# Neural Network Regression

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

# Neural network form and a toy example

A multi-layer feed-forward neural network is a nonlinear map of $\mathbf{x} \in \Re^p$ to one or more outputs through functions of linear combinations of functions of linear combinations of ... of functions of linear combinations of coordinates of $\mathbf{x}$.

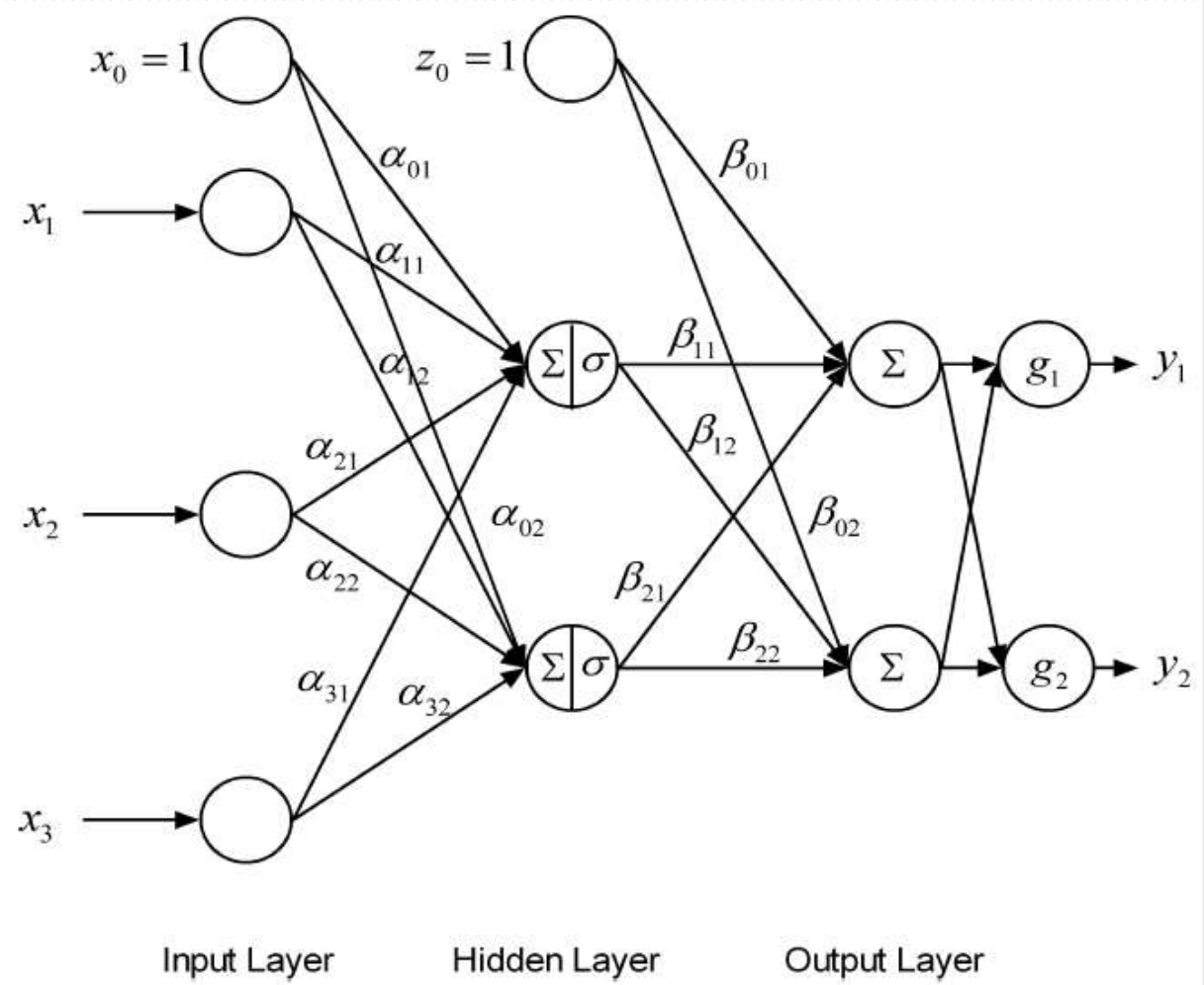
The next slide represents a single hidden layer feed-forward neural net with 3 inputs 2 hidden nodes and 2 outputs. It stands for a function of $\mathbf{x}$ defined by setting

$$z_1 = \sigma\left(\alpha_{01} \cdot 1 + \alpha_{11} x_1 + \alpha_{21} x_2 + \alpha_{31} x_3\right)$$
$$z_2 = \sigma\left(\alpha_{02} \cdot 1 + \alpha_{12} x_1 + \alpha_{22} x_2 + \alpha_{32} x_3\right)$$

and then

$$y_1 = g_1\left(\beta_{01} \cdot 1 + \beta_{11} z_1 + \beta_{21} z_2, \beta_{02} \cdot 1 + \beta_{12} z_1 + \beta_{22} z_2\right)$$
$$y_2 = g_2\left(\beta_{01} \cdot 1 + \beta_{11} z_1 + \beta_{21} z_2, \beta_{02} \cdot 1 + \beta_{12} z_1 + \beta_{22} z_2\right)$$

# Toy example network representation



# Functional forms employed

In SEL/regression contexts, identity functions (dependent upon a single argument, rather than multiple ones of them) are common chocies for the $g$s. In classification problems, in order to make estimates of class probabilities, the functions $g$ often exponentiate one argument and divide by a sum of such terms for all arguments. Originally, the most common choice of $\sigma$ at hidden nodes was the (sigmoidal-shaped) logistic function

$$\sigma(u) = \frac{1}{1 + \exp(-u)}$$

or the (completely equivalent in this context) hyperbolic tangent function

$$\sigma(u) = \tanh(u) = \frac{\exp(u) - \exp(-u)}{\exp(u) + \exp(-u)}$$

These functions are differentiable at $u = 0$, so that for small $\alpha$'s the functions of $\mathbf{x}$ entering the $g$s in a single hidden layer network are nearly linear. For large $\alpha$s the functions are nearly step functions.

# Functional forms employed: ReLUs

More recently, sigmoidal forms for the activation function have declined in popularity. Instead, the hinge or positive part function

$$\sigma(u) = \max(u, 0) = u_+$$

is often used. In common parlance, this makes the hidden nodes "rectified linear units" (ReLUs). Note that this choice makes functions of **x** entering an output layer piece-wise linear and continuous (not at all an unreasonable form).

# Consequences of functional forms

In light of the nature of the forms used for $\sigma(u)$, it is not surprising that there are universal approximation theorems that guarantee that any continuous function on a compact subset of $\Re^p$ can be approximated to any degree of fidelity with a single layer feed-forward neural net with enough nodes in the hidden layer. This is both a blessing and a curse. It promises that these forms are quite flexible It also promises that there must be both over-fitting and identifiability issues inherent in their use (the latter in addition to the identifiability issues already inherent in the symmetric nature of the functional forms assumed for the predictors). Typically some form of regularization must be used to mitigate the over-fitting possibility.