SOME NOTES ON THE PROBABILISTIC SEMANTICS OF LOGISTIC FUNCTION PARAMETERS IN NEURAL NETWORKS.

Todd R. Davies. Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025 USA[1]

A common choice for the activation function that is computed at each node in a model neural network is one that is given by the logistic equation

$$\alpha_j = \frac{1}{1 + \exp[-(\sum_{i \in \mathbf{I_j}} w_{ij} x_i + \theta_j)/T]}$$

in which α_j is the activation value (in [0,1]) for the jth unit (node), $\mathbf{I_j}$ is the set $\{i_1, \ldots, i_n\}$ of inputs to j, w_{ij} is the weight (in $\Re$) connecting an input i to unit j, x_i is the state (usually binary or in [0,1]) of i, θ_j is the threshold or bias (in $\Re$) for j, and T is the temperature or gain parameter (in $\Re^+$, assumed herein to be 1). The activation value α_j is usually thought of as representing either the computed probability of the proposition or event ascribed to j as its content, or the probability that j, as a binary unit, will turn on (i.e., that its state x_j will be 1 rather than 0). The weight and threshold terms are often not given a clear probabilistic interpretation by those who use them in models, but intuitively a weight w_{ij} expresses the strength of evidence in favor of (or against) the proposition represented by x_j that is provided by $x_i = 1$, and the threshold θ_j expresses the (positive or negative) tendency for x_j to go to 1 in the absence of any of the inputs being on or "true". If α_j is identified with the probability of an event or proposition B_j (which could be identical to the event $x_j = 1$) given the input state vector (or vector of truth values) $\mathbf{x} = (x_{i_1}, \ldots, x_{i_n})$ then the logit transformation of the probability $p(B_j \mid \mathbf{x})$ defines an alternative form of the logistic equation (with $T = 1$), namely,

$$\text{logit}[p(B_j \mid \mathbf{x})] = \ln \frac{p(B_j \mid \mathbf{x})}{1 - p(B_j \mid \mathbf{x})} = \sum_{i \in \mathbf{I_j}} w_{ij} x_i + \theta_j .$$

If we are to interpret networks with logistic units in probabilistic terms, or if we wish to know how to translate probability statements into parameter values for such networks, then two questions that arise are (1) How can activation values, weights, and thresholds be expressed as functions of probabilities?, and (2) What assumptions about probabilities are embodied in the logistic model? Little if any attention appears to have been paid to these problems by statisticians, probably because logistic regression models are used primarily for prediction and the exact analysis of coefficients is not of great interest (Cox, 1970). Hinton & Sejnowski (1983) discuss possible definitions for the weight and threshold parameters for the case of one input state x_i, under the assumptions that α_j represents $p(B_j)$, and that $x_i = 1$ represents the presence of input evidence A_i, and $x_i = 0$ represents its absence $\neg A_i$. The definitions on which they settle, for the restricted case of $\mid \mathbf{I_j} \mid = 1$, amount to the following:

$$\theta_j = \text{logit}[p(B_j)] + \sum_{i \in \mathbf{I_j}} \ln \frac{p(\neg A_i \mid B_j)}{p(\neg A_i \mid \neg B_j)}, \quad w_{ij} = \ln \frac{p(A_i \mid B_j)}{p(A_i \mid \neg B_j)} - \ln \frac{p(\neg A_i \mid B_j)}{p(\neg A_i \mid \neg B_j)}.$$

A similar semantics has been proposed by Geffner & Pearl (1987). A very useful feature of these definitions, as Hinton and Sejnowski point out, is that the function for weights is symmetric in A_i and B_j, implying that $w_{ij} = w_{ji}$. The semantics therefore provides a justification for the symmetry constraints often imposed by network architectures.

The definitions given above may be difficult to understand intuitively. However, if we extend the analysis of Hinton and Sejnowski to the case of multiple inputs, then an equivalent form can be given for these equations that defines θ_j as a function of one probability value, and requires just one additional probability value to be specified for each w_{ij}, viz,

$$\theta_j = \text{logit}[p(B_j \mid \neg A_{i_1}, \ldots, \neg A_{i_n})], \quad w_{ij} = \text{logit}[p(B_j \mid A_i \text{ and } \neg A_k \text{ for all } k \text{ s.t. } k \in \mathbf{I_j},\ k \neq i)] - \theta_j .$$

The parameters can thus be defined as functions of the probability of a hypothesis given evidence, rather than the reverse, and this may be easier to think about. Because the logit mapping is one-to-one we can recover these probabilities from the parameter values. Extension to multiple inputs also allows us to prove the following: *For θ_j and w_{ij} defined as above, $p(B_j \mid \mathbf{x})$ can be computed as a logistic function of $\sum_{i \in \mathbf{I_j}} w_{ij} x_i + \theta_j$ for each input vector $\mathbf{x}$ iff*

$$p(\mathbf{x} \mid B_j) = \prod_{i \in \mathbf{I_j}} p(x_i \mid B_j) \text{ and } p(\mathbf{x} \mid \neg B_j) = \prod_{i \in \mathbf{I_j}} p(x_i \mid \neg B_j).$$

In words, assumptions of conditional independence among the components of $\mathbf{x}$ for both B_j and $\neg B_j$ are necessary and sufficient conditions for exact representation of the probability model in a logistic function. From the definitions given above for w_{ij} and θ_j one can derive a system of linear equations that significantly overdetermine the parameters and probabilities. When the conditional independence assumptions are violated, error-minimization (e.g. orthogonalization, least-squares) can be applied to the system to determine an optimal interpretation or assignment for the parameters. The analysis of violations suggests how much error-correction, in the form of informational redundancy, should be built into the network to achieve high accuracy.

References:

Cox, D. R. (1970) *The Analysis of Binary Data*, Metheun & Co. Ltd., London.

Geffner, H. & Pearl, J. (1987) On the probabilistic semantics of connectionist networks, *IEEE First International Conference on Neural Networks*, San Diego, June 21-24, 1987, pp. II-187-195.

Hinton, G. E. & Sejnowski, T. J. (1983) Optimal perceptual inference, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, D.C., June 19-23, 1983, pp. 448-453.

[1] Also affiliated with the Department of Psychology, Stanford University.