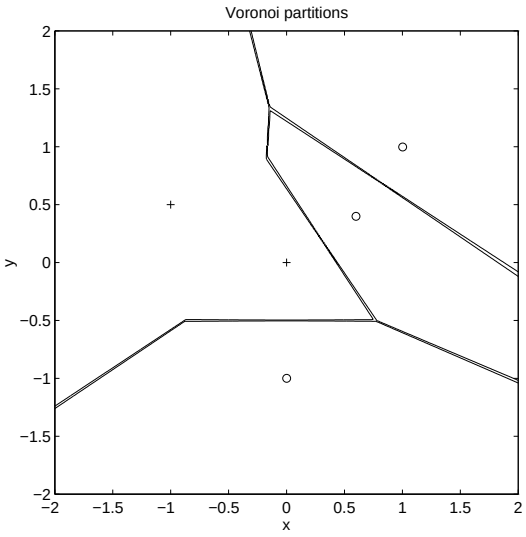

Figure 1: Separation of four classes by locally linear class-to-complement boundaries. One class has two exemplars, indicated with crosses; the other classes have one exemplar (denoted by circles). The close curves approximate the exact equi-class probability contours, which is the union of all class-to-complement boundaries. By construction these classes are GCM-trainable.

class-to-class boundaries. In H exemplars can be found given an euclidean metric for these classes (such as in the 2 layer case). Under the mapping $I \xrightarrow{\phi} H$ these exemplars e_j have inverse images $\phi^{-1}(e_j)$. The dimension of these inverse images is in general $\dim(I) - \dim(H)$ (the number of input units minus the number of hidden units) if $\dim(I) - \dim(H) \geq 0$, while if $\dim(I) < \dim(H)$ these images do not necessarily exist, i.e. they may mark a tendency as in the case described above. The required metric in I is complex but independent of the class.

The interpretation of the MLP results in terms of the GRT (the acceptance of higher order class boundaries) can be based on two observations: (1) the exact form of class boundaries also depend on the representation and is itself not a good measure for the complexity of the classification problem, and (2) the human stimulus representation involves a cascade of mappings from the actual stimulus to its representation X in I. Consequently the shape of the class boundaries in the classification problem is the result of many other preceding transformations, and mathematical boundary complexity is not a good measure for perceptual boundary complexity.

3. Error measures

Since the output of the MLP (especially the 2-LP) after training can straightforwardly be interpreted in terms of GCM and to a smaller extent of GRT, the type of error criterion used in training deserves more interest. Here we will discuss four types of error function that each has a specific

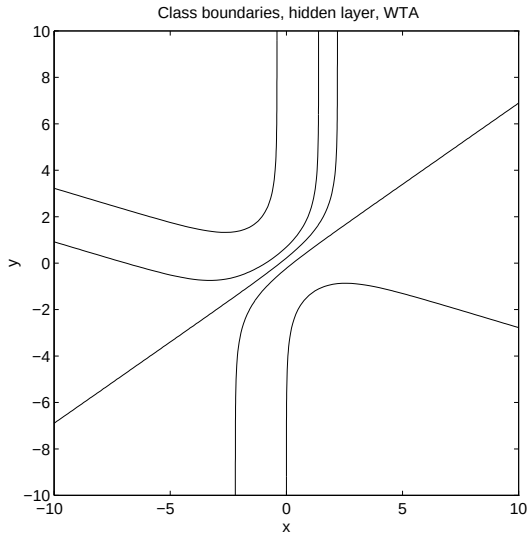

Figure 2: Examples (dimension 2) of class-to-class boundaries in case of a 3-Layer Perceptron assuming the winner-takes-all (WTA) strategy. All class boundaries are 'rounded-off' versions of the locally linear class boundaries of the type as shown in figure 1.

statistical interpretation. The notation will be as follows: X and $\tilde{Y}$ will denote the input and (normalized) output vector of the MLP of dimension M and N respectively. Each stimulus is presented K times. The symbol Y denotes the desired probability vector, observed during training.
The training data can be represented in different ways. For example, if the seven responses on a stimulus are $C_2, C_4, C_4, C_2, C_1, C_4, C_4$ respectively, training data can be described by counts: $(1, 2, 0, 4)$, or by probabilities: $(0.14, 0.29, 0.00, 0.57)$, or by events: $(0, 1, 0, 0), \ldots, (0, 0, 0, 1)$. In the event-representation, one emphasizes a correct class response for each stimulus, whereas the probabilistic representation focusses on the match of the observed and desired class probability vectors. During training, the connections of the MLP are adjusted so as to optimize E_{tot}

$$E_{tot} = \sum_{\forall X} E(X)$$

in which $E(X)$ is a function measuring the difference between the predicted ($\tilde{Y}$) and desired (Y) output. Each stimulus has equal weight in the total error. We normalize Y and $\tilde{Y}$ and write $Y = (y_1, \ldots, y_N)$ and $\tilde{Y} = (\tilde{y}_1, \ldots, \tilde{y}_N)$. $\sum y_i = \sum \tilde{y}_i = 1$. The different types of error criteria are based on: (1) the least squared error (LSE), (2) the log likelihood ratio (LLR); (3) the log likelihood (LL); and (4) the likelihood ratio (LR). We will now discuss the four error types in combination with the different types for database representation.

Case LSE
If the MLP is trained to minimize the LSE-error, then it is a well-known result that the optimum MLP is independent of the representation of the data. This is even the case for Mahalanobis distances in general. This can be seen as follows. Let $Y_1, \ldots, Y_K$ be the K (0, 1) response vectors in the training set for a stimulus X. For general Mahalanobis matrix G,

$$E_{tot} = \sum_X \sum_i (\tilde{Y} - Y_i)^t G (\tilde{Y} - Y_i)$$

and

$$\sum_X K \left(\tilde{Y} - \frac{\sum_i Y_i}{K} \right)^t G \left(\tilde{Y} - \frac{\sum_i Y_i}{K} \right)$$

lead to the same optimal MLP-solution, since both expressions differ only a constant independent of $\tilde{Y}$.

Case LL
Let k_i denote the number of responses C_i for the stimulus X, and $\sum k_i = K$. In fact, $k_i = K y_i$. For the log likelihood, we have per stimulus X in case of the probabilistic representation (to be optimized)

$$E(X) = \log \left(\frac{K!}{k_1! \cdots k_N!} \right) + \sum_i y_i K \log(\tilde{y}_i)$$

and in case of the event-representation (to be optimized)

$$E(X) = \sum_{\forall Y} \log(\tilde{y}_{i_{max}})$$

where $i_{max} = argmax_i k_i$ (the value of i for which $k_i = 1$) depends only on Y. $\sum_{\forall Y}$ indicates the sum over all responses Y to X. Since the term $\sum_i y_i K \log(\tilde{y}_i)$ exactly collects the right terms as are added in the expression $\sum_{\forall Y} \log(\tilde{y}_{i_{max}})$ (note that $k_i = y_i K$), these expressions differ a constant independent on $\tilde{Y}$, and henceforth minimizations lead to the same MLP-solutions.

Case LLR
In this case, we have to compare

$$E(X) = \sum_i y_i K \left(\log(\tilde{y}_i) - \log(y_i) \right)$$

with

$$E(X) = \sum_{\forall Y} \log(\tilde{y}_{i_{max}})$$

and since $\sum_i y_i K \log(y_i)$ is independent on $\tilde{Y}$, we again get identical optimization results.

Both the LLR and LL error measure have the property of resulting in a high error for the entire training set due to just one bad response. In modelling, this may be an undesirable property, depending on the desired statistical properties of the classifier. If one aims to model the responses for all X simultaneously, the LL and LLR are appropriate error measures that can be used. If one aims at optimization of the average match between the observed (desired) and predicted probability vectors, the following measure LR may be an alternative.

Case LR

This measure has been used successfully by Smits & ten Bosch (1996) to model the response behaviour of subjects in a (/p/, /t/, /k/)-classification experiment. In contrast with the LLR and LL, the LR does not show the property of resulting in a low score for the whole training set due to just one bad response.

It is straightforward to see that in this case the event-representation and the probabilistic representation do not yield an identical MLP. In the probabilistic representation we obtain

$$E(X) = \prod_i \frac{\tilde{y}_i^{Ky_i}}{y_i^{Ky_i}}$$

which yields the optimization of the match between predicted and observed probability *vectors*. The event-representation would read

$$E(X) = \sum_{\forall Y} \tilde{y}_{i\,max}$$

leading to the optimization of the probability of the correct class, given a randomly drawn stimulus X.

In the first case, the statistical properties of $E_{tot} = \sum_X E(X)$ are known if K is sufficiently large. It is well-known that for larger K the distribution of the LLR $\log(E(X))$ tends to a χ^2-distribution with $df = K - 1$ degrees of freedom, independent of X. Consequently, the distribution of $E(X)$ tends to a distribution independent of X. Applying the Central Limit theorem (Stuart & Ord, 1993, ch. 8), we conclude that E_{tot}/n_{stim} tends to be normally distributed, which is useful for cross validation techniques (see also Smits & ten Bosch, 1996).

4. Interpretation of the error measures in terms of GRT and GCM. Conclusion

The error measures can be interpreted within the context of GRT and GCM as follows. In general, the tranparancy of this interpretation depends on the topology of the MLP: the smaller the MLP, the more transparant the relation with the GRT and GCM. The MLP-solution can always be formally rewritten in terms of the GRT context. If the MLP contains no hidden layer, the GRT should allow locally linear class-to-complement boundaries; if there is a hidden layer the boundaries have a highly complex structure. The relation between MLP and GCM is the more explicit. If there is no hidden layer, the MLP results can be directly transformed in GCM-terms with a limited number of exemplars per class and the euclidean metric. If there is a hidden layer, the GCM should allow the concept of virtual exemplars (exemplars not in the data set itself, indicating 'directions' or 'tendencies'), and the presence of infinite exemplar sets.

In specific cases, the relation between the MLP-solution and GCM can be drawn more explicitly, for example in the LSE-case. LSE basically assumes a gaussian distribution of the reponses $(\tilde{y}_1, \ldots, \tilde{y}_N)$ with mean $(y_1, \ldots, y_N)$ and equal (diagonal) variance independent of X. If all variances are equal, all class-to-class boundaries are linear and the solution can be interpreted as modelled by a simple exemplar model. The LSE, however, is a bad measure if the separation between the classes is high in the training set, since the statistical assumptions (equal variances everywhere) do not hold for probabilities close to zero or one.

The LL and LLR model do a better job if the observed separation between the classes is high. Both LL and LLR are insensitive to the representation of the training data.

The forth measure (LR) is a useful alternative if statistical outlayers (stimuli with bad class prediction) would deteriorate the match of the entire stimulus set. Rather than focussing on the joint probability of predicting the correct class for all stimuli simultaneously, if deals with the average match of predicted and observed class. The LR, however, is sensitive to the representation of the training data. The likelihood ratio LR is suggested as an error measure that is useful in cases where one desires to optimize the average probabilistic match between desired and observed class probabilities. The reader is referred to Smits & ten Bosch (1996) and ten Bosch & Smits (forthcoming) for more details.

References

Ashby, F.G., and Townsend, J.T. (**1986**). 'Varieties of perceptual independence.' Psychological Review, vol. 93. pp. 154–179.

Maddox, W.T., and Ashby, F.G. (**1993**). 'Comparing decision bound and exemplar models of categorization.' Perception & Psychophysics, vol 53, pp. 49–70.

McKinley, S.C., and Nosofsky, R.M. (**1995**). 'Investigations of exemplar and decision bound models in large, ill-defined category structures.' Journal of Experimental Psychology: Human Perception and Performance, vol. 21. pp. 128–148.

Bosch, L.F.M. ten, and Smits, R. (*forthcoming*). 'On the cost function for fuzzy classification.' To be submitted to Journal of Mathematical Psychology.

Smits, R. and ten Bosch, L. (**1996**) 'The multi-layer perceptron as a model of human categorization behavior'. Submitted to Journal of Mathematical Psychology.

Nosofsky, R.M. (**1986**). 'Attention, similarity, and the identification-categorization relationship'. Journal of Experimental Psychology: General. Vol. 115, pp. 39–57.

Richard, M.D., and Lippmann, F.F. (**1991**). 'Neural network classifiers estimate Bayesian a posteriori probabilities.' Neural Computation 3, pp. 461 – 483.

Stuart, A., and Ord, J.K. (*1993*). *Kendall's advanced Theory of Statistics*. John Wiley & Sons, New York, Toronto.