Abstract

Neural Networks as a way to specify nonparametric regression and classification models.

Feed Forward Neural Nets Models

Feed forward Neural Nets are also known as multilayer perceptrons or backpropagation networks. The figure shows a network with a layer of 4 hidden units.

The outputs are computed from the following formulas,

\[ g_k(x) = b_k + \sum_j v_{jk} h_j(x) \]  \hspace{1cm} (1)

\[ h_j(x) = \tanh(a_j + \sum_i u_{ij} x_i) \]  \hspace{1cm} (2)

where \( \{a_j\}, \{b_k\}, \{u_{ij}\}, \{v_{jk}\} \) are the parameters of the network. The parameters with one index are known as biases and those with two indices are known as weights. We assume that \( x = (x_1, \ldots, x_d) \in \mathbb{R}^d, h(x) = (h_1(x), \ldots, h_l(x)) \in \mathbb{R}^l, g(x) = (g_1(x), \ldots, g_p(x)) \in \mathbb{R}^p \). The hyperbolic tangent,

\[ \tanh(z) = \frac{\sinh(z)}{\cosh(z)} = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{1 - e^{-2z}}{1 + e^{-2z}} \]

is an example of a sigmoid function. A sigmoid is a non-linear function, \( s(z) \), that goes through the origin, approaches \(+1\) as \( z \to \infty \) and approaches \(-1\) as \( z \to -\infty \).
Figure 2: Hyperbolic Tangent

It is known since 1989 (only) that as the number of hidden units increases, any function defined on a compact set can be approximated by linear combinations of sigmoids.

Multilayer perceptrons are often used as flexible models for nonparametric regression and classification. Given data,

\[(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\]

with,

\[y^{(k)} = g(x^{(k)}, \theta) + e^{(k)}\]

where

\[e^{(1)}, e^{(2)}, \ldots, e^{(n)}\] are iid with \(Ee^{(k)} = 0\)

Hence, the \(g\) is the regression of \(y\) on \(x\), i.e.,

\[E(y|x, \theta) = g(x, \theta) \text{ with } \theta \in \Theta\]

The multilayer perceptrons provide a practical way to define the functions \(g\) with high dimensional parameter spaces \(\Theta\). We take \(\theta = \{\{a_j\}, \{b_k\}, \{u_{ij}\}, \{v_{jk}\}\}\). The objective is to find the predictive distribution of a new target vector \(y\), given the examples \(D = ((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)}))\) and the new vector of inputs \(x\), i.e.,

\[f(y|x, D) = \int f(y|x, \theta) \pi(\theta|D) d\theta\]

Under the assumption of quadratic loss, the best guess for \(y\) will be its mean,

\[\hat{y} = E(y|x, D) = \int g(x, \theta) \pi(\theta|D) d\theta\]
These estimates can be approximated by MCMC by sampling \( \theta^{(1)}, \ldots, \theta^{(N)} \) from the posterior and then computing empirical averages,

\[
\hat{g}_N = \frac{1}{N} \sum_{j=1}^{N} g(x, \theta^{(j)})
\]

**Useful Priors on Feed Forward Networks**

In the absence of specific information, the following assumptions about the prior \( \pi(\theta) \) are reasonable,

1. The components of \( \theta \) are independent and symmetric about 0.
2. Parameters of the same kind have the same a priori distributions, i.e.,
   \[
   a_1, a_2, \ldots \text{ id} \\
   b_1, b_2, \ldots \text{ id} \\
   u_{1j}, u_{2j}, \ldots \text{ iid for all } j \\
   v_{1k}, v_{2k}, \ldots \text{ iid for all } k
   \]

With these assumptions if \( \text{var}_x(v_{jk}) = l^{-1} \sigma_v^2 < \infty \) then by the Central Limit Theorem, as \( l \to \infty \) the prior on the output units converges to a Gaussian process. Gaussian processes are characterized by their covariance functions and they are often considered inadequate for modeling complex inter-dependence of the outputs. To avoid the Gaussian trap, it is convenient to use a priori distributions for the components of \( \theta \) that have infinite variance.

A practical choice (used by Neal) is to take,

\[
v_{jk} \text{ as t-distribution } \propto \left(1 + \frac{v_{jk}^2}{\alpha \sigma_v^2}\right)^{-\frac{(\alpha+1)/2}{2}} \text{ with } 0 < \alpha < 2
\]

Furthermore, if we take \( \sigma_v = w_r l^{-1/\alpha} \) then the resulting prior will converge as \( l \to \infty \) to a symmetric stable process of index \( \alpha \).

Recall that \( Z_1, Z_2, \ldots, Z_n \) iid with distribution symmetric about 0 are said to be stable of index \( \alpha \) if

\[
\frac{Z_1 + \ldots + Z_n}{n^{1/\alpha}} \text{ has the same law as } Z_1
\]

A distribution is said to be in the domain of attraction of a stable law if properly normalized sums of independent observations from this distribution, converge in law to a stable distribution. Hence, distributions with finite variance are in the domain of attraction of the Gaussians. It is also well known that distributions with tails going to zero as \( |x|^{-\alpha+1} \) as \( |x| \to \infty \) are in the domain of attraction of stable laws of index \( \alpha \) which justifies the choice of t-dist above.
Postulating an Energy for the Net

An alternative approach without priors (apparently...) is to postulate directly and Energy function for the network, e.g.,

\[ E(\theta, \gamma) = \frac{1}{n} \sum_{k=1}^{n} L(y^{(k)}, g(x^{(k)}, \theta)) + \gamma \| \theta \|^2 \]

where \( L(y, z) \) is the assumed loss when we estimate \( y \) with \( z \). Typical choices are \( L(y, z) = R(||y - z||) \) for some nondecreasing function \( R \) and some norm \( || \cdot || \). Then choose \( \theta \) to minimize this Energy function. Often, the smoothness parameter \( \gamma > 0 \) is chosen by Cross-Validation or by plain trial and error.

For complicated multi modal energy functions, a combination of simulated annealing with a classical gradient method (such as conjugate gradients) have been the most successful.