Chapter 3

Methodology

3.1 Definition

The network system considered in this study is shown in Figure 3.1. It is a feed-forward neural network with one hidden layer and one output node.

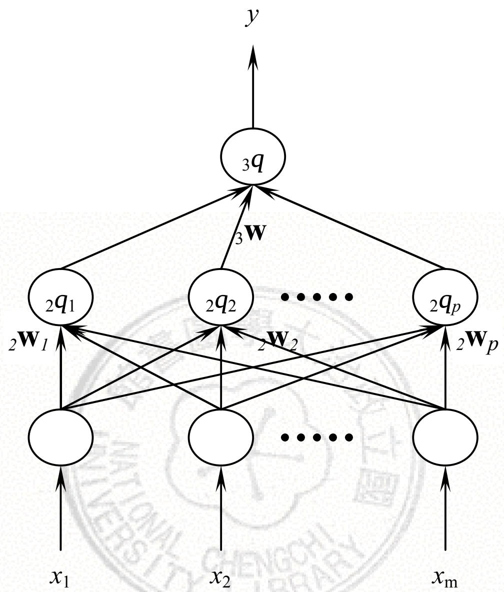

Figure 3.1: The ANN with one hidden layer and one output node.

In Figure 3.1, y denotes the output value of the neural network, and $\mathbf{x}^{\mathrm{t}} \equiv (x_1, x_2, \ldots, x_{\mathrm{m}})$ where x_i denotes the i-th outside stimulus input, with i from 1 to m. ${}_2\mathbf{w}_j^{\mathrm{t}} \equiv ({}_2w_{j1}, {}_2w_{j2}, \ldots, {}_2w_{jm})$ stands for the weights between the j-th hidden node and the input layer, with j from 1 to p, and ${}_3\mathbf{w}^{\mathrm{t}} \equiv ({}_3w_1, {}_3w_2, \ldots, {}_3w_p)$ stands for the weights between the output node and all hidden nodes. The following tanh(t) activation function is used in hidden nodes and the linear activation function is used in the output node.

$$\tanh(t) \equiv \frac{e^t - e^{-t}}{e^t + e^{-t}} \tag{3.1}$$

For the c-th input ${}_c\mathbf{x}$, the hidden node activation value ${}_ch_j$ and its output value ${}_cy$ are computed as follows:

$$ {}_c h_j = \tanh\left(\sum_{i=1}^{m} {}_2 w_{ji}\, {}_c x_i + {}_2 q_j\right) \tag{3.2} $$

$$ {}_c y = \sum_{j=1}^{p} {}_3 w_j\, {}_c h_j + {}_3 q \tag{3.3} $$

3.2 Method of Extracting Rules from Neural Networks

3.2.1 The Approximation of Hidden Node Activation Function

To extract comprehensible rules from the ANN with the tanh(t) activation function, we use the following function $g(t)$ to approximate tanh(t):

$$ g(t) \equiv \begin{cases} 1 & \text{if } t \geq k \\ b_1 t + b_2 t^2 & \text{if } 0 \leq t \leq k \\ b_1 t - b_2 t^2 & \text{if } -k \leq t \leq 0 \\ -1 & \text{if } t \leq -k \end{cases} \tag{3.4} $$

where $(b_1, b_2, k) \equiv \arg(\min_{b_1, b_2, k} \int_{-\infty}^{\infty} (\tanh(t) - g(t))^2 \mathrm{d}t$ subject to $b_1 k + b_2 k^2 = 1$). Then, using the numerical analysis of Sequential Quadratic Programming (The MathWorks, Inc. 2002), we obtain $b_1 \cong 1.0020101308531$, $b_2 \cong -0.251006075157012$, $k \cong 1.99607103795966$, and $\min_{b_1, b_2, k} \int_{-\infty}^{\infty} (\tanh(t) - g(t))^2 \mathrm{d}t \cong 0.00329781871956464$. Note that $g(t)$ is continuous at the boundaries of four regions (k £ t, 0 £ t £ k, $-k$ £ t £ 0, t £ $-k$), because we set $\lim_{t \to k^-} b_1 t + b_2 t^2 = 1$, $\lim_{t \to 0^+} b_1 t + b_2 t^2 = 0$, $\lim_{t \to 0^-} b_1 t - b_2 t^2 = 0$, and $\lim_{t \to -k^+} b_1 t - b_2 t^2 = -1$.

For the j-th hidden node, let $t_j \equiv {}_2\mathbf{w}_j^{t}\mathbf{x}$. Thus $\tanh(t_j + {}_2 q_j)$ can be approximated with $g(t_j + {}_2 q_j)$, which is defined by

$$ g(t_j + {}_2 q_j) = \begin{cases} 1 & \text{if } t_j \geq k - {}_2 q_j \\ (b_1\, {}_2 q_j + b_2\, {}_2 q_j^{\,2}) + (b_1 + 2 b_2\, {}_2 q_j) t_j + b_2 t_j^2 & \text{if } -{}_2 q_j \leq t_j \leq k - {}_2 q_j \\ (b_1\, {}_2 q_j - b_2\, {}_2 q_j^{\,2}) + (b_1 - 2 b_2\, {}_2 q_j) t_j - b_2 t_j^2 & \text{if } -k - {}_2 q_j \leq t_j \leq -{}_2 q_j \\ -1 & \text{if } t_j \leq -k - {}_2 q_j \end{cases} \tag{3.5} $$

In other words, for the j-th hidden node, the activation value is approximated with a form of single-variate polynomial in each of four separate regions in the t_j space. For example, if $-{}_2 q_j$ £ t_j £ $k - {}_2 q_j$, then $\tanh(t_j + {}_2 q_j)$ is approximated with $b_2\, {}_2 q_j^{\,2} + b_1\, {}_2 q_j + (b_1$

$+ 2\, b_2\, {}_2q_j)\, t_j + b_2 t_j^2$.

To better represent the condition, let's introduce some notations. Set i_j be 1, if the condition k - ${}_2q_j$ £ ${}_2\mathbf{w}_j^t\mathbf{x}$ holds; 2, if the condition $-{}_2q_j$ £ ${}_2\mathbf{w}_j^t\mathbf{x}$ £ k - ${}_2q_j$ holds; 3, if the condition $-k$ - ${}_2q_j$ £ ${}_2\mathbf{w}_j^t\mathbf{x}$ £ $-{}_2q_j$ holds; and 4, if the condition ${}_2\mathbf{w}_j^t\mathbf{x}$ £ $-k$ - ${}_2q_j$ holds. Also, set $\mathrm{w}_{j1} \equiv {}_2\mathbf{w}_j^t$, $\mathrm{w}_{j2} \equiv \begin{bmatrix} {}_2\mathbf{w}_j^t \\ -{}_2\mathbf{w}_j^t \end{bmatrix}$, $\mathrm{w}_{j3} \equiv \begin{bmatrix} {}_2\mathbf{w}_j^t \\ -{}_2\mathbf{w}_j^t \end{bmatrix}$, $\mathrm{w}_{j4} \equiv -{}_2\mathbf{w}_j^t$, $\mathrm{u}_{j1} \equiv k - {}_2q_j$, $\mathrm{u}_{j2} \equiv \begin{bmatrix} -{}_2q_j \\ -k + {}_2q_j \end{bmatrix}$,

$\mathrm{u}_{j3} \equiv \begin{bmatrix} -k - {}_2q_j \\ {}_2q_j \end{bmatrix}$, $\mathrm{u}_{j4} \equiv k + {}_2q_j$, $g_{j1}(t_j) \equiv 1$, $g_{j2}(t_j) \equiv (b_1\, {}_2q_j + b_2\, {}_2q_j^2) + (b_1 + 2\, b_2\, {}_2q_j)\, t_j + b_2 t_j^2$, $g_{j3}(t_j) \equiv (b_1\, {}_2q_j - b_2\, {}_2q_j^2) + (b_1 - 2\, b_2\, {}_2q_j)\, t_j - b_2\, t_j^2$, and $g_{j4}(t_j) \equiv -1$. Then, when the i_j-th condition $\mathrm{w}_{ji_j}\mathbf{x}$ ³ u_{ji_j} holds, the activation value of the j-th hidden node is approximated with $g_{ji_j}(t_j)$. Furthermore, $y' \equiv {}_3q + \sum_{j=1}^{p} {}_3w_j \tanh(t_j + {}_2q_j)$ is approximated with ${}_3q + \sum_{j=1}^{p} {}_3w_j\, g_{ji_j}(t_j)$.

Let $i \equiv [i_1, i_2, ..., i_p]$ with $i_j \in \{1, 2, 3, 4\}\ \forall j = 1, 2, ..., p$. Thus, the conditions associated with p hidden nodes can be expressed as ${}_1\mathbf{A}_i\mathbf{x}$ ³ ${}_1\mathbf{b}_i$, where

$$ {}_1\mathbf{A}_i \equiv \begin{bmatrix} \mathrm{w}_{1i_1} \\ \mathrm{w}_{2i_2} \\ \vdots \\ \mathrm{w}_{pi_p} \end{bmatrix} \tag{3.6} $$

$$ {}_1\mathbf{b}_i \equiv \begin{bmatrix} \mathrm{u}_{1i_1} \\ \mathrm{u}_{2i_2} \\ \vdots \\ \mathrm{u}_{pi_p} \end{bmatrix} \tag{3.7} $$

For example, the condition [$-{}_2q_j$ £ ${}_2\mathbf{w}_j^t\mathbf{x}$ £ $k - {}_2q_j\ \forall j = 1, 2, ..., p$] can be expressed as ${}_1\mathbf{A}_i\mathbf{x}$ ³ ${}_1\mathbf{b}_i$ with $i_j = 2$ for every j.

In addition, the independent variables may have some extra constraints corresponding to the application. The extra constraints are usually linear as follows.

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{im}x_m \geq b_{2i},\ i = 1, 2, \ldots, n_2 \tag{3.8}$$

Let

$${}_2\mathbf{A} \equiv \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_2 2} & a_{n_2 2} & \cdots & a_{n_2 m} \end{bmatrix} \tag{3.9}$$

$${}_2\mathbf{b} \equiv \begin{bmatrix} b_{21} \\ b_{22} \\ \vdots \\ b_{2n_2} \end{bmatrix} \tag{3.10}$$

Thus the extra constrains are expressed as

$${}_2\mathbf{A}\,\mathbf{x}\ {}^{3}\ {}_2\mathbf{b} \tag{3.11}$$

Therefore, the full constrains associated with the i-th region are

$$\mathbf{A}_i\,\mathbf{x}\ {}^{3}\ \mathbf{b}_i \tag{3.12}$$

where $\mathbf{A}_i \equiv \begin{bmatrix} {}_1\mathbf{A}_i \\ {}_2\mathbf{A} \end{bmatrix}$ and $\mathbf{b}_i \equiv \begin{bmatrix} {}_1\mathbf{b}_i \\ {}_2\mathbf{b} \end{bmatrix}$.

In sum, when the approximation is applied to a layered feed-forward neural network, there are 4^p separate regions in the input space where the corresponding output value y' is approximated in a form of multivariate polynomial. The i-th region is $\{\mathbf{x}|\ \mathbf{A}_i\,\mathbf{x}\ {}^{3}\ \mathbf{b}_i\}$, and its associated y' equates ${}_3q + \sum_{j=1}^{p} {}_3w_j\, g_{ji_j}(t_j)$. In other words, there is a rule associated with each separate region in the input space:

$$\text{If } \mathbf{A}_i\,\mathbf{x} \geq \mathbf{b}_i, \text{ then } y' = {}_3q + \sum_{j=1}^{p} {}_3w_j\, g_{ji_j}(t_j) \tag{3.13}$$

$\mathbf{A}_i\,\mathbf{x}\ {}^{3}\ \mathbf{b}_i$ is a convex polyhedral set in the input space because $\mathbf{A}_i\,\mathbf{x}\ {}^{3}\ \mathbf{b}_i$ consists of linear inequality constraints. Furthermore, $\mathbf{A}_i\,\mathbf{x}\ {}^{3}\ \mathbf{b}_i$ has a feasible solution if and only if the linear programming (LP) problem (3.14) has an optimal solution.

$$\begin{aligned} &\text{Minimize: } \textit{constant} \\ &\text{Subject to: } \mathbf{A}_i\,\mathbf{x}\ {}^{3}\ \mathbf{b}_i \end{aligned} \tag{3.14}$$

If equation (3.14) has an optimal solution, then the corresponding rule exists.

Otherwise, the rule fails to exist.

3.2.2 The Differential Analysis of Rules

Since $t_j \equiv {}_2\mathbf{w}_j^{\mathrm{t}}\,\mathbf{x} = \sum_{i=1}^{m} {}_2w_{ji}x_i$ and

$$g_j(t_j) = \begin{cases} 1 & \text{if } t_j \ge k - {}_2q_j \\ (b_1\,{}_2q_j + b_2\,{}_2q_j^{\,2}) + (b_1 + 2b_2\,{}_2q_j)t_j + b_2 t_j^{\,2} & \text{if } -{}_2q_j \le t_j \le k - {}_2q_j \\ (b_1\,{}_2q_j - b_2\,{}_2q_j^{\,2}) + (b_1 - 2b_2\,{}_2q_j)t_j - b_2 t_j^{\,2} & \text{if } -k - {}_2q_j \le t_j \le -{}_2q_j \\ -1 & \text{if } t_j \le -k - {}_2q_j \end{cases} . \tag{3.15}$$

Thus

$$\frac{\partial g_j(t_j)}{\partial x_k} = \begin{cases} 0 & \text{if } t_j > k - {}_2q_j \\ {}_2w_{jk}(b_1 + 2b_2\,{}_2q_j) + 2\,{}_2w_{jk}\,b_2 t_j & \text{if } -{}_2q_j < t_j < k - {}_2q_j \\ {}_2w_{jk}(b_1 - 2b_2\,{}_2q_j) - 2\,{}_2w_{jk}\,b_2 t_j & \text{if } -k - {}_2q_j < t_j < -{}_2q_j \\ 0 & \text{if } t_j < -k - {}_2q_j \end{cases} \tag{3.16}$$

$$\frac{\partial^2 g_j(t_j)}{\partial x_l \partial x_k} = \begin{cases} 0 & \text{if } t_j > k - {}_2q_j \\ 2\,{}_2w_{jl}\,{}_2w_{jk}\,b_2 & \text{if } -{}_2q_j < t_j < k - {}_2q_j \\ -2\,{}_2w_{jl}\,{}_2w_{jk}\,b_2 & \text{if } -k - {}_2q_j < t_j < -{}_2q_j \\ 0 & \text{if } t_j < -k - {}_2q_j \end{cases} \tag{3.17}$$

$$\frac{\partial^3 g_j(t_j)}{\partial x_r \partial x_l \partial x_k} = 0 \tag{3.18}$$

where $r, l, k = 1, 2, ..., m$. $y' = {}_3q + \sum_{j=1}^{p} {}_3w_j\, g_j(t_j)$. Thus

$$\frac{\partial y'}{\partial x_k} = \sum_{j=1}^{p} {}_3w_j \frac{\partial g_j(t_j)}{\partial x_k} \tag{3.19}$$

$$\frac{\partial^2 y'}{\partial x_l \partial x_k} = \sum_{j=1}^{p} {}_3w_j \frac{\partial^2 g_j(t_j)}{\partial x_l \partial x_k} \tag{3.20}$$

$$\frac{\partial^3 y'}{\partial x_r \partial x_l \partial x_k} = \sum_{j=1}^{p} {}_3w_j \frac{\partial^3 g_j(t_j)}{\partial x_r \partial x_l \partial x_k} = 0 \tag{3.21}$$

For example, for the i-th region with $i_j = 2\ \forall\ j = 1, 2, ..., p$, $y' = {}_3q + \sum_{j=1}^{p} {}_3w_j\, g_{ji_j}(t_j)$,

and

$$\frac{\partial y'}{\partial x_k} = \sum_{j=1}^{p} {}_3w_j \frac{\partial g_j(t_j)}{\partial x_k} = \sum_{j=1}^{p} {}_3w_j\,{}_2w_{jk}\,(b_1 + 2b_2\,{}_2q_j^2) + \sum_{j=1}^{p} 2\,{}_3w_j\,{}_2w_{jk}\,b_2\,\mathbf{w}_j^{\mathrm{t}}\,\mathbf{x} \tag{3.22}$$

$$\frac{\partial^2 y'}{\partial x_l \partial x_k} = \sum_{j=1}^{p} {}_3w_j \frac{\partial^2 g_j(t_j)}{\partial x_l \partial x_k} = \sum_{j=1}^{p} 2\ {}_3w_j\,{}_2w_{jl}\,{}_2w_{jk}\,b_2 \tag{3.23}$$

$$\frac{\partial^3 y'}{\partial x_r \partial x_l \partial x_k} = 0 \tag{3.24}$$

If $\left.\frac{\partial y'}{\partial x_k}\right|_{\mathbf{x} \in \{\mathbf{x} | \mathbf{A}_i \mathbf{x} \geq \mathbf{b}_i\}}$ ≥ 0, the optimal solution of the LP problem (3.25) shall be greater than zero. Similarly, if $\left.\frac{\partial y'}{\partial x_k}\right|_{\mathbf{x} \in \{\mathbf{x} | \mathbf{A}_i \mathbf{x} \geq \mathbf{b}_i\}}$ < 0, the optimal solution of the LP problem (3.26) shall be less than zero.

$$\text{Minimize: } \frac{\partial y'}{\partial x_k}$$

$$\text{Subject to: } \mathbf{A}_i \mathbf{x} \ 3 \ \mathbf{b}_i \tag{3.25}$$

$$\text{Maximize: } \frac{\partial y'}{\partial x_k}$$

$$\text{Subject to: } \mathbf{A}_i \mathbf{x} \ 3 \ \mathbf{b}_i \tag{3.26}$$

If $\frac{\partial y'}{\partial x_k}$ is a linear equation, we can adopt the Simplex method to solve LP problems (3.25) and (3.26). Such LP problems can analyze if $\frac{\partial y'}{\partial x_k}$ is great or less than zero for every point in the region, $\mathbf{A}_i \mathbf{x} \ 3 \ \mathbf{b}_i$, without any dataset.

Note that the differentiations of y' are not defined at $t_j = -k + {}_2q_j, {}_2q_j$, or $k + {}_2q_j$. Also, since $\frac{\partial^3 y'}{\partial x_r \partial x_l \partial x_k}$ always equals zero, this approximation loses the information of higher order differentials.

After differential analyses, we can derive (differential) features via applying the sign test on our extracted rules. Take as illustration of the sign test of the relationship between y' and x_k. Let the significance level of the test be a (generally equals 0.05 or 0.01). The null hypothesis H_0 is that there is not a relationship between y' and x_k., while the alternative hypothesis H_1 is that there is a negative relationship between y'

and x_k. (that is, $\frac{\partial y'}{\partial x_k} < 0$). If H_0 is true, the conditional probability Pr($\frac{\partial y'}{\partial x_k} < 0 \mid H_0$) equals 0.5. If H_1 is true, the conditional probability Pr($\frac{\partial y'}{\partial x_k} < 0 \mid H_1$) is great than 0.5. Let n^{-} be the count of the maximal value of $\frac{\partial y'}{\partial x_k}$ associated with n_e (the number of extracted rules) LP problems stated in (3.27) that are less than 0.

$$\text{Maximize: } \frac{\partial y'}{\partial x_k}$$

$$\text{Subject to: } \mathbf{A}_i\ \mathbf{x} \ge \mathbf{b}_i \tag{3.27}$$

If H_0 is true, then n^{-} has a binomial distribution, $b(n_e, 0.5)$. We reject H_0 and accept H_1 at significant a if only if n^{-} is greater than n_0^{-}, where $\sum_{n^{-}=n_0^{-}}^{n_e} C_{n^{-}}^{n_e} 0.5^{n^{-}} \le a$.

3.2.3 The Rule Extraction Process

We summarize the rule extraction process in the Table 3.1.

Table 3.1: The Rule Extraction Process

1. Give a trained ANN with tanh(t) activation functions in hidden layer.

$$y = \sum_{j=1}^{p} {}_3w_j \tanh(\sum_{i=1}^{m} {}_2w_{ji} x_i + {}_2q_j) + {}_3q$$

2. Use $g_j(t_j)$ to approximation the $\tanh(t_j + {}_2q_j)$.

$$g_j(t_j)=\begin{cases} 1 & if\ t_j \ge k - {}_2q_j \\ (b_1\, {}_2q_j + b_2\, {}_2q_j^{\,2}) + (b_1 + 2b_2\, {}_2q_j)t_j + b_2 t_j^{\,2} & if\ -{}_2q_j \le t_j \le k - {}_2q_j \\ (b_1\, {}_2q_j - b_2\, {}_2q_j^{\,2}) + (b_1 - 2b_2\, {}_2q_j)t_j - b_2 t_j^{\,2} & if\ -k - {}_2q_j \le t_j \le -{}_2q_j \\ -1 & if\ t_j \le -k - {}_2q_j \end{cases}$$

where $t_j \equiv {}_2\mathbf{w}_j^{\,t}\,\mathbf{x} = \sum_{i=1}^{m} {}_2w_{ji} x_i$, $b_1 \cong 1.0020101308531$, $b_2 \cong -0.251006075157012$, and $k \cong 1.99607103795966$.

3. Give the extra conditions as follows:

$$\mathbf{A}_2\,\mathbf{x}\ {}^{3}\ \mathbf{b}_2$$

Then, the full constrains associated with the i-th region are

$$\mathbf{A}_i\,\mathbf{x}\ {}^{3}\ \mathbf{b}_i$$

where $\mathbf{A}_i \equiv \begin{bmatrix} \mathbf{A}_{1i} \\ \mathbf{A}_2 \end{bmatrix}$ and $\mathbf{b}_i \equiv \begin{bmatrix} \mathbf{b}_{1i} \\ \mathbf{b}_2 \end{bmatrix}$.

4. Get 4^p potential rules as follows:

$$\text{If } \mathbf{A}_i\,\mathbf{x} \geq \mathbf{b}_i, \text{ then } y' = {}_3q + \sum_{j=1}^{p} {}_3w_j\, g_{ji_j}(t_j)$$

5. Extract our interesting rules via determining if the following optimal problem has an optimal solution.

Minimize: *constant*

Subject to: $\mathbf{A}_i\,\mathbf{x}\ {}^3\ \mathbf{b}_i$

6. For each extracted rule, we check if $\left.\dfrac{\partial y'}{\partial x_k}\right|_{\mathbf{x}\in\{\mathbf{x}|\mathbf{A}_i\mathbf{x}\geq\mathbf{b}_i\}}$ ≥ or < 0, via solving the following optimal problems.

Minimize: $\dfrac{\partial y'}{\partial x_k}$

Subject to: $\mathbf{A}_i\,\mathbf{x}\ {}^3\ \mathbf{b}_i$

Maximize: $\dfrac{\partial y'}{\partial x_k}$

Subject to: $\mathbf{A}_i\,\mathbf{x}\ {}^3\ \mathbf{b}_i$

And we can easily determine if $\left.\dfrac{\partial^2 y'}{\partial x_l \partial x_k}\right|_{\mathbf{x}\in\{\mathbf{x}|\mathbf{A}_i\mathbf{x}\geq\mathbf{b}_i\}}$ ≥ or < 0.

7. Generalize these important differential features via the sign test from these differential analyses in Step 6.