บทที่ 2

ทฤษฎีและสถิติที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงการเปรียบเทียบวิธีการพยากรณ์ในการวิเคราะห์ความถดถอยพหุคูณด้วยวิธีริดจ์รีเกรสชันและวิธีที่ใช้หลักการของโครงข่ายประสาทเทียม ซึ่งรายละเอียดต่างๆ มีดังนี้

วิธีริดจ์รีเกรสชัน

เราสามารถเขียนตัวแบบทั่วไปในการวิเคราะห์ความถดถอยพหุคูณเชิงเส้นได้ดังนี้

(2.1) $$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

เมื่อ $\underset{\sim}{y}$ คือ เวกเตอร์ของตัวแปรตามขนาด n x 1

X คือ เมทริกซ์ของตัวแปรอิสระขนาด n x (p+1)

$\underset{\sim}{\beta}$ คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยพหุคูณขนาด (p+1) x 1

$\underset{\sim}{\varepsilon}$ คือ เวกเตอร์ค่าความคลาดเคลื่อนที่เกิดขึ้นขนาด n x 1

โดยที่ $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$, $\mathrm{cov}(\underset{\sim}{\varepsilon}) = \sigma^2 I_n$

n คือ ขนาดตัวอย่าง

และ p คือ จำนวนตัวแปรอิสระ

จากตัวแบบทั่วไป มักนิยมใช้วิธีกำลังสองน้อยสุดในการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณ ซึ่งอยู่ในรูปของ

(2.2) $$\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}$$

เมื่อ $\hat{\underset{\sim}{\beta}}$ เป็นตัวประมาณไม่เอนเอียงที่มีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดเมื่อเปรียบเทียบกับตัวประมาณไม่เอนเอียงตัวอื่นๆ แต่ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กัน คือมีสภาพที่ไม่เหมาะสม (ill-condition) ซึ่งเกิดขึ้นบ่อยครั้งในทางปฏิบัติ และทำให้การประมาณ $\underset{\sim}{\beta}$ ด้วยวิธีกำลังสองน้อยสุดไม่ได้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด ดังนั้นจึงควรตรวจสอบประสิทธิภาพของตัวประมาณที่ได้

ซึ่งในการพิจารณาประสิทธิภาพของตัวประมาณค่า $\underset{\sim}{\beta}$ แบ่งการพิจารณาเป็นสองส่วน คือเมทริกซ์ความแปรปรวนร่วมของตัวประมาณ $\underset{\sim}{\beta}$ และค่าเฉลี่ยกำลังสองระยะทางจาก $\hat{\underset{\sim}{\beta}}$ ไปยัง $\underset{\sim}{\beta}$ เราสามารถเขียนเมทริกซ์ความแปรปรวนร่วมของ $\hat{\underset{\sim}{\beta}}$ ได้ดังนี้

(2.3) $$\operatorname{cov}(\hat{\underset{\sim}{\beta}}) = \sigma^2 (X'X)^{-1}$$

ให้ L_1 คือ ระยะทางจาก $\hat{\underset{\sim}{\beta}}$ ไปยัง $\underset{\sim}{\beta}$ ดังนั้น

(2.4) $$L_1^2 = (\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})'(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})$$

และได้ค่าเฉลี่ยกำลังสองระยะทางจาก $\hat{\underset{\sim}{\beta}}$ ไปยัง $\underset{\sim}{\beta}$ อยู่ในรูปของ

(2.5) $$\begin{aligned} E(L_1^2) &= \sigma^2 \mathit{trace}(X'X)^{-1} \\ &= E\,[(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})'(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})] \\ &= E\,[\hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}} - 2\hat{\underset{\sim}{\beta}}'\underset{\sim}{\beta} + \underset{\sim}{\beta}'\underset{\sim}{\beta}] \\ &= E(\hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}}) - \underset{\sim}{\beta}'\underset{\sim}{\beta} \end{aligned}$$

(2.6) $$E(\hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}}) = \underset{\sim}{\beta}'\underset{\sim}{\beta} + \sigma^2 \mathit{trace}(X'X)^{-1}$$

เมื่อ $\underset{\sim}{\varepsilon}$ มีการแจกแจงปกติ จะได้ว่า

(2.7) $$\mathit{Var}(L_1^2) = 2\sigma^4 \mathit{trace}\{(X'X)^{-2}\}$$

จากสมการ (2.3), (2.5) และ (2.7) เราจะเห็นได้ว่า $\operatorname{cov}(\hat{\underset{\sim}{\beta}})$, $E(L_1^2)$ และ $\mathit{Var}(L_1^2)$ ต่างก็มีฟังก์ชัน $(X'X)^{-1}$ อยู่ในสมการเช่นเดียวกัน ดังนั้นเพื่อความสะดวกเราจะแปลงฟังก์ชัน $(X'X)^{-1}$ ให้อยู่ในรูปของฟังก์ชันค่าเฉพาะ (eigenvalue function) โดยใช้คุณสมบัติของค่าเฉพาะ กล่าวคือ

ถ้า λ_i เป็นค่าเฉพาะของเมทริกซ์ $X'X$ แล้ว $\sum_{i=1}^{p} \lambda_i = \mathit{trace}(X'X)$; i = 1, 2, ..., p เมื่อ p เป็นจำนวนตัวแปรอิสระ

กำหนดให้ค่าเฉพาะของเมทริกซ์ $X'X$ มีค่าดังนี้

$$(\lambda_{max} = \lambda_1) \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq (\lambda_p = \lambda_{min}) > 0$$

จากสมการ (2.5) สามารถเขียนให้อยู่ในรูปฟังก์ชันค่าเฉพาะได้ดังนี้

$$E(L_1^2) = \sigma^2 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i}\right) \tag{2.8}$$

จากสมการ (2.7) และคุณสมบัติของค่าเฉพาะ เราสามารถเขียนให้อยู่ในรูปฟังก์ชันค่าเฉพาะได้ดังนี้

$$Var(L_1^2) = 2\sigma^4 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i}\right)^2 \tag{2.9}$$

ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูงจะทำให้ $|X'X|$ มีค่าเข้าใกล้ศูนย์ เนื่องจาก $|X'X|$ มีค่าเท่ากับผลคูณของค่าเฉพาะของเมทริกซ์ $X'X$ จึงส่งผลให้ค่าเฉพาะบางค่าของเมทริกซ์ $X'X$ มีค่าน้อยมาก พิจารณาสมการ (2.8) และ (2.9) จะเห็นได้ว่าค่า $E(L_1^2)$ และ $Var(L_1^2)$ มีค่าสูงตามไปด้วย นั่นคือ การประมาณค่า $\underset{\sim}{\beta}$ ด้วยวิธีกำลังสองน้อยสุดจะให้ค่าเฉลี่ยกำลังสองที่สูงขึ้นและ $\underset{\sim}{\hat{\beta}}$ เป็นตัวประมาณที่ไม่เหมาะสม

ในปี ค.ศ. 1970 โฮเอิร์ล (Hoerl) และ เคนนาร์ด (Kennard) ได้เสนอวิธีริดจ์รีเกรสชันเพื่อใช้ในการวิเคราะห์ความถดถอยพหุคูณในกรณีที่เกิดพหุสัมพันธ์ในตัวแปรอิสระ โดยการบวกค่าคงที่มากกว่าศูนย์กับสมาชิกทุกตัวในแนวทแยงมุมของเมทริกซ์ $X'X$ เพื่อลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของการประมาณ $\underset{\sim}{\hat{\beta}}$ และให้ค่าที่ต่ำกว่าวิธีกำลังสองน้อยสุด

ให้ $\underset{\sim}{\hat{\beta}}_R$ เป็นตัวประมาณสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีริดจ์รีเกรสชัน เราสามารถเขียนได้ดังนี้

$$\underset{\sim}{\hat{\beta}}_R = (X'X + kI)^{-1} X' \underset{\sim}{y} \quad ; \; k > 0 \tag{2.10}$$

$$= W X' \underset{\sim}{y} \quad ; \; W = (X'X + kI)^{-1} \tag{2.11}$$

ในการศึกษานี้จะใช้การประมาณค่า k ด้วยวิธีค้นหาลำดับ (binary search) ซึ่งจะได้กล่าวถึงในลำดับต่อไป

จากสมการปกติของวิธีกำลังสองน้อยสุด $X'X\underset{\sim}{\beta} = X'\underset{\sim}{y}$ เราสามารถเขียนตัวประมาณ $\underset{\sim}{\hat{\beta}}_R$ ในรูปของตัวประมาณ $\underset{\sim}{\hat{\beta}}$ ได้ดังนี้

(2.12) $$\underset{\sim}{\hat{\beta}}_R = [I + k(X'X)^{-1}]^{-1}\underset{\sim}{\hat{\beta}}$$

(2.13) $$= Z\underset{\sim}{\hat{\beta}} \qquad ; \quad Z = [I + k(X'X)^{-1}]^{-1}$$

คุณสมบัติที่จำเป็นของ $\underset{\sim}{\hat{\beta}}_R$, W และ Z เพื่อใช้ในการเปรียบเทียบค่าความคลาดเคลื่อนกับวิธีกำลังสองน้อยสุด มีดังนี้

1) ให้ $\xi_i(W)$ และ $\xi_i(Z)$ เป็นค่าเฉพาะของ W และ Z ตามลำดับ ซึ่งคำนวณได้จากสมการลักษณะเฉพาะ (characteristic equations)

$$\left| W - \xi_i I \right| = 0$$

และ $$\left| Z - \xi_i I \right| = 0$$

เราจะได้ค่าเฉพาะของ W และ Z ซึ่งอยู่ในรูปของ

$$\xi_i(W) = \frac{1}{(\lambda_i + k)}$$

และ $$\xi_i(Z) = \frac{1}{(\lambda_i + k)}$$

2) ซึ่งสามารถจัดรูป Z ให้อยู่ในรูปของ W ได้ดังนี้

$$Z = I - k(X'X + kI)^{-1}$$

$$= I - kW$$

3) $\hat{\underset{\sim}{\beta}}_R$ มีค่าน้อยกว่า $\hat{\underset{\sim}{\beta}}$ เมื่อ k > 0 กล่าวคือ

$$\hat{\underset{\sim}{\beta}}_R{}' \hat{\underset{\sim}{\beta}}_R < \hat{\underset{\sim}{\beta}}' \hat{\underset{\sim}{\beta}}$$

พิจารณาตัวประมาณ $\hat{\underset{\sim}{\beta}}_R$ จากสมการ (2.13) เมทริกซ์ $X'X$ และ Z มีคุณสมบัติเป็นบวกแน่นอน (symmetric positive definite) จะได้

$$\hat{\underset{\sim}{\beta}}_R{}' \hat{\underset{\sim}{\beta}}_R \le \xi^2_{max}(Z) \hat{\underset{\sim}{\beta}}' \hat{\underset{\sim}{\beta}}$$

เมื่อ $\xi_{max}(Z) = \dfrac{\lambda_i}{(\lambda_i + k)}$ โดยที่ λ_i เป็นค่าเฉพาะที่มีค่าสูงสุดของเมทริกซ์ $X'X$

พิจารณา คุณสมบัติในข้อ 1) และ 2) ของ Z จะได้ ว่า Z = I เมื่อ k = 0 และ Z มีค่าเข้าใกล้ 0 เมื่อ $k \to \infty$

ถ้าให้ $\phi(\hat{\underset{\sim}{\beta}})$ เป็นผลบวกความคลาดเคลื่อนกำลังสองของตัวประมาณ $\hat{\underset{\sim}{\beta}}$ จะได้

$$\phi(\hat{\underset{\sim}{\beta}}) = \underset{\sim}{y}' \underset{\sim}{y} - \hat{\underset{\sim}{\beta}}' X' \underset{\sim}{y} \tag{2.14}$$

ให้ $\phi(\hat{\underset{\sim}{\beta}}_R)$ เป็นผลบวกความคลาดเคลื่อนกำลังสองของตัวประมาณ $\hat{\underset{\sim}{\beta}}_R$ จะได้

$$\phi(\hat{\underset{\sim}{\beta}}_R) = (\underset{\sim}{y} - X\hat{\underset{\sim}{\beta}}_R)'(\underset{\sim}{y} - X\hat{\underset{\sim}{\beta}}_R)$$

เราสามารถเขียน $\phi(\hat{\underset{\sim}{\beta}}_R)$ ในรูปของ $\underset{\sim}{y}' \underset{\sim}{y}$ และ $X' \underset{\sim}{y}$ ได้ดังนี้

$$\begin{aligned} \phi(\hat{\underset{\sim}{\beta}}_R) &= \underset{\sim}{y}' \underset{\sim}{y} - \hat{\underset{\sim}{\beta}}' X' \underset{\sim}{y} - k\hat{\underset{\sim}{\beta}}_R{}' \hat{\underset{\sim}{\beta}}_R \\ &= \phi(\hat{\underset{\sim}{\beta}}) - k\hat{\underset{\sim}{\beta}}_R{}' \hat{\underset{\sim}{\beta}}_R \end{aligned} \tag{2.15}$$

จากสมการ (2.15) เราสามารถสรุปได้ว่าผลบวกความคลาดเคลื่อนกำลังสองที่ได้จากวิธีริดจ์รีเกรสชัน มีค่าน้อยกว่าผลบวกความคลาดเคลื่อนกำลังสองที่ได้จากวิธีกำลังสองน้อยสุด

คุณสมบัติของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีริดจ์รีเกรสชัน

ในการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณ $\underset{\sim}{\beta}$ ด้วยวิธีกำลังสองน้อยสุด จะได้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยดังนี้

$$MSE(\underset{\sim}{\hat{\beta}}) = \mathrm{cov}(\underset{\sim}{\hat{\beta}})$$

$$= \sigma^2 (X'X)^{-1}$$

เนื่องจาก $E(\underset{\sim}{\hat{\beta}}) = \underset{\sim}{\beta}$ ดังนั้นการประมาณค่าพารามิเตอร์ $\underset{\sim}{\beta}$ ด้วยวิธีกำลังสองน้อยสุด จะได้ตัวประมาณ $\underset{\sim}{\hat{\beta}}$ ที่ไม่เอนเอียง แต่การประมาณค่าพารามิเตอร์ $\underset{\sim}{\beta_R}$ ในวิธีริดจ์รีเกรสชันจะได้ตัวประมาณ $\underset{\sim}{\hat{\beta}_R}$ ที่เอนเอียง กล่าวคือ $E(\underset{\sim}{\hat{\beta}_R}) = Z\underset{\sim}{\beta}$ และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยดังนี้

$$E[L_1^2(k)] = E\,[(\underset{\sim}{\hat{\beta}_R} - \underset{\sim}{\beta})'(\underset{\sim}{\hat{\beta}_R} - \underset{\sim}{\beta})]$$

$$= E\,[(\underset{\sim}{\hat{\beta}} - \underset{\sim}{\beta})'Z'Z(\underset{\sim}{\hat{\beta}} - \underset{\sim}{\beta})] + (Z\underset{\sim}{\beta} - \underset{\sim}{\beta})'(Z\underset{\sim}{\beta} - \underset{\sim}{\beta})$$

$$= \sigma^2 \mathit{trace}\big[(X'X)^{-1}Z'Z\big] + \underset{\sim}{\beta}'(Z-I)'(Z-I)\underset{\sim}{\beta}$$

$$= \sigma^2\Big[\mathit{trace}(X'X+kI)^{-1} - k\,\mathit{trace}(X'X+kI)^{-2}\Big]$$

$$+\ k^2\,\underset{\sim}{\beta}'(X'X+kI)^{-2}\,\underset{\sim}{\beta}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2\,\underset{\sim}{\beta}'(X'X+kI)^{-2}\,\underset{\sim}{\beta}$$

$$= \gamma_1(k) + \gamma_2(k)$$

โดยที่ $\gamma_2(k)$ เป็นระยะทางจาก $Z\underset{\sim}{\beta}$ ไปยัง $\underset{\sim}{\beta}$ มีค่าเป็นศูนย์เมื่อ $k = 0$ และ Z เป็นเมทริกซ์เอกลักษณ์ ดังนั้นอาจพิจารณาเทอม $\gamma_2(k)$ เป็นค่าเอนเอียงกำลังสองของตัวประมาณ $\underset{\sim}{\hat{\beta}_R}$ เมื่อ $k > 0$ ส่วนเทอม $\gamma_1(k)$ เป็นผลรวมความแปรปรวนของการประมาณพารามิเตอร์ $\underset{\sim}{\beta}$ สามารถเขียนในรูปของ $\underset{\sim}{y}$ ได้ดังนี้

$$\underset{\sim}{\hat{\beta}_R} = Z\underset{\sim}{\hat{\beta}}$$

$$= Z(X'X)^{-1}X'\underset{\sim}{y}$$

ดังนั้น

$$Var(\underset{\sim}{\hat{\beta}_R}) = Z(X'X)^{-1}X'\,Var(\underset{\sim}{y})\;X(X'X)^{-1}Z'$$

(2.16) $$= \sigma^2 Z(X'X)^{-1}Z'$$

เราสามารถเขียน Z และ $(X'X)^{-1}$ ในรูปของค่าเฉพาะได้ดังนี้

$$trace(X'X)^{-1} = \sum_{i=1}^{p}\frac{1}{\lambda_i}$$

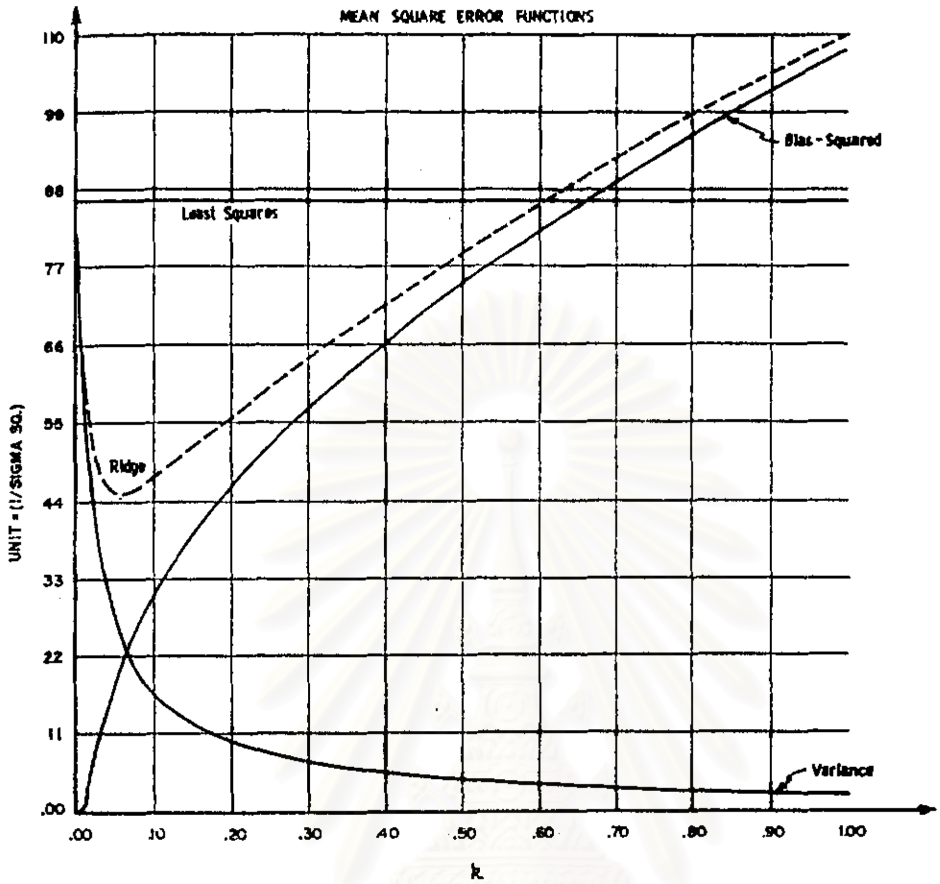
$$trace(Z) = \sum_{i=1}^{p}\frac{\lambda_i}{\lambda_i + k}$$

จากสมการ (2.16) จะได้ว่า

$$\sigma^2 trace\; Z(X'X)^{-1}Z' = \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i + k)^2}$$

ดังนั้น

$$E\,[L_1^2(k)] = Var(\underset{\sim}{\hat{\beta}_R}(k)) + [Bias\;(\underset{\sim}{\hat{\beta}_R}(k))]^2$$



รูปที่ 2.1 แสดงลักษณะความสัมพันธ์ระหว่างความแปรปรวน ความเอนเอียงกำลังสอง และพารามิเตอร์ k

(ที่มา : Authur E. Hoerl and Robert W. Kennard. "Ridge Regression : Biased Estimation for Nonorthogonal Problems." *Technometrics.*, February 1970., page 61.)

จากรูปที่ 2.1 จะเห็นได้ว่าเมื่อ k มีค่าเพิ่ม ความแปรปรวนจะมีค่าลดลง คือ ค่าพารามิเตอร์ k ผกผันกับความแปรปรวนของตัวประมาณ $\hat{\underset{\sim}{\beta}}_R$ และแปรผันตามค่าความเอนเอียงกำลังสองของตัวประมาณ $\hat{\underset{\sim}{\beta}}_R$

พิจารณาที่กราฟเส้นประ และกราฟเส้นตรง จะเห็นได้ว่ามีค่า k บางค่าที่ทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณ $\hat{\underset{\sim}{\beta}}_R$ มีค่าน้อยกว่าค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณ $\hat{\underset{\sim}{\beta}}$

เราสามารถหาขีดจำกัดของความแปรปรวนและความเอนเอียงกำลังสองของตัวประมาณ $\underset{\sim}{\hat{\beta}}_R$ โดยการหาขีดจำกัดของอนุพันธ์อันดับที่ 1 ของฟังก์ชัน $\gamma_1(k)$ และ $\gamma_2(k)$ เทียบกับค่าพารามิเตอร์ k ได้ดังนี้

$$\lim_{k\to 0^+}\left(\frac{\partial \gamma_1}{\partial k}\right) = -2\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i^2}$$

$$\lim_{k\to 0^+}\left(\frac{\partial \gamma_2}{\partial k}\right) = 0$$

ดังนั้นฟังก์ชัน $\gamma_1(k)$ มีอนุพันธ์เป็นลบมีค่าลู่เข้าสู่ $-2\sigma^2\sum_{i=1}^{p}(1/\lambda_i^2)$ เมื่อ $k \to 0^+$ และ $X'X$ เป็นเมทริกซ์เชิงตั้งฉาก และมีค่าลู่เข้าสู่ $-\infty$ เมื่อ $X'X$ มีสภาพไม่เหมาะสม คือตัวแปรอิสระมีความสัมพันธ์กัน ส่วนฟังก์ชัน $\gamma_2(k)$ มีอนุพันธ์เท่ากับศูนย์ เมื่อ $k \to 0^+$ กล่าวคือ เมื่อ $k > 0$ จะมีความเอนเอียงกำลังสองของตัวประมาณ $\underset{\sim}{\hat{\beta}}_R$ เล็กน้อย แต่ความแปรปรวนของตัวประมาณ $\underset{\sim}{\hat{\beta}}_R$ จะมีค่าลดลงมาก จึงทำให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีริดจ์รีเกรสชันมีค่าลดลงและมีค่าน้อยกว่าค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่ได้จากวิธีกำลังสองน้อยสุด

การประมาณค่าพารามิเตอร์ k

ในการวิจัยของจิรายุส หุ่นนตรี (2534) ได้เปรียบเทียบการประมาณค่าพารามิเตอร์ k ด้วยวิธี HKB, TZE และการค้นหาแบบทวิ (Binary Search) ผลสรุปที่ได้ คือวิธีการค้นหาแบบทวิจะให้ค่าการประมาณพารามิเตอร์ k ดีที่สุด แต่ในงานวิจัยของ ธันยากร ต้นชลขันธ์ (2538) ได้ชี้ให้เห็นความไม่เหมาะสมของวิธีนี้ เนื่องจากวิธีค้นหาแบบทวิมีข้อจำกัด คือจะใช้ได้กับข้อมูลที่มีลักษณะแบบไม่ต่อเนื่อง และใช้ในการค้นหาที่มีลักษณะเรียงลำดับเท่านั้น ซึ่งธันยากร ต้นชลขันธ์ ได้นำเสนอวิธีการประมาณค่าพารามิเตอร์ k แบบลำดับ (Sequential Search) ซึ่งเป็นการค้นหาข้อมูลในทางคอมพิวเตอร์เช่นเดียวกับวิธีค้นหาแบบทวิ แต่สามารถใช้ได้กับข้อมูลที่มีลักษณะทั้งแบบต่อเนื่อง และแบบไม่ต่อเนื่อง และข้อมูลที่จะค้นหาจะเรียงลำดับหรือไม่เรียงลำดับก็ได้ แต่ข้อเสียของวิธีการนี้ คือ อาจใช้เวลานาน ถ้า k ที่เหมาะสมอยู่ลำดับไกล แต่ในทางปฏิบัติค่า k จะอยู่ช่วง (0,1) และจากคุณสมบัติของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณ $\underset{\sim}{\hat{\beta}}_R$ เราจะได้ว่า k ที่เหมาะสมจะอยู่ห่างจากศูนย์ไม่มาก ผู้วิจัยจึงได้ใช้วิธีการค้นหาแบบลำดับในการประมาณค่าพารามิเตอร์ k

การประมาณค่าพารามิเตอร์ k โดยวิธีค้นหาลำดับ [1]

กำหนดให้ k_{opt} คือ ค่า k ที่ทำให้ $MSE(\hat{\underset{\sim}{\beta}}_R(k))$ มีค่าต่ำสุด $\hat{\underset{\sim}{\beta}}_R$

k_{int} คือ ค่า k ณ จุดเริ่มต้นในการค้นหาข้อมูล

k_d คือ ค่า k ณ ที่มีการเว้นช่วงห่างของค่า k เท่ากับ d

k_{BEFd} คือ ค่า k ณ จุดที่อยู่ก่อนค่า k ที่เหมาะสม มีการเว้นช่วงห่างของค่า k เท่ากับ d

k_{AFTd} คือ ค่า k ณ จุดที่อยู่หลังค่า k ที่เหมาะสม มีการเว้นช่วงห่างของค่า k เท่ากับ d

และ k_{STOP} คือ ค่า k ที่ใช้ตรวจสอบการหยุดค้นหาค่า k ที่เหมาะสมเมื่อทำการค้นหาข้อมูลทางด้านขวาของค่า k ที่เหมาะสม

ขั้นที่ 1 กำหนดให้ค่า $k_{int} = 0.0$ และค่า $d = 0.01$

ขั้นที่ 2 คำนวณค่า $k_{01} = k_{int} + d$

ขั้นที่ 3 คำนวณหา $MSE(\hat{\underset{\sim}{\beta}}_R(k_{int}))$ และหาค่า $MSE(\hat{\underset{\sim}{\beta}}_R(k_{01}))$ แล้วทำการเปรียบเทียบ

ก) ถ้า $MSE(\hat{\underset{\sim}{\beta}}_R(k_{01})) > MSE(\hat{\underset{\sim}{\beta}}_R(k_{int}))$

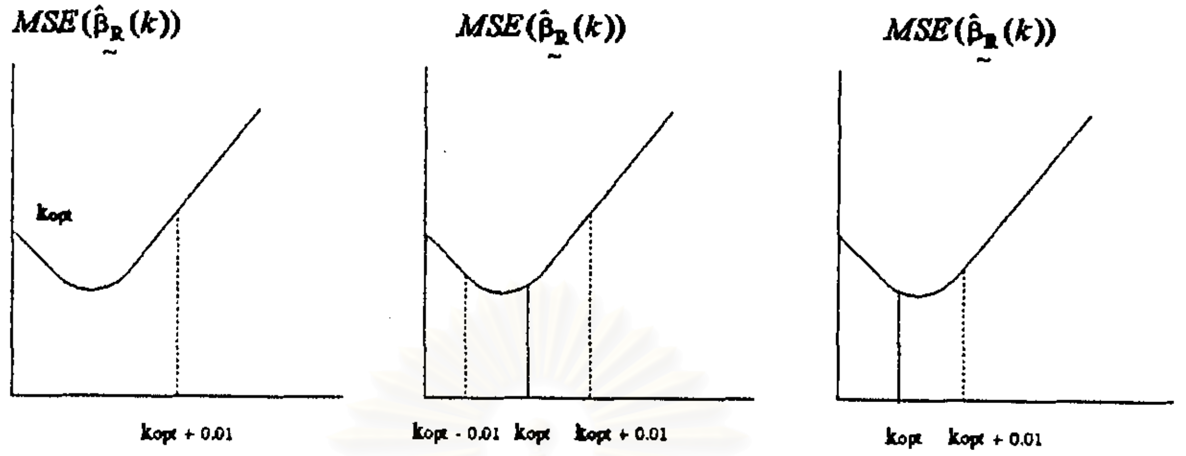
จะได้ว่า $k_{01} = k_{int}$ ยุติการประมวลผล

ข) ถ้า $MSE(\hat{\underset{\sim}{\beta}}_R(k_{01})) \le MSE(\hat{\underset{\sim}{\beta}}_R(k_{int}))$ เราจะกำหนดให้ $k_{int} = k_{01}$ และทำการวิเคราะห์ขั้นที่ 2 ใหม่จนกว่าจะเป็นดังกรณี ก)

เนื่องจากลักษณะของ $MSE(\hat{\underset{\sim}{\beta}}_R(k))$ เป็นรูปโค้งหงาย การหาจุดที่ต่ำสุดจึงขึ้นอยู่กับการกำหนดช่วงห่างของค่า k ซึ่งลักษณะความเป็นไปได้ของการพบจุดที่ต่ำสุดของช่วงห่างที่ไม่ละเอียดพอ ซึ่งแสดงได้ดังกรณีที่ 1-3

---

[1]ธันยากร ตันชลขันธ์, “การเปรียบเทียบการประมาณค่าสัมประสิทธิ์การถดถอยพหุโดยวิธีกำลังสองน้อยที่สุด วิธีริดจ์รีเกรสชัน และวิธีที่ใช้หลักการของริดจ์และสไตน์ ในกรณีที่เกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ” (วิทยานิพนธ์ปริญญามหาบัณฑิตสาขาวิชาสถิติ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2538), หน้า 216.



<u>กรณีที่ 1</u> <u>กรณีที่ 2</u> <u>กรณีที่ 3</u>

<u>กรณีที่ 1</u> ถ้าเรากำหนดช่วง d ห่างเกินไป จุดที่จะได้จุดต่ำสุดอาจเป็นจุดเริ่มแรกที่กำหนดให้ k_{opt} = 0.0 ซึ่งในกรณีนี้แก้ปัญหาได้โดยให้มีการกำหนดช่วงห่างของการค้นหาให้ละเอียดขึ้น และทำการค้นหาข้อมูลแบบลำดับจากจุดเริ่มต้น k_{opt} = 0.0 จนถึงจุดที่ k_{opt} = 0.0 + 0.01 ใหม่อีกครั้ง โดยใช้ช่วงห่างของการคำนวณเป็น d = 0.001

<u>กรณีที่ 2</u> ถ้าจุดของค่า k ที่ค้นหาได้ตกอยู่หลังค่า k ที่เหมาะสมจริงๆ จึงจะเกิดกรณีนี้ขึ้น แก้ปัญหาโดยการเพิ่มช่วงห่างของการค้นหาให้ละเอียดขึ้นและจะค้นหาข้อมูลทางด้านซ้ายของค่า k_{opt} อีกครั้งโดยการค้นหาข้อมูลแบบลำดับจากจุด k_{opt} - 0.01 ไปจนถึงจุด k_{opt} โดยใช้ช่วงห่างของการค้นหาเป็น d = 0.001

<u>กรณีที่ 3</u> ถ้าจุดของค่า k ที่ค้นหาตกอยู่ก่อนหน้า k ที่เหมาะสมจริงๆ จึงจะเกิดกรณีนี้ แก้ปัญหาโดยการเพิ่มช่วงห่างของการค้นหาให้ละเอียดขึ้น และจะค้นหาข้อมูลทางด้านขวาของค่า k_{opt} อีกครั้งโดยจะค้นหาข้อมูลแบบลำดับจากจุด k_{opt} ไปจนถึงจุด k_{opt} + 0.01 โดยใช้ช่วงของการค้นหาเป็น d = 0.001

ดังนั้นจะทำการค้นหาค่า k ที่เหมาะสมอีกครั้งโดยกำหนดช่วงความห่างของการค้นหาใหม่ คือ d = 0.001 โดยการตรวจสอบก่อนว่าค่า k_{opt} ที่ได้มีค่าเท่ากับศูนย์หรือไม่ ถ้ามีค่าเท่ากับศูนย์ก็คือกรณีที่ 1 และทำการค้นหาข้อมูลแบบลำดับเฉพาะทางด้านขวาของค่า k_{opt} แต่ถ้าค่า k_{opt} ไม่เท่ากับศูนย์ จะทำการค้นหาข้อมูลแบบลำดับทางซ้ายของค่า k_{opt} อีกครั้งซึ่งขั้นตอนต่างๆ ในการค้นหามีดังนี้

ให้ d = 0.001

<u>ขั้นที่ 1</u> ตรวจสอบว่าค่า k_{opt} เท่ากับศูนย์หรือไม่

ก) ถ้า $k_{opt} = 0$ ให้ทำการวิเคราะห์ขั้นที่ 3 (ค้นหาทางด้านขวาของค่า k_{opt})

ข) ถ้า $k_{opt} \neq 0$ (มากกว่า 0) ให้ทำการวิเคราะห์ขั้นที่ 2 (ค้นหาทางด้านซ้ายของค่า k_{opt})

<u>ขั้นที่ 2</u> ให้ $k_{BEF01} = k_{opt} - 0.01$

ก) ให้ $k_{001} = k_{BEF01} + d$ เปรียบเทียบค่า k_{001} กับ k_{opt}

ถ้า $k_{001} \geq k_{opt}$ ให้ทำการวิเคราะห์ขั้นที่ 3

ถ้า $k_{001} < k_{opt}$ คำนวณหา $MSE(\hat{\underset{\sim}{\beta}}_R(k_{BEF01}))$ และหา

$MSE(\hat{\underset{\sim}{\beta}}_R(k_{001}))$ แล้วทำการเปรียบเทียบ

ถ้า $MSE(\hat{\underset{\sim}{\beta}}_R(k_{001})) > MSE(\hat{\underset{\sim}{\beta}}_R(k_{BEF01}))$

จะได้ว่า $k_{opt} = k_{BEF01}$ ยุติการประมวลผล

ถ้า $MSE(\hat{\underset{\sim}{\beta}}_R(k_{001})) \leq MSE(\hat{\underset{\sim}{\beta}}_R(k_{BEF01}))$

เราจะกำหนด $k_{BEF01} = k_{001}$ แล้วทำการวิเคราะห์ข้อ ก) ในขั้นที่ 2 จนกว่า

จะเข้ากรณีที่ $k_{001} \geq k_{opt}$ หรือ $MSE(\hat{\underset{\sim}{\beta}}_R(k_{001})) > MSE(\hat{\underset{\sim}{\beta}}_R(k_{BEF01}))$

<u>ขั้นที่ 3</u> ให้ $k_{STOP} = k_{opt} + 0.01$

ก) ให้ $k_{001} = k_{opt}$

ข) ให้ $k_{AFT01} = k_{001} + d$ เปรียบเทียบค่า k_{AFT01} กับ k_{STOP}

ถ้า $k_{AFT01} \geq k_{STOP}$ ให้ยุติการประมวลผล

ถ้า $k_{AFT01} \geq k_{STOP}$ เราจะคำนวณหา $MSE(\hat{\underset{\sim}{\beta}}_R(k_{AFT01}))$และคำนวณหา

$MSE(\hat{\underset{\sim}{\beta}}_R(k_{001}))$ แล้วทำการเปรียบเทียบ

ถ้า $MSE(\hat{\underset{\sim}{\beta}}_R(k_{AFT01})) > MSE(\hat{\underset{\sim}{\beta}}_R(k_{001}))$

จะได้ว่า kopt = k001 และยุติการประมวลผล

ถ้า $MSE(\hat{\underset{\sim}{\beta}}_R(k_{AFT01})) \leq MSE(\hat{\underset{\sim}{\beta}}_R(k_{001}))$

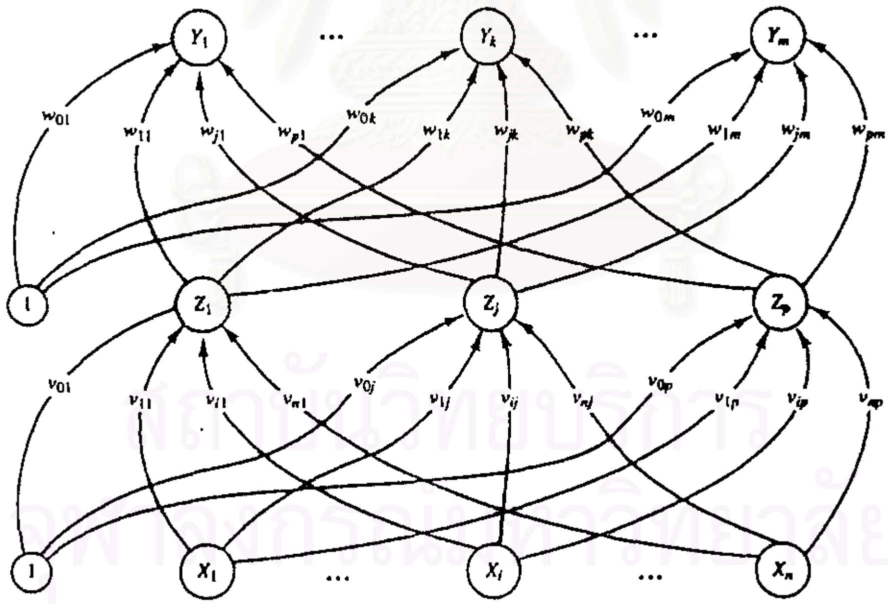
เราจะกำหนด $k_{001} = k_{AFT01}$ และทำการวิเคราะห์ข้อ ข) ในขั้นที่ 3 จนกว่า

จะเข้ากรณีที่ $k_{AFT01} \geq k_{STOP}$ หรือ $MSE(\hat{\underset{\sim}{\beta}}_R(k_{AFT01})) > MSE(\hat{\underset{\sim}{\beta}}_R(k_{001}))$

เนื่องจาก ณ จุดที่ค่า $d = 0.001$ และ $d = 0.0001$ จะให้ค่า $MSE(\hat{\underset{\sim}{\beta}}_R(k))$ ที่มีค่าใกล้เคียงกัน ดังนั้นผู้วิจัยจึงไม่ทำการขยายช่วงค่า d จาก $d = 0.001$ เป็น $d = 0.0001$

วิธีโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมเป็นอีกแนวทางหนึ่งในการแก้ปัญหาทางคอมพิวเตอร์ ซึ่งเป็นปัญหาซับซ้อนจนไม่สามารถสรุปเป็นขั้นตอนที่ชัดเจนได้ การนำหลักการของโครงข่ายประสาทเทียมมาใช้ทางสถิติโดยเฉพาะในการพยากรณ์ เนื่องจากโครงข่ายประสาทเทียมมีความสามารถในการจำชุดข้อมูลเข้า และออกได้เป็นอย่างดี สามารถปรับปรุงตัวแบบที่เหมาะสมได้ด้วยตัวเอง และยังมีความสามารถในการพยากรณ์ข้อมูลที่ไม่เคยเห็นได้เป็นอย่างดี ถ้าจำนวนตัวอย่างที่นำมาฝึกสอนแก่โครงข่ายมีมาก และวิธีการสอนดีพอ ข้อแตกต่างกับวิธีทางสถิติ คือไม่มีการอธิบายตัวแบบ อธิบายการเปลี่ยนแปลงค่าของตัวแปรตามเมื่อตัวแปรอิสระเปลี่ยนไปและไม่มีการทดสอบสมมติฐานในการพยากรณ์ เนื่องจากสัมประสิทธิ์การถดถอยเป็นแบบไม่เชิงเส้น ไม่มีสัมประสิทธิ์ของตัวแปรอิสระแต่ละตัว นั่นคือโครงข่ายประสาทเทียมไม่มีพารามิเตอร์ที่มีคุณสมบัติทางสถิติ และเพื่อง่ายต่อการทำความเข้าใจโครงข่ายประสาทเทียม จะอธิบายโดยใช้ตัวแบบโครงข่ายประสาทเทียม



รูปที่ 2.2 แสดงตัวแบบโครงข่ายประสาทเทียมหนึ่งชั้นซ่อนในรูปทั่วไป

(ที่มา : Laurene Fausett. "Fundamentals of Neural Networks : Architectures, algorithms and applications." New Tersey : Prentice - Hall, Inc., 1994. page 291.)

เมื่อ X_i คือ นิวรอนชั้นนำเข้า (input layer) หน่วยที่ i ; i = 1, 2, ..., n

Z_j คือ นิวรอนชั้นซ่อน (hidden layer) หน่วยที่ j ; j = 1, 2, ..., p

Y_k คือ นิวรอนชั้นผลลัพธ์ (output layer) หน่วยที่ k ; k = 1, 2, ..., m

v_{ij} คือ ค่าถ่วงน้ำหนักในการเชื่อมนิวรอน X_i กับนิวรอน Z_j

w_{jk} คือ ค่าถ่วงน้ำหนักในการเชื่อมนิวรอน Z_j กับนิวรอน Y_k

v_{0j} คือ ค่าความเอนเอียงของนิวรอน Z_j

w_{0k} คือ ค่าความเอนเอียงของนิวรอน Y_k

โครงข่ายที่เชื่อมนิวรอน Z_j ทั้งหมด แทนด้วย z_in_j : $z\_in_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij}$

ค่าที่ส่งออกจากนิวรอน Z_j แทนด้วย z_j : $z_j = f(z\_in_j)$

โครงข่ายที่เชื่อมนิวรอน Y_k ทั้งหมด แทนด้วย y_in_k : $y\_in_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$

และ ค่าที่ส่งออกจากนิวรอน Y_k แทนด้วย y_k : $y_k = f(y\_in_k)$

ตัวแบบทั่วไปอยู่ในรูปของ

$$\hat{y}_k = f\left(\sum_{j=1}^{p}\left(f\left(\sum_{i=1}^{n} x_i v_{ij}\right) w_{jk}\right)\right)$$

เมื่อ $\hat{y}_k$ คือ ค่าพยากรณ์ผลลัพธ์ของนิวรอนหน่วยที่ k

x_i คือ ข้อมูลนำเข้าในนิวรอนหน่วยที่ i

v_{ij} คือ ค่าถ่วงน้ำหนักในการเชื่อมนิวรอน X_i กับนิวรอน Z_j

w_{jk} คือ ค่าถ่วงน้ำหนักในการเชื่อมนิวรอน Z_j กับนิวรอน Y_k

และ $f(x) = \dfrac{1}{1+\exp(-x)}$ คือ Logistic Function

ในโครงข่ายประสาทเทียมมีการกระทำ 2 แบบ คือ

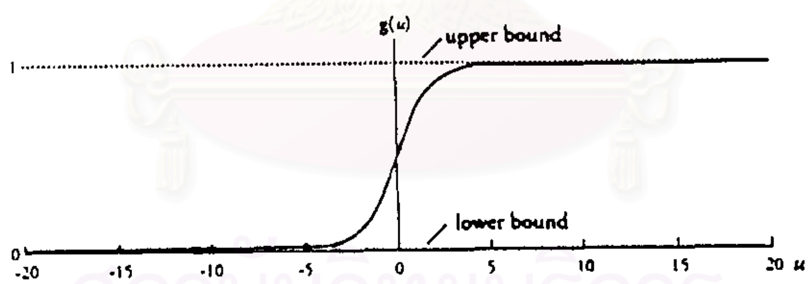
1. แบบการส่ง (mapping mode)
2. แบบการเรียนรู้ (learning mode)

แบบการส่ง (mapping mode)

การส่งเป็นการผ่านค่าไปข้างหน้าจากนิวรอนในชั้นนำเข้า ไปสู่นิวรอนในชั้นผลลัพธ์ ฟังก์ชันการส่งในโครงข่ายประสาทเทียมสามารถให้ค่าถ่วงน้ำหนักที่ถูกต้องกับฟังก์ชันรูปร่างต่างๆ จำนวนมาก หรืออาจกล่าวได้ว่าฟังก์ชันการส่งสามารถเปลี่ยนรูปร่างได้ตามเป้าหมายใดๆ ก็ได้ จนได้ชื่อว่า "a universal approximator" รวมทั้งยังได้มีการพิสูจน์ทางสถิติ [2] โดยใช้ฟังก์ชันซิกมอยด์ (sigmoid function) ซึ่งเป็นโครงสร้างของฟังก์ชันการส่งของโครงข่าย มีลักษณะเป็นฟังก์ชันเพิ่มทางเดียว เป็นฟังก์ชันการส่ง

ในโครงข่ายประสาทเทียมจะมีการคำนวณในชั้นซ่อนและชั้นผลลัพธ์เท่านั้น โดยคำนวณผลรวมค่าถ่วงน้ำหนักของชั้นซ่อน แทนค่าในฟังก์ชันการส่ง เพื่อให้ได้ค่าที่เหมาะสม ฟังก์ชันการส่งที่ใช้กันมากที่สุดในโครงข่ายประสาทเทียม คือ ฟังก์ชันโลจิสติก (logistic function) เนื่องจากสามารถหาอนุพันธ์ได้สำหรับทุกๆ ค่าของโดเมนเพื่อใช้ในการเรียนรู้ และสามารถเขียนอนุพันธ์อันดับหนึ่งให้อยู่ในรูปของฟังก์ชันเองได้ด้วย คือ

$$f'(x) = f(x)\,(1 - f(x))$$



รูปที่ 2.8 แสดงฟังก์ชันความหนาแน่นของฟังก์ชันโลจิสติก

---

[2] Hornik, Kurt, Maxwell Stinchcombe and Halbert White, "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks* 2, No.5 (1989) : 359-366.

Irie, Bunpei, and Sei Miyake, "Capabilities of Three-layered Perceptrons," *IEEE International Conference on Neural Networks, I* (1988) : 161-172.

Cybenko, G, "Approximation by Superpositions of a Sigmoidal Function," *Mathematical Control, Signals and Systems*, 2 (1989) : 303-314.

ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x)=\frac{1}{1+\exp(-x)}$$

แบบการเรียนรู้ (learning mode)

สิ่งสำคัญที่มีผลกับการแพร่ย้อนกลับของความคลาดเคลื่อนในโครงข่ายประสาทเทียม คือ การเปลี่ยนแปลงของค่าถ่วงน้ำหนัก ดังนั้นเราจำเป็นต้องทราบค่าอนุพันธ์ย่อยของความคลาดเคลื่อนเมื่อเทียบกับค่าถ่วงน้ำหนัก

อนุพันธ์ย่อยของความคลาดเคลื่อนกำลังสองของโครงข่ายประสาทเทียมเทียบกับค่าถ่วงน้ำหนักของโครงข่ายประสาทเทียมโดยใช้กฎลูกโซ่ (chain rule) ได้ดังนี้

$$\frac{\partial E}{\partial w_{jk}}=\frac{\partial E}{\partial y_k}\cdot\frac{\partial y_k}{\partial (y\_in_k)}\cdot\frac{\partial (y\_in_k)}{\partial w_{jk}}$$

$$\frac{\partial E}{\partial v_{ij}}=\left(\sum_{k=1}^{m}\frac{\partial E}{\partial y_k}\cdot\frac{\partial y_k}{\partial (y\_in_k)}\cdot\frac{\partial (y\_in_k)}{\partial z_j}\right)\cdot\frac{\partial z_j}{\partial (z\_in_j)}\cdot\frac{\partial (z\_in_j)}{\partial v_{ij}}$$

เมื่อ E คือ ความคลาดเคลื่อนกำลังสอง

v_{ij} คือ ค่าถ่วงน้ำหนักในการเชื่อมนิวรอน X_i กับนิวรอน Z_j

w_{jk} คือ ค่าถ่วงน้ำหนักในการเชื่อมนิวรอน Z_j กับนิวรอน Y_k

v_{0j} คือ ค่าความเอนเอียงของนิวรอน Z_j

w_{0k} คือ ค่าความเอนเอียงของนิวรอน Y_k

z_in_j คือ โครงข่ายที่เชื่อมนิวรอน Z_j ทั้งหมด : $z\_in_j = v_{0j}+\sum_{i=1}^{n}x_i v_{ij}$

z_j คือ ค่าที่ส่งออกจากนิวรอน Z_j : $z_j = f(z\_in_j)$

y_in_k คือ โครงข่ายที่เชื่อมนิวรอน Y_k ทั้งหมด : $y\_in_k = w_{0k}+\sum_{j=1}^{p}z_j w_{jk}$

และ y_k คือ ค่าที่ส่งออกจากนิวรอน Y_k : $y_k = f(y\_in_k)$

ส่วนประกอบของอนุพันธ์เหล่านี้คำนวณโดยผ่านโครงข่ายแบบย้อนกลับ และแบบส่งค่าไปข้างหน้า แสดงได้ดังตารางต่อไปนี้

ส่งผ่านไปข้างหน้า (อ่านขึ้น)	ส่งผ่านไปข้างหลัง (อ่านลง)
$E = \frac{1}{2}(y_k - t_k)^2$	$\frac{\partial E}{\partial y_k} = (y_k - t_k)$
$y_k = f(y\_in_k)$	$\frac{\partial y_k}{\partial (y\_in_k)} = y_k(1-y_k)$
$y\_in_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$	$\frac{\partial (y\_in_k)}{\partial w_{jk}} = \begin{cases} 1 & j = 0 \\ z_j & j = 1,\ldots,p \end{cases}$ $\frac{\partial (y\_in_k)}{\partial z_j} = w_{jk}$
$z_j = f(z\_in_j)$	$\frac{\partial z_j}{\partial (z\_in_j)} = z_j(1-z_j)$
$z\_in_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij}$	$\frac{\partial (z\_in_j)}{\partial v_{ij}} = \begin{cases} 1 & i = 0 \\ x_i & i = 1,\ldots,n \end{cases}$

เมื่อ t_k คือ ผลลัพธ์เป้าหมายของนิวรอนหน่วยที่ k

กำหนดให้

$$p_k = \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial (y\_in_k)}$$

เราจะได้ว่า

$$p_k = (y_k - t_k) y_k (1 - y_k)$$

ดังนั้นอนุพันธ์ย่อยของความคลาดเคลื่อนกำลังสองของโครงข่ายประสาทเทียมกับค่าถ่วงน้ำหนักที่เชื่อมนิวรอนในชั้นผลลัพธ์ คือ

$$\frac{\partial E}{\partial w_{jk}} = \begin{cases} p_k & ; \; j = 0 \\ p_k z_j & ; \; j = 1, \ldots, p \end{cases}$$

กำหนดให้

$$q_j = \left(\sum_{k=1}^{m} \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial (y\_in_k)} \cdot \frac{\partial (y\_in_k)}{\partial z_j} \right) \cdot \frac{\partial z_j}{\partial (z\_in_j)}$$

เราจะกล่าวได้ว่า

$$q_j = \left(\sum_{k=1}^{m} p_k w_{jk} \right) z_j \, (1 - z_j)$$

ดังนั้นอนุพันธ์ย่อยของความคลาดเคลื่อนกำลังสองของโครงข่ายประสาทเทียมกับค่าถ่วงน้ำหนักที่เชื่อมนิวรอนในชั้นซ่อน คือ

$$\frac{\partial E}{\partial v_{ij}} = \begin{cases} q_j & ; \; i = 0 \\ q_j x_i & ; \; i = 1, \ldots, n \end{cases}$$

ขั้นตอนการเรียนรู้แบบแพร่ย้อนกลับ (Backpropagation algorithm)

1. กำหนดค่าถ่วงน้ำหนักเริ่มต้น (initial weight) ให้แก่ทุกนิวรอนในชั้นนำเข้า เป็นเลขสุ่มในช่วง (-1,1)

2. ในแต่ละนิวรอนนำเข้า (X_i , i = 1, 2, ..., n) ได้รับค่า x_i และส่งค่าไปยังทุกๆ นิวรอนในชั้นซ่อน

3. แต่ละนิวรอนในชั้นซ่อน (Z_j , j = 1, 2, ..., p) มีค่ารวมถ่วงน้ำหนักเท่ากับ

$$z\_in_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij}$$

จากนั้นเราจะคำนวณค่าฟังก์ชันกระตุ้น (activation function) ได้ค่าส่งออกในนิวรอนชั้นซ่อนเท่ากับ

$$z_j = f(z\_in_j)$$

และส่งค่า z_j ไปยังทุกๆ นิวรอนในชั้นผลลัพธ์

4. แต่ละนิวรอนในชั้นผลลัพธ์ (Y_k , k = 1, 2, ..., m) มีค่ารวมถ่วงน้ำหนักเท่ากับ

$$y\_in_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$$

จากนั้นเราคำนวณค่าฟังก์ชันกระตุ้น ได้ค่าส่งออกในนิวรอนชั้นผลลัพธ์เท่ากับ

$$y_k = f(y\_in_k)$$

5. เปรียบเทียบผลต่างของค่า $(t_k - y_k)$ ถ้ามีค่าน้อยกว่าหรือเท่ากับที่กำหนดไว้ก็จบการเรียนรู้ มิฉะนั้นทำขั้นที่ 6 ต่อ

6. แต่ละนิวรอนในชั้นผลลัพธ์คำนวณความคลาดเคลื่อนได้เท่ากับ

$$\delta_k = (t_k - y_k) f'(y\_in_k)$$

จากนั้นเราจะคำนวณค่าถ่วงน้ำหนักที่ทำการปรับได้ ซึ่งอยู่ในรูปของ

$$\Delta w_{jk} = \alpha \delta_k z_j$$

เมื่อ α คือ อัตราการเรียนรู้ มีค่าอยู่ในช่วง (0,1]

และคำนวณค่าเอนเอียงที่ทำการปรับได้ ซึ่งอยู่ในรูปของ

$$\Delta w_{0k} = \alpha \delta_k$$

แล้วส่งค่า δ_k ไปยังทุกๆ นิวรอนในชั้นซ่อน

7. แต่ละชั้นซ่อนมีผลรวมค่าถ่วงน้ำหนักของความคลาดเคลื่อนเท่ากับ

$$\delta\_in_j = \sum_{k=1}^{m} \delta_k w_{jk}$$

จากนั้นเราจะคำนวณฟังก์ชันกระตุ้นของค่าความคลาดเคลื่อนเท่ากับ

$$\delta_j = \delta\_in_j f'(z\_in_j)$$

และจะคำนวณค่าถ่วงน้ำหนักที่ทำการปรับได้ ซึ่งอยู่ในรูปของ

$$\Delta v_{ij} = \alpha \delta_j x_i$$

แล้วคำนวณค่าความเอนเอียงที่ทำการปรับได้ ซึ่งอยู่ในรูปของ

$$\Delta v_{0j} = \alpha \delta_j$$

8. ปรับค่าถ่วงน้ำหนักที่เชื่อมนิวรอนในชั้นซ่อนกับนิวรอนในชั้นผลลัพธ์ ด้วยค่าความคลาดเคลื่อน และค่าความเอนเอียง ได้ค่าถ่วงน้ำหนักใหม่เท่ากับ

$$w_{jk}(new) = w_{jk}(old) + \Delta w_{jk}$$

และทำการปรับค่าถ่วงน้ำหนักที่เชื่อมนิวรอนในชั้นนำเข้ากับนิวรอนในชั้นซ่อน ด้วยค่าความคลาดเคลื่อน และค่าความเอนเอียง ได้ค่าถ่วงน้ำหนักใหม่เท่ากับ

$$v_{ij}(new) = v_{ij}(old) + \Delta v_{ij}$$

9. กลับไปทำขั้นที่ 3

การแปลงข้อมูลโดยการแปลงที่ใช้การยกกำลัง (Power Transformation)

กรณีที่ความคลาดเคลื่อนมีการแจกแจงแบบเบ้จะทำการแปลงข้อมูลโดยการแปลงที่ใช้การยกกำลัง เพื่อที่จะทำให้ข้อสมมติของความเป็นปกติใช้ได้ ซึ่งในปี ค.ศ. 1964 บอกซ์ (Box) และคอกซ์ (Cox) ได้เสนอการแปลงข้อมูลซึ่งอยู่ในรูปของ

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{(\lambda)} - 1}{\lambda} & ; \ \lambda \neq 0 \\ \ln y_i & ; \ \lambda = 0 \end{cases}$$

ซึ่งการแปลงรูปตัวแปรตามในลักษณะนี้ทำให้ $y_i^{(\lambda)}$ มีการแจกแจงปกติโดยประมาณซึ่งขั้นตอนการแปลงรูปมีดังนี้

1) จากสมการ $\underset{\sim}{y} = X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}$ คำนวณค่ามัชฌิมเรขาคณิต (geometric mean) ของ $\underset{\sim}{y}$ ได้จาก $G = \sqrt[n]{y_1 y_2 \ldots y_n}$ มาหาร y_i i = 1, 2, ..., n ทำให้ได้ $y_{i*} = y_i / G$; i = 1, 2, ..., n และจะได้ว่า $\sum_{i=1}^{n} \ln(y_{i*}) = 0$

2) จากสมการ $\frac{y_{i*}^{(\lambda)} - 1}{\lambda} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ ซึ่ง บอกซ์และคอกซ์ ถือว่ากรณีเช่นนี้ $\underset{\sim}{\varepsilon} \sim N(0,\sigma^2)$ ดังนั้นฟังก์ชันความหนาแน่นร่วมของ $\underset{\sim}{y}$ คือ

$$\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n} (y_{i*}^{(\lambda)} - \underset{\sim}{X}'\underset{\sim}{\beta})^2\right] \cdot J$$

เมื่อ J คือ จาโคเบียนของการแปลงจากตัวแปร y_{i*} เป็น $y_{i*}^{(\lambda)}$ ดังนั้น

$$J = \prod_{i=1}^{n} y_{i*}^{(\lambda-1)}$$

จะได้ว่า ฟังก์ชัน log-likelihood อยู่ในรูปของ

$$\ln L = \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (y_{i*}^{(\lambda)} - \underset{\sim}{X}'\underset{\sim}{\beta})^2$$

ทำการถดถอย $y_{i*}^{(\lambda)}$ บน X_i และคำนวณผลรวมกำลังสองของค่าความคลาดเคลื่อน $\hat{\sigma}^2(\lambda)$ โดยที่

$$\hat{\sigma}^2(\lambda) = \frac{1}{n}(\underset{\sim}{y}_*^{(\lambda)} - \underset{\sim}{X}\hat{\beta}(\lambda))'(\underset{\sim}{y}_*^{(\lambda)} - \underset{\sim}{X}\hat{\beta}(\lambda))$$

3) เลือกค่า λ ที่ให้ค่า $\hat{\sigma}^2(\lambda)$ ที่น้อยที่สุด จะได้ λ ที่เป็น maximum likelihood ของ λ