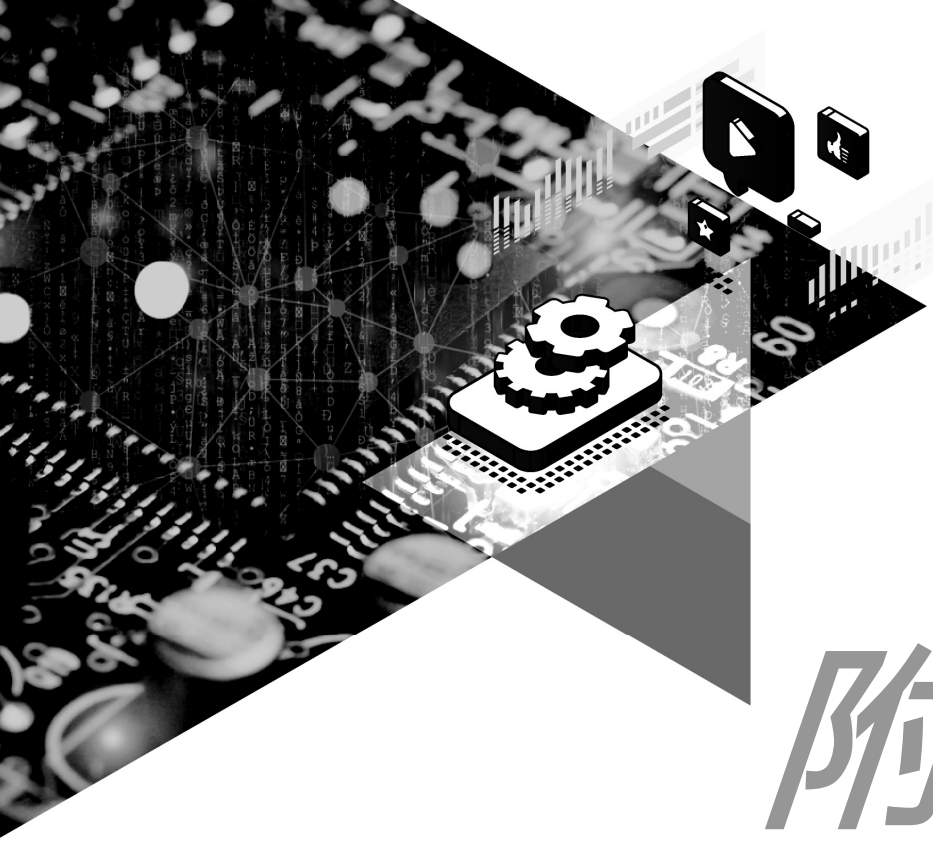

附錄D

無母數迴歸與分類

Nonparametric Regression and Classification

前章我們介紹了解決迴歸問題的「線性迴歸」以及解決分類問題的「線性分類器」，對於這些線性模型來說，模型的建構與計算不僅相對簡易，模型的解釋與推論也相當完備。然而，線性模型存在一大缺陷，就是預測能力是有限的，線性模型受限於線性以及給定的分配假設。實務上，特徵與目標通常存在複雜的非線性關係，因此無論我們使用何種線性模型，一般來說，在預測表現上都不如非線性模型。但這並不代表線性模型不重要，許多非線性模型是基於這些線性模型之上所發展而成的。在數據科學中，我們更期望的是能兼顧模型的預測力與解釋力（非線性模型的解釋力通常較低），因此如何同時使用線性模型與非線性模型在「預測力與解釋力之間權衡」，是人工智慧的主軸之一（後續章節將介紹「可解釋人工智慧」）。此外，非線性模型由於其彈性較大，通常具備一個至多個的超參數需要設定，因此如何調整超參數使得模型在「偏誤與變異之間權衡」，也是建構良好非線性模型的一大關鍵。

於前一章節的「線性分類器」中，我們已介紹了兩個非線性的模型，是屬於生成模型的「樸素貝氏分類器」以及「二次判別分析」（假設多元常態分配，但類別間可不等的變異）。本章節以及後續將依序介紹常用且重要的非線性模型，分別為「**無母數迴歸**」（non-parametric regression）、「**支持向量機**」（support vector machine, SVM）、「**決策樹與集成學習**」（decision tree and ensemble learning）以及「**類神經網路與深度學習**」（neural network and deep learning）。

由於大多數的線性模型假設了機率分配（線性迴歸假設了誤差服從常態分配、羅吉斯迴歸假設目標變數服從二項分配、線性判別分析假設各個類別的特徵服從多變量常態分配），這樣的模型我們稱之為「**有母數**」（parametric）方法。相反地，我們將介紹的這些非線性模型，大多數為「**無母數**」（nonparametric）方法，其中「**核密度估計**」（kernel density estimation）是無母數估計方法中重要的觀念之一。「核密度估計」是以無母數方式「**區域地**」（locally）且「**平滑地**」（smoothly）估計機率密度函數，若將這樣的思維拓展至迴歸問題中，即是本章介紹的「無母數迴歸」方法。我們將依序介紹「無母數迴歸」家族方法包括了基於「核密度」的「**K 最近鄰法**」（K-nearest neighbors, KNN）與「**核迴歸與區域線性迴歸**」（kernel regression and local linear regression），以及基於「樣條」的「**迴歸樣條**、**自然樣條**、**平滑樣條**以及**區域適應性迴歸樣條**」（regression splines,

nature splines, smoothing splines, and locally adaptive regression splines)，以及由單變量延伸至多變量的「**加法模型**」(additive models)與進一步考慮交互作用的「**多變量適應性迴歸樣條**」(multivariate adaptive regression splines, MARS)。

D.1 無母數思維：核密度估計

在典型的有母數統計方法中，我們可能假設某個變數服從常態分配或其他的機率分配，然而在實務中，該變數真正的分配可能並非如此單純，可能是由多個分配組合而成的；又或是我們根本無先驗知識得知數據服從的分配，常藉由數據視覺化中的「**直方圖**」(histogram)，協助我們觀察數據的分配。圖 D.1 為紅酒製造案例中目標變數品質的直方圖，其呈現的分配相當複雜。然而，直方圖呈現的分布可能過於顛簸，而這時就需仰賴能估計出更平滑的「核密度估計」方法。「核密度估計」是一個估計機率密度函數的無母數方法，此方法並未對數據假設任何分配，而是基於數據本身的密度直接估計而出的。

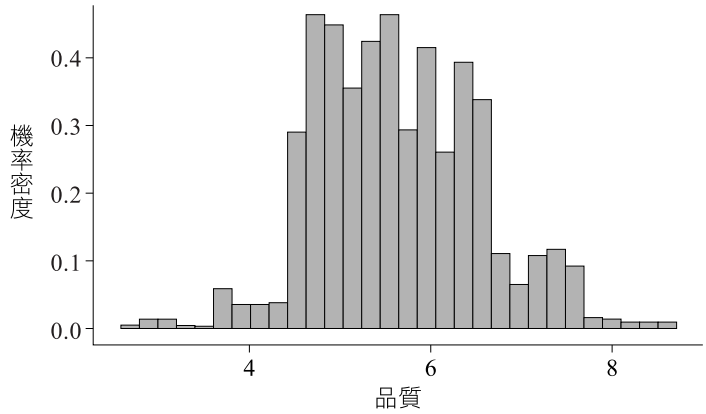

圖 D.1　紅酒製造案例的品質直方圖

D.1.1 直方圖

實際上，在數據視覺化中的直方圖內含核密度估計的思維，直方圖所建構的機率密度函數$\hat{f}$一般可表示如公式(D.1)所示。

$$\hat{f}(x) = \underbrace{\frac{1}{n}}_{\text{縮放(scaling)}} \cdot \underbrace{\frac{\sum_{i=1}^{n} I\left(|x_i - x| \le \frac{h}{2}\right)}{h}}_{\text{長條的高度(height of bar)}} \tag{D.1}$$

其中n為樣本數，x_i為第i個樣本，h為「**帶寬**」（bandwidth）表示直方圖中直方的寬度，x為帶寬的中心點。而$I(|x_i - x| \le \frac{h}{2})$為「**指標函數**」（indicator function），表示樣本x_i是否在特定的直方內（$h/2$為左半邊與右半邊寬度），因此上式可計算出每個直方中心的密度，而得到直方圖。如圖 D.2 所示，我們由紅酒數據中抽取十個樣本來繪製直方圖，其中 x 軸為品質 y 軸為機率密度，在給定的帶寬下（直方的寬度），多個樣本落於相同直方因而密度較大。因此，我們可將公式(D.1)的$|x_i - x|$理解為每個直方中心點得到每個樣本x_i對其的貢獻，而這就是核密度估計的核心思維：「**計算每個樣本點對於任意點**（此處直方圖是狹義的每個直方的中心點）**的貢獻**」。也就是說，當任意點附近的樣本越多，則其得到的貢獻也越大。然而，每個樣本對於任意點的貢獻該如何決定？是否能滿足越靠近則貢獻越多的原則？此外，使用直方圖估計機率密度函數存在兩大問題，一是估計的函數相當顛簸（機率密度變動大），這是由於使用的是指標函數；二是落於直方邊緣的樣本即便密度高，容易被分散於兩個直方內。

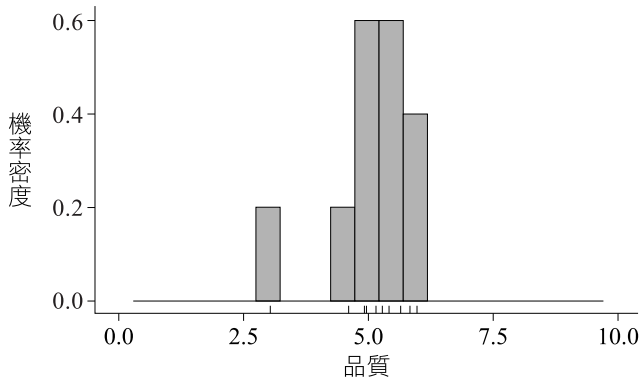

圖 D.2　樣本密度與直方圖

D.1.2　核密度估計

在直方圖中的貢獻是依據「指標函數」所決定的，落於給定靠近樣本

的範圍賦予 1 其餘賦予 0；而核密度估計即是將樣本的貢獻拓展成一般化的任意函數，而這個賦予每個樣本貢獻（權重）的函數稱之為「**核函數**」（kernel function）。因此，我們可以說直方圖是使用「指標函數」作為核函數的核密度估計。一般來說，以核密度估計的機率密度函數可表示如公式(D.2)所示。

$$\hat{f}(x) = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}{h} \tag{D.2}$$

其中x為任意樣本點，K為核函數是一個賦予了每個樣本特定權重的函數，h為帶寬。若我們使用與圖 D.2 相同的數據，但選用「常態分配函數」作為核函數，結果如圖 D.3 核密度函數所示。此等同於賦予每個樣本一個相同的常態分配，而估計出的機率密度函數也將等同於加總所有常態分配並除上樣本數。

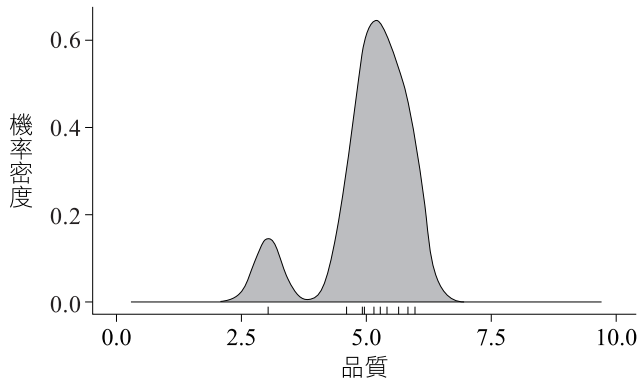

圖 D.3　以常態分配函數作為核函數的核密度函數

D.1.2.1　帶寬選擇

帶寬會對我們的估計造成什麼影響呢？如圖 D.4 所示，延續前述案例，若我們以相同的常態核函數估計，當帶寬越大時（左圖），常態核函數的變異越大，估計出的機率密度函數的平滑程度也越大；相反地，當帶寬越小時（右圖），常態核函數的變異越小，估計出的機率密度函數越曲折。若從另一角度觀之，帶寬的選擇對於機率密度函數的估計，就好比於超參數的選擇對於模型配適，取決於「**偏誤與變異**」兩者間的權衡。

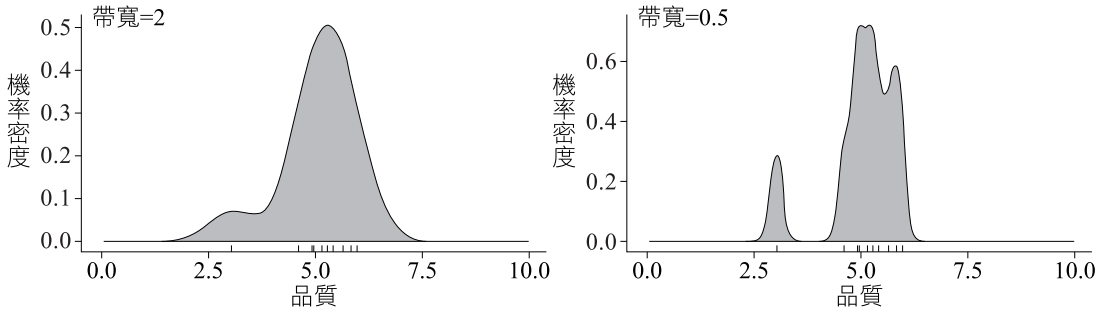

圖 D.4　常態核函數在不同帶寬下的估計表現

D.2　K 最近鄰法

延續「核密度估計」的思維，「K **最近鄰法**」假設任意樣本應與其鄰近樣本有著高度相關（區域性），並將此思維應用至監督式學習的迴歸與分類問題中。更具體的說，「K 最近鄰法」包含了「K **最近鄰分類器**」（KNN classifier）以及「K **最近鄰迴歸**」（KNN regression），分別是對任意樣本使用最近的 K 個樣本來判斷自身應該為何種類別或估計其數值。

D.2.1　最近鄰分類器

在分類問題中，「K 最近鄰分類器」預測某一樣本x_0屬於某一類別j的機率如公式(D.3)所示。

$$\Pr(Y = j|X = x) = \frac{1}{K}\sum_{i\in \mathbf{N}} I(y_i = j) \tag{D.3}$$

其中N為屬於x附近的$|N| = K$個樣本所形成的子集，$I(y_i = j)$為一個指標函數，若挑選出的樣本類別屬於給定的類別j則為 1；反之為 0。而K個近鄰樣本子集的收集，是由計算原有數據與特定數據特徵的「歐氏距離」（Euclidean distance），如公式(D.4)所示（也可使用其他的距離函數）。

$$d(x_i, x) = \sqrt{\sum_{j=1}^{p}(x_i^j - x^j)^2} \tag{D.4}$$

並由所有距離排序後，挑選出距離最小的K個樣本。

若以視覺化呈現，「K 最近鄰分類器」分類流程如圖 D.5 所示。首先，給定一個待分類的樣本（黑色圓點）；其次，我們計算該樣本與原始數據所有樣本的距離並排序；接著，在給定的鄰居數 K 下，找出距離最近的 K 個樣本子集；最後，並投票它們所屬的類別，選擇票數最高的那一個作為預測值（等於求出各別的機率後選擇最大者）。K 最近鄰分類器演算法如表 D.1 所示。

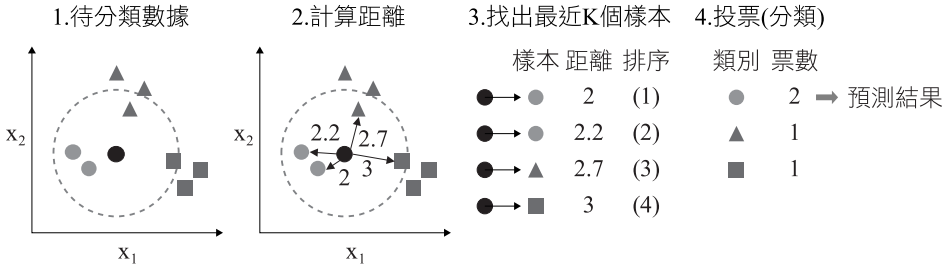

圖 D.5　K 最近鄰分類器

表 D.1　K 最近鄰分類器演算法

演算法 K 最近鄰分類器（KNN classifier）
輸入：訓練集$S=(x_1,y_1),\ldots,(x_n,y_n)$，測試樣本$x$，鄰居數$K$ 輸出：測試樣本的預測值$\hat{y}$ 1. For $i=1$ to n: 2. 計算訓練集與測試樣本的距離： $$d(x_i,x)=\sqrt{\sum_{j=1}^{p}(x_i^j-x^j)^2};$$ 3. End 4. 排序後找出距離最小的K個樣本，並令這個樣本子集為n 5. 求出每一類別的機率： $$\Pr(Y=j\|X=x)=\frac{1}{K}\sum_{i\in N}I(y_i=j);$$ 6. 找出機率最大的類別作為預測值$\hat{y}$ 7. Return K 最近鄰分類器的預測值$\hat{y}$

D.2.2　K 最近鄰迴歸

在迴歸問題中，「K 最近鄰迴歸」預測樣本x的預測值如公式(D.5)所示。

$$\hat{f}(x) = \frac{1}{K}\sum_{i \in N} y_i \tag{D.5}$$

其中N為屬於x附近的$|N| = K$個樣本所形成的子集（同樣是計算歐氏距離並排序挑選出），y_i為數據中一樣本的目標值，因此經由上述過程我們可以得到給定任意樣本的預測值。舉個單一特徵的迴歸問題案例，如圖 D.6 所示，在給定一樣本x_0下，找出鄰近的K個樣本後，計算它們函數值的平均即可得到預測值$\hat{f}(x_0)$，圖中虛線為實際函數、實線為「K 最近鄰迴歸」預測函數。

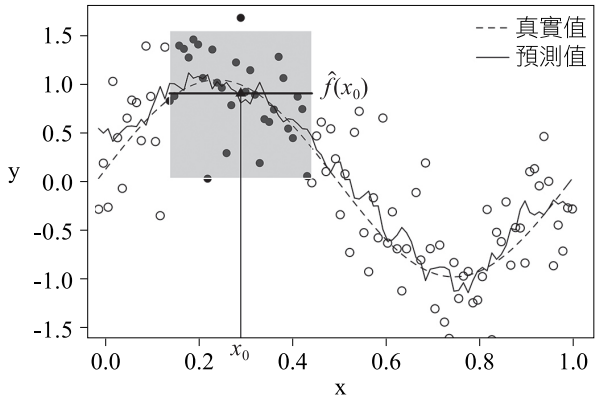

圖 D.6　K 最近鄰迴歸

在「K 最近鄰法」中，超參數「K（鄰居數）」決定了模型的偏誤與變異之間的權衡，當 K 個數越大時越「**一般化**」（generalization）（偏誤越大而變異越小），較容易欠缺配適；相反地，當 K 個數越小時越「**特殊化**」（specification）（偏誤越小而變異越大），較容易過度配適。一般來說，「K 最近鄰法」的**優點**包含(1)實作與解釋相當容易；(2)對於複雜的非線性關係有不錯的預測力（對於每個區域都有很好的配適）；(3)相對較少的超參數（K 個數以及距離函數）需要被調整。其**缺點**則包含(1)通常在迴歸問題上平滑程度相當低（與直方圖相似），難以建構出平滑且配適較佳的迴歸曲線，這是由於直接以相同的權重選定 K 個樣本作為平滑；(2)受到維度詛咒的影響使得預測力降低（維度增加時樣本之間的相似度就會劇烈地下降）；(3)受到數據量的影響使得預測值的計算時間增加（每一預測樣

本需與所有訓練樣本計算相似度）。後續我們將介紹強化「**平滑度**」與「**預測力**」的進階無母數迴歸模型。

D.3 核迴歸與區域線性迴歸

前述的「K 最近鄰法」在迴歸問題上，以近鄰的樣本直接取平均使得迴歸曲線相當不平滑（與「核密度估計」中直方圖所遇到的是相同問題），因此，「**核迴歸**」（kernel regression）將「核函數」（kernel function）引入「K 最近鄰迴歸」使模型不僅考慮了「區域性」，更囊括了「平滑性」；而「**區域線性迴歸**」（local linear regression）則是進一步考慮核函數在邊界上的截斷，並以線性模型配適加以修正。

D.3.1 核迴歸

在考慮單一特徵（單變量）的情境下（後續會說明多變量的配適），從純粹的「K 最近鄰迴歸」引入強化平滑性的「核函數」後，便成為了區域且平滑地加權平均的「**核迴歸**」，其預測樣本x的預測值如公式(D.6)所示。

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{(x - x_i)^2}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{(x - x_i)^2}{h}\right)} \tag{D.6}$$

其中x_i與y_i為數據中一樣本的特徵與目標值，K為核函數賦予了每個樣本特定權重，h則為核函數的帶寬。以相同的案例來說（圖 D.6），核迴歸預測的結果如圖 D.7 所示，在給定一樣本特徵x_0下，近鄰樣本被選用「Epanechnikov 核」的核函數賦予不同大小的權重，計算這些近鄰樣本目標值的核函數加權平均後，即可得到預測值$\hat{f}(x_0)$。實際上，上述的「K 最近鄰迴歸」是「核迴歸」選用「均勻核」的特例，因此我們比較此兩結果可發現（圖 D.6 與圖 D.7），選用「Epanechnikov 核」的「核迴歸」在預測函數的平滑性有更好的效果。由此觀之，「核密度估計」是使用核函數（將數據加權）估計機率密度函數，而在「K 最近鄰法」中則是使用投票或平均估計特徵與目標的函數關係，因此兩者皆充分蘊含了無母數中的「區域性」思維。

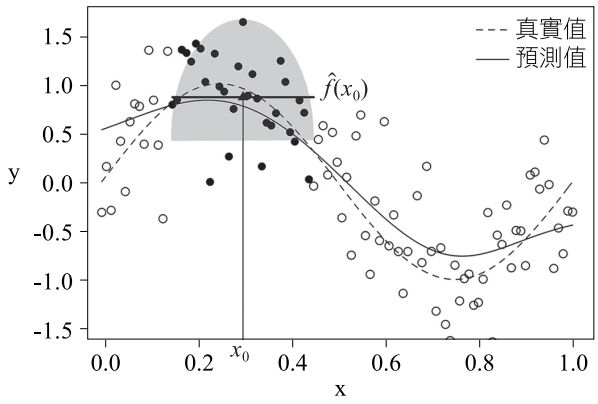

圖 D.7　以 Epanechnikov 核函數的核迴歸

D.3.2　區域線性迴歸

在上述的「核迴歸」存在著一大缺陷，當我們使用核函數進行加權平滑時，在邊界上核函數會被截斷，這樣被截斷的核函數具有**不對稱性**（asymmetry），使得在邊界的函數估計存在嚴重的**偏誤**（biased）。如圖 D.8 左圖所示，圖中最左邊的預測函數估計因為有著被截斷的核函數，使得模型的預測值被**高估**（overestimated），我們可以看到圖中最右邊也有一樣的情形，而這個問題該如何被解決呢？首先，將原有的「核迴歸」由公式(D.6)轉換成一個最佳化的形式，如公式(D.7)所示。

$$\begin{aligned}\hat{f}(x) &= \underset{f(x)}{\arg\min} \sum_{i=1}^{n} K\left(\frac{(x-x_i)^2}{h}\right)(y_i - f(x))^2 \\ &= \underset{\theta_x}{\arg\min} \sum_{i=1}^{n} K\left(\frac{(x-x_i)^2}{h}\right)(y_i - \theta_x)^2\end{aligned} \tag{D.7}$$

其中預測值$\hat{f}(x)$在上述「核迴歸」最佳化的式子中改以常數$\hat{\theta}_x$表示，因而我們可以改用一個區域化的「線性迴歸」配適取代原先區域的「常數」配適，也就是令函數$f(x)$為線性迴歸$\alpha_x + \beta_x x$，以減低邊界估計的偏誤，於是一個「**區域線性迴歸**」（local linear regression）的最佳化形式可表示如公式(D.8)所示。

$$\underset{\alpha_x,\beta_x}{\arg\min} \sum_{i=1}^{n} \underbrace{K\left(\frac{(x-x_i)^2}{h}\right)}_{\text{權重(weight)}} \underbrace{(y_i - \alpha_x - \beta_x x)^2}_{\text{最小平方法}} \tag{D.8}$$

上式意味著我們對於任意點的預測，皆可視為求解一個以核函數作為權重的「**加權線性迴歸**」(weighted linear regression) 問題。以相同的案例來說 (圖 D.6)，如圖 D.8 右圖所示，在給定靠近邊界的樣本點x_0下，近鄰樣本被核函數賦予不同大小的權重，並將這些權重賦予到「加權線性迴歸」中校正，因此我們可以觀察在邊際上的預測值比起「核迴歸」有更好的預測效果 (與圖 D.8 左圖相比)，沒有明顯的偏誤。

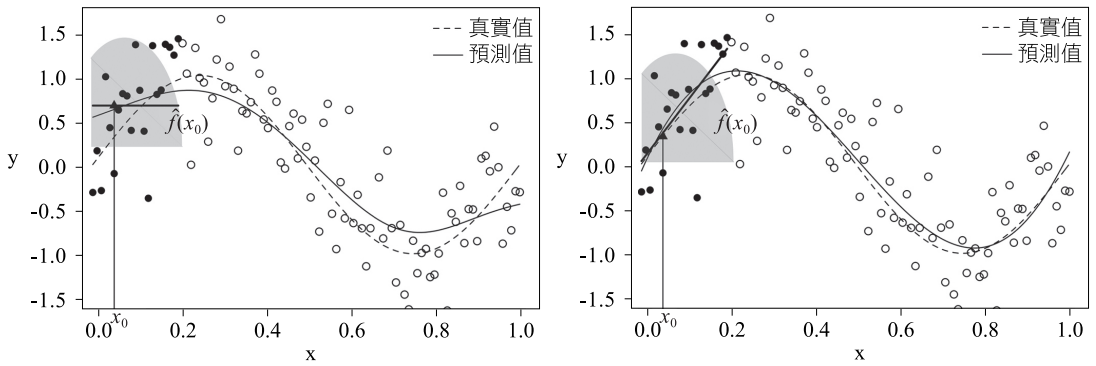

圖 D.8 被截斷的核函數造成邊界估計的偏誤與區域線性迴歸

D.4 迴歸樣條、自然樣條、平滑樣條以及區域適應性迴歸樣條

前述所介紹的無母數迴歸方法「K 最近鄰迴歸」、「核迴歸」與「區域線性迴歸」均對近鄰樣本使用「核函數」進行區域的平滑，而此小節我們將以另一個角度建構無母數迴歸模型，是使用多個子模型「**基函數**」(basis function) 對整個函數進行「**分段**」的建模，其中包含了「**迴歸樣條**」(regression splines)、「**自然樣條**」(nature splines)、「**平滑樣條**」(smoothing splines) 以及「**區域適應性迴歸樣條**」(locally adaptive regression splines)。

首先，我們考慮單一特徵並將整個迴歸模型拆解以多個「基函數」的線性加總表示，也就是一個「**線性基函數擴張模型**」(linear basis expansion model)，如公式(D.9)所示，

$$f(x) = \sum_{j=1}^{k} \beta_j h_j(x) \tag{D.9}$$

其中h_i是第j個給定的基函數，而這些基函數可以是多項式、非線性、交互作用或是分段的函數轉換，而β_i是特徵經基函數j轉換後新特徵的迴歸係數。此方法巧妙之處在於將原始特徵經過多個不同的轉換後，亦可以使用線性迴歸的方式估計各個新特徵的迴歸係數，也就是使用線性的手法估計一個複雜度極高的非線性模型。更精確地說，在無母數迴歸中，我們強調的是「區域性」與「平滑性」的兩大核心思維，因此主要以「**分段**」與「**多項式**」函數作為基函數，並且我們這些高度平滑且分段模型被稱之為「**樣條**」(splines)。此外，由於我們以多個特徵更精確地描述原有單一特徵的非線性，使模型具備足夠的自由度（彈性）配適複雜的非線性關係，但這也同時增加了模型過度配適發生的可能，因此在「線性基函數擴張模型」中，模型自由度會是我們需特別注意的一項因子。

D.4.1 迴歸樣條

「**迴歸樣條**」(regression splines)，又被稱作「**分段多項式函數**」(piecewise polynomial function)，是將特徵的值域以給定的「**節點**」(knot) 分段成多個連續的區間，並且在每個區間各別配適一個多項式函數。若我們以p個節點分段成$p+1$區間，則這些區間可以表示如公式(D.10)所示，

$$(-\infty, \mathrm{t}_1], [\mathrm{t}_1, \mathrm{t}_2], \dots, [\mathrm{t}_p, \infty) \tag{D.10}$$

並於這些區間配適k階的多項式。以下我們將逐步說明如何在各區間建構零次（常數）、一次（斜率）以及推廣到k次多項式的「迴歸樣條」，並且維持節點左右的「**連續性**」與「**平滑性**」的重要關鍵。

首先，我們可在每個區段以一個最簡單的常數配適，我們令三個基函數為各區間的指標函數，如公式(D.11)所示。

$$h_1(x) = I(x < \mathrm{t}_1), \quad h_2(x) = I(\mathrm{t}_1 < x < \mathrm{t}_2), \quad h_3(x) = I(\mathrm{t}_2 < x) \tag{D.11}$$

而實際上這三個基函數的迴歸係數估計等於各區間的平均。接著，我們進一步在每個區段以一個線性函數配適，因此除了原先截距項的三個基函數外，還需另外三個一次項的基函數，如公式(D.12)所示，

$$\begin{aligned} &h_4(x) = I(x < \mathrm{t}_1) \cdot x, \quad h_5(x) = I(\mathrm{t}_1 < x < \mathrm{t}_2) \cdot x, \\ &h_6(x) = I(\mathrm{t}_2 < x) \cdot x \end{aligned} \tag{D.12}$$

然而，使用這六個基函數的模型有極高的可能發生節點左右「不連續」（discontinuous）的情形。然而對於一個函數配適而言，我們更偏好一個連續（甚至是平滑）的函數，如圖 D.9 左圖所示。因此為確保函數的連續性，我們可以令上述案例的第二、三段區間的基函數（$h_2(x), h_3(x)$）等同於第一段區間的基函數（等價於直接移除此兩基函數），使整個函數使用相同的常數截距項。為了有更直接的表示方式，我們可改用下列四個基函數表示，如公式(D.13)所示。

$$h_1(x) = 1, \quad h_2(x) = x, \quad h_3(x) = (x - \mathrm{t}_1)_+, \; h_4(x) = (x - \mathrm{t}_2)_+ \qquad \text{(D.13)}$$

其中「**分段線性**」（piecewise linear）的基函數$(x - \mathrm{t})_+$表示減去節點後取大於等於零的部分，也就是$(x - \mathrm{t})_+ = I(x > \mathrm{t}) \cdot (x - \mathrm{t})$，如圖 D.9 右圖所示。換言之，原先的模型是由各區間以一個相同的截距以及三個不同斜率的分段基函數所建構而成，而上式則將三個不同斜率的基函數改以「逐步疊加」的方式表示，基函數$h_1(x)$合併了三個常數基函數（公式(D.11)）代表整體模型的截距，基函數$h_2(x)$代表整個模型的主斜率，而基函數$h_3(x)$則將節點ξ_1後的斜率疊加上去，並依此類推。這樣的表示方法精簡了原先指標函數的形式（公式(D.12)），使得在實作上特徵於基函數的轉換更容易實踐。

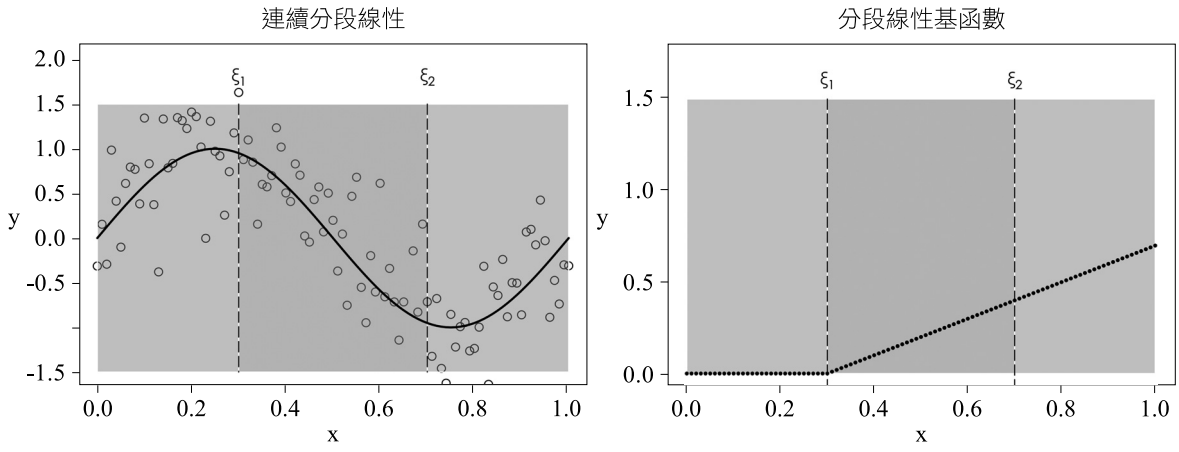

圖 D.9　連續分段線性與分段線性基函數

然而，上述在各區間內以一次多項式的配適顯然還欠缺配適，因此我們可以使用更高階的多項式來增加模型複雜度與平滑度。若我們以一個三次多項式配適各區間，因各區間的常數造成的不連續性使我們令所有「零次項」的基函數相等，為了確保函數的平滑程度，我們可進一步強化函數的一階與二階導數連續性，也就是令各區間的一次項與二次項的基函數相等，使各節點兩邊斜率與凸凹性更加一致，使節點具有高度的平滑性，如圖 D.10 所示。實際上，這就是一個「**三次樣條**」（cubic spline）模型（「迴歸樣條」的特例），我們可將此模型的所有基函數表示如公式(D.14)所示，

$$h_1(x) = 1, \quad h_2(x) = x, \quad h_3(x) = x^2, \quad h_4(x) = x^3,$$
$$h_5(x) = (x - \mathrm{t}_1)_+^3, \quad h_6(x) = (x - \mathrm{t}_2)_+^3 \tag{D.14}$$

也就是原先三次多項式與三個分段的模型一共有$4 \times 3 = 12$個基函數，而我們令零次到二次的多項式相等，扣除了兩個節點中零次到二次的多項式一共有$3 \times 2 = 6$個基函數，因此最後得到了上述6個基函數的「三次樣條」。若以另一個角度想，實際上我們是建構一個不分段的三次多項式模型，而在某個節點後變換其三次項的迴歸係數。

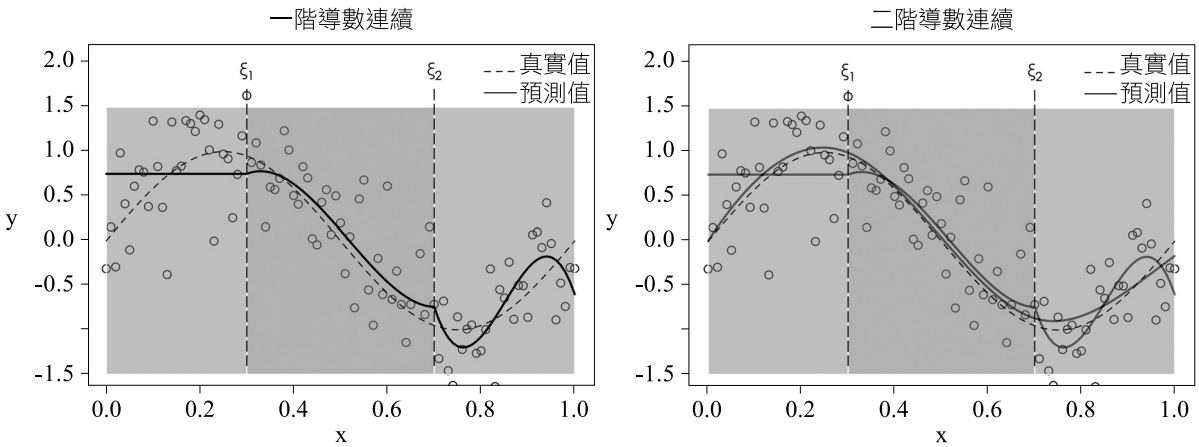

圖 D.10　分段多項式節點的平滑程度

有了上述的思維後，我們便可推廣到k次多項式與q個節點的「迴歸樣條」，此模型的所有基函數如公式(D.15)所示，

$$h_{j+1}(x) = x^j, \qquad j = 0, \dots, k$$
$$h_{k+\mathrm{m}+1}(x) = (x - \mathrm{t}_s)_+^k, \qquad s = 1, \dots, q \tag{D.15}$$

因此，「迴歸樣條」的模型可以表示如公式(D.16)所示。

$$f(x) = \sum_{j=1}^{k+q+1} \beta_j h_j(x) \tag{D.16}$$

其中，$h_1(x), \dots, h_{k+q+1}(x)$為公式(D.15)的所有基函數，而$\beta_1, \dots, \beta_{k+q+1}$為經過基函數轉換後新特徵的迴歸係數。接著，若我們定義基函數矩陣G，此模型的迴歸係數β與函數$f(x)$的估計式可表示如公式(D.17)所示。

$$\hat{\beta} = (G^T G)^{-1} G^T y$$
$$\hat{f}(x) = G(G^T G)^{-1} G^T y \tag{D.17}$$

顯然地，這是一個共有$k + q + 1$個基函數的迴歸模型，而上述提到的「三次樣條」就是$k = 3$的情形，也是常被使用的低階「迴歸樣條」，這是由於「三次樣條」的平滑程度使得節點已足夠難以用肉眼判別了。然而「迴歸樣條」卻存在兩大問題：第一，在邊界（最左節點的左側與最右節點的右側）的估計由於**外插**（extrapolation）且配適高次多項式的緣故，導致估計上產生較大的「變異」(正好與「核迴歸」在邊界上的問題相反，「核迴歸」則是存在估計上的「偏誤」)；第二，q個節點不易挑選，這是因為特徵與目標變數的函數關係不見得有明顯的轉折點，即便使用「**貪婪方法**」（greedy methods）去找尋也是十分耗費運算資源的，因而發展「自然樣條」與「平滑樣條」兩方法就上述兩問題進行改善。

D.4.2 自然樣條

「**自然樣條**」(nature splines）為改善「迴歸樣條」在邊界上具有較大的變異，分別給予邊界兩節點外的區域自由度較小的限制。一個項次為k且節點數為q的「自然樣條」f可表示如下所述：

- **節點內**（interior）：f在區間$[\mathrm{t}_1, \mathrm{t}_2], \dots, [\mathrm{t}_{q-1}, \mathrm{t}_q]$的多項式項次為$k$
- **節點外**（exterior）：f在區間$(-\infty, \mathrm{t}_1], [\mathrm{t}_q, \infty)$的多項式項次為$(k-1)/2$
- **限制**（constraints）：f在所有節點$\mathrm{t}_1, \dots, \mathrm{t}_q$的$1, \dots, k-1$階導數均相同

由於第二項條件的緣故，「自然樣條」只定義在項次k為奇數下，而此模型的基函數個數可計算如公式(D.18)所示（回顧「迴歸樣條」的基函數個數等於$(k+q+1)$）。

$$\underbrace{(k+1)\cdot(q-1)}_{\text{節點內(interior)}}+\underbrace{\left(\frac{(k-1)}{2}+1\right)}_{\text{節點外(exterior)}}\cdot 2-\underbrace{k\cdot q}_{\text{限制(constraints)}}=q \tag{D.18}$$

若我們以項次k為 3 的「自然樣條」為例，其兩邊界的基函數各為$k=2$，因此須滿足線性的限制，也就是說三次的「自然樣條」實際上是邊界被限制為線性的「三次樣條」。

D.4.3 平滑樣條

在「自然樣條」修正了邊界上的變異後，接著介紹的是避開節點挑選的「**平滑樣條**」（smoothing splines）。「平滑樣條」將所有樣本點$x_1,\dots,x_n$全部視為節點配適一個「自然樣條」，並同時以正則化方法進行節點的挑選（可參閱「特徵挑選與維度縮減」章節中的正則化方法）。壓縮模型中基函數的迴歸係數以控制模型複雜度，避免模型的過度配適。若以最佳化的視角來看，輸入為$x_1,\dots,x_n$且介於區間$[0,1]$的「平滑樣條」模型估計如公式(D.19)所示，

$$\hat{f}(x)=\underset{f(x)}{\arg\min}\ \underbrace{\sum_{i=1}^{n}(y_i-f(x_i))^2}_{\text{residual sum of square}}+\underbrace{\lambda\int_0^1 f^{(m)}(x)^2dx,}_{\text{penalty}} \tag{D.19}$$

$$\text{其中}m=\frac{(k+1)}{2}$$

其中$f^{(m)}(\cdot)$為函數f的m階導數，超參數λ是一個大於零的**懲罰**（penalty）係數。在上式中，前者為殘差最小平方代表模型配適程度，後者為懲罰項則為函數導數的平方和，負責控制模型複雜度，使得模型能在變異與偏誤間取得權衡（以平方和作為懲罰權重在正則化方法中屬於脊迴歸的L_2 懲罰）。當懲罰項權重λ趨近於零時，會配適出一個極為複雜的非線性模型，而當懲罰項權重λ趨近於無窮時，則會配適出一個極為簡單的模型，這是由於模型無法容忍任何較高次的函數。若以一個項次$k=3$的「三次平滑樣條」為例，如公式(D.20)所示，

$$\hat{f}(x) = \underset{f(x)}{\arg\min} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 f''(x)^2 dx \tag{D.20}$$

在上式中計算函數的二階導數平方和作為懲罰項。因此，若我們將懲罰權重λ設定趨近於無窮時，此模型會無法容忍任何二階導數因而被限制成一個線性迴歸模型。

若將「**平滑樣條**」的估計（公式(D.19)）以基函數表示（與「自然樣條」相同一共有n個），則其迴歸係數估計式可表示如公式(D.21)所示。

$$\hat{\beta} = \underset{\beta \in \mathrm{R}^n}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \beta_j h_j(x_i) \right)^2 + \lambda \int_0^1 \left(\sum_{j=1}^{n} \beta_j h_j^{(m)}(x) \right)^2 dx \tag{D.21}$$

接著我們定義基函數矩陣H與懲罰項矩陣Ω中的元素如公式(D.22)所示，

$$H_{ij} = h_j(x_i) \text{ and } \Omega_{ij} = \int_0^1 h_i^{(m)}(x) h_j^{(m)}(x) dx \tag{D.22}$$

因此，「平滑樣條」的迴歸係數估計式公式(D.22)可被重新改寫如公式(D.23)所示。

$$\hat{\beta} = \underset{\beta \in \mathrm{R}^n}{\arg\min} \|y - N\beta\|_2^2 + \lambda \beta \Omega \beta \tag{D.23}$$

上述所呈現的是一個「**脊迴歸**」（ridge regression）的形式。以圖 D.11 為例，圖中圓圈為真實的非線性數據，我們配適並比較不同自由度的「平滑樣條」，左圖為自由度為 19〔該自由度是由模型以**線性平滑器**（linear smoothers）表示後計算其線性權重矩陣的跡（trace）所得到的〕所估計出的模型，其僅滿足了左邊區域的平滑度，但右邊區域則被過度平滑了；中圖則為自由度為 50 所估計出的模型，相反地，其僅滿足了右邊區域的平滑度，左邊區域則欠缺平滑。這說明了「平滑樣條」雖能調整模型整體的平滑度，但不易配適不同區域具有不一致平滑度的數據。

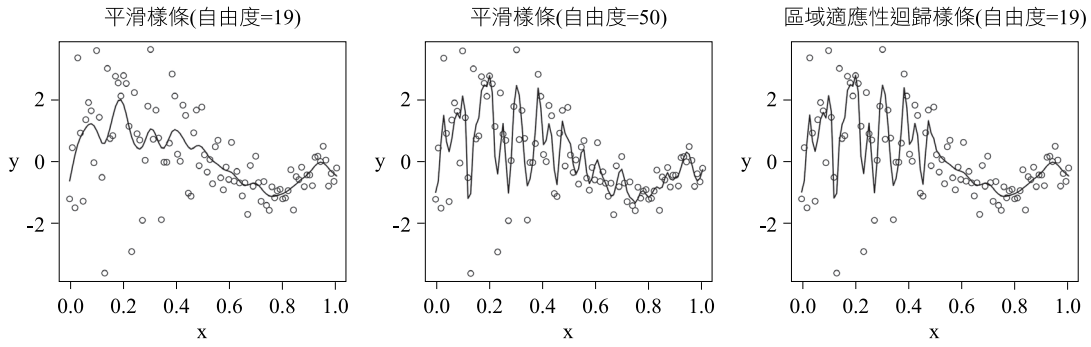

圖 D.11　平滑樣條與區域適應性迴歸樣條的模型配適

D.4.4　區域適應性迴歸樣條

基於「平滑樣條」僅能調整整體平滑度的限制，「**區域適應性迴歸樣條**」（locally adaptive regression splines）則能在每個區域能有不同且更合適的平滑度。若將「區域適應性迴歸樣條」以基函數表示，其迴歸係數估計式可表示如公式(D.24)所示。

$$\hat{\beta} = \underset{\beta \in \mathrm{R}^n}{\arg\min} \underbrace{\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \beta_j h_j(x_i) \right)^2}_{\text{residual sum of square}} + \underbrace{\lambda \mathrm{TV} \left\{ \left(\sum_{j=1}^{n} \beta_j h_j(x) \right)^{(k)} \right\}}_{\text{penalty}} \tag{D.24}$$

其中TV為「**總變異**」（total variation），超參數λ是一個大於零的懲罰係數。我們可發現實際上此模型與「平滑樣條」（公式(D.19)）非常相似，差異在於懲罰項並非模型「導數平方和」，而是模型「導數總變異」。接著，由於基函數是由多個樣條構成的（包含節點$\{\mathrm{t}_1, \ldots, \mathrm{t}_{n-k-1}\}$），其導數以及總變異的推導如公式(D.25)所示。

$$\begin{aligned} \left(\sum_{j=1}^{n} \beta_j h_j(x) \right)^{(k)} &= k! + k! \sum_{j=k+2}^{n} \beta_j \, I\left(x \geq \mathrm{t}_{j-k-1}\right), \\ \Longrightarrow \mathrm{TV} \left\{ \left(\sum_{j=1}^{n} \beta_j h_j(x) \right)^{(k)} \right\} &= k! \sum_{j=k+2}^{n} |\beta_j| \end{aligned} \tag{D.25}$$

因此，「區域適應性迴歸樣條」的迴歸係數估計式公式(D.24)可被重新改寫

如公式(D.26)所示。

$$\hat{\beta} = \underset{\beta \in \mathrm{R}^n}{\arg\min} \|y - H\beta\|_2^2 + \lambda k! \sum_{j=k+2}^{n} |\beta_j| \tag{D.26}$$

上述所呈現的是一個「**套索迴歸**」(lasso regression) 的形式，只有k階多項式的$k+1$個迴歸係數並未進行懲罰。承接圖 D.11 的案例，圖 D.11 右圖則是以自由度為 19 所建構的「區域適應性迴歸樣條」，此模型滿足了左邊區域與右邊區域有著不一樣的平滑程度，與實際函數更加貼近，這解釋了「區域適應性迴歸樣條」能同時滿足不同區域的不同平滑程度。

D.5 加法模型與多變量適應性迴歸樣條

由前述介紹基於「核函數」以及「樣條」的兩大類無母數迴歸模型，已能對「單一特徵」的迴歸問題達到很好的配適能力 (尤其是考慮區域不同平滑度的「區域適應性迴歸樣條」)。然而實際上多數的數據皆包含多個特徵 (也就是多變量)，因此接著討論無母數迴歸模型由單變量拓展至多變量的模型建構。多變量的模型考慮維度為p的所有特徵與目標值的函數關係，可描述如公式(D.27)所示。

$$f(x) = f(x_1, \ldots, x_p) \tag{D.27}$$

D.5.1 加法模型

「**加法模型**」(additive models) 是將模型中特徵與目標值的關係拆解並相加匯總，如公式(D.28)所示，

$$f(x) = f_1(x_1) + \cdots + f_p(x_p) \tag{D.28}$$

其中每個維度的函數f_j可以是任意一種前述的無母數迴歸模型。然而，此模型假設各特徵與目標值之間的關係是相互獨立的，因此受限於無法考慮特徵之間潛藏的「**交互作用**」(interaction)。「加法模型」一般是以「**向後配適演算法**」(backfitting algorithm) 進行建模，此方法首先決定使用的無母數迴歸方法 (可各特徵獨立選擇)；其次，獨立且重複地估計每一個特

徵x_j與部分的目標值r的函數f_j，如公式(D.29)所示，

$$\hat{f}_j = f_j(x_j, r) \tag{D.29}$$

而其中部分的目標值r指的是原始目標值扣除其餘特徵（除特徵j以外）預測值的總和，如公式(D.30)所示。

$$r_i = y_i - \sum_{l \neq j} \hat{f}_l(x_{il}),\ \ i = 1, \ldots, n \tag{D.30}$$

這說明了「向後配適演算法」每次對特定特徵的估計時將扣除其餘特徵的影響，並與剩餘的殘量建模。「向後配適演算法」如表 D.2 所示。

表 D.2　向後配適演算法

演算法　向後配適演算法（Backfitting Algorithm）

輸入：訓練集$S = (x_1, y_1), \ldots, (x_n, y_n)$
輸出：估計的加法模型預測值$\hat{f}$

1. 初始化設定所有特徵模型的預測值皆為零$\hat{f}_j \overset{\text{set}}{=} 0$，平均預測值為$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i$
2 Cycle $j = 1$ to p:
3. 計算原始目標值扣除其餘特徵預測值的和：

$$r_i = y_i - \hat{\alpha} - \sum_{l \neq j} \hat{f}_l(x_{il}),\ \ i = 1, \ldots, n;$$

4. 估計該特徵與部分目標值的無母數迴歸模型：

$$\hat{f}_j = f_j(x_j, r);$$

5. 扣除平均值的殘量：

$$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{n}\sum_{i=1}^{n} \hat{f}_j(x_{ij});$$

6. Until 每個特徵的模型$\hat{f}_j$的改變量小於某個閾值；
7. 加總所有模型得出估計的加法模型：

$$\hat{f}\,(x) = \hat{\alpha} + \hat{f}_1(x_1) + \cdots + \hat{f}_p(x_p);$$

Return 估計的加法模型預測值$\hat{f}$

D.5.2　多變量適應性迴歸樣條

「**多變量適應性迴歸樣條**」（multivariate adaptive regression splines, MARS）屬於「加法模型」的延伸方法，此模型以多變量的迴歸樣條建模，並同時考量特徵之間的「交互作用」，突破「加法模型」原有的限

制。此外，此模型可被視為廣義的「逐步迴歸」(可參閱後續章節「特徵挑選與維度縮減」)以及在「決策樹」在迴歸問題上的變形版(可參閱章節「決策樹與集成學習」)。

「多變量適應性迴歸樣條」同樣是由多個基函數所組成的，更精確地說，是成對的「分段線性基函數」，如公式(D.31)所示。

$$(x-t)_+ = \begin{cases} x-t, & \text{if } x > \mathrm{t} \\ 0, & \text{otherwise} \end{cases} \text{ 與 } (\mathrm{t}-x)_+ = \begin{cases} \mathrm{t}-x, & \text{if } x < t \\ 0, & \text{otherwise} \end{cases} \qquad \text{(D.31)}$$

上式又被稱為「**反射對**」(reflected pair)，而選擇線性的基函數是為了簡化高維度模型的配適與其複雜度。因此，我們可將每個特徵X_j以及其各別樣本x_{ij}作為的節點並建構出一個「反射對」的集合$\mathcal{C}$，如公式(D.32)所示，

$$C = \left\{ \left(X_{\mathrm{j}} - t\right)_+, \left(t - X_j\right)_+ \right\} \text{ 其中 } \mathrm{t} \in \left\{x_{1j}, \dots, x_{Nj}\right\}, j = 1, \dots, p \qquad \text{(D.32)}$$

其中此集合$\mathcal{C}$一共有np個基函數，有了此集合我們便可對這些基函數進行組合與挑選。

「多變量適應性迴歸樣條」的模型配適可分為「**建構**」與「**縮減**」兩階段。首先，在模型的**建構階段**，使用的是「**向前逐步挑選**」(forward stepwise selection)。在每一個階段，我們將候選集合$\mathcal{C}$中的反射對與已建構模型$\mathcal{M}$中的每一個基函數h_l進行乘積「**組合**」(初始為常數基函數$h_0(x) = 1$)，並將他們視為候選反射對，如公式(D.33)所示。

$$\begin{aligned} &\hat{\beta}_{M+1} h_l(x) \cdot \left(X_j - t\right)_+ + \hat{\beta}_{M+2} h_l(x) \cdot \left(t - X_j\right)_+, \\ &\text{其中 } \left\{ \left(X_j - t\right)_+, \left(t - X_j\right)_+ \right\} \in \mathcal{C}, \qquad h_l \in \mathcal{M} \end{aligned} \qquad \text{(D.33)}$$

其中$\hat{\beta}_{M+1}$與$\hat{\beta}_{M+2}$為一候選反射對的迴歸係數，並且是與已建構模型$\mathcal{M}$中其餘基函數的迴歸係數一同估計的結果。接著，將考慮不同候選反射對的模型進行比較，並「**挑選**」使得模型誤差減少最多的反射對加入已建構模型$\mathcal{M}$中，並重複「**組合與挑選**」的流程，直到模型誤差不再減少。舉例而言，假設我們已挑選了其中一組反射對進入模型，例如是第 2 個特徵的第 9 個節點，因此現階段的模型$\mathcal{M}$是由三個基函數h_m所構成的，如公式(D.34)所示，

$$h_0(x) = 1,$$
$$h_1(x) = (X_2 - x_{92}), \tag{D.34}$$
$$h_2(x) = (x_{92} - X_2)$$

因而候選基函數的所有乘積組合（候選集合$\mathcal{C}$與已建構模型$\mathcal{M}$）可表示如公式(D.35)所示，

$$h_m(x) \cdot (X_j - t)_+ \text{ 與 } h_m(x) \cdot (t - X_j)_+, \qquad t \in \{x_{ij}\} \tag{D.35}$$

因此，新的組合可能是單獨的分段線性基函數，也可能是低階或高階的交互作用。假設最後挑選使得模型誤差減少最多的組合為交互作用$(X_1 - x_{31}) \cdot (x_{92} - X_2)$，其視覺化後呈現如圖 D.12 所示，使得能考慮特徵之間（X_1與X_2）同時對目標值的影響。

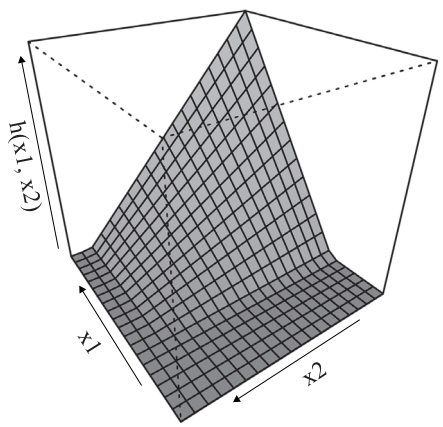

圖 D.12　多變量適應性迴歸樣條

經由上述模型建構後，我們會得到一個極為複雜的模型，這是因為我們僅考慮最小化模型的訓練誤差，因此下一步我們將**縮減**模型（減低模型複雜度）。因此在模型的**縮減階段**，使用的是「**向後逐步挑選**」（backward stepwise selection），也就是從複雜模型中逐一移除每一個基函數來觀察模型誤差的變化，而（樣本內）誤差的評估可使用「**廣義交叉驗證**」（generalized cross-validation, GCV），如公式(D.36)所示，

$$GCV(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(x_i)}{1 - \frac{M(\lambda)}{n}}\right)^2 \quad \text{(D.36)}$$

其中$M(\lambda)$代表有效的參數個數（基函數個數加上節點個數）。因而模型的縮減是以最小化「廣義交叉驗證」作為目標。

最後，我們來討論「多變量適應性迴歸樣條」選用「分段線性基函數」以及「向前逐步挑選」來建構模型的理由。首先，「**分段線性基函數**」具有很強的區域特性，此基函數僅在節點的某一邊發揮其配適模型的效果，而其餘的區域並未受到任何配適的作用，這使其可以很精簡地配適整個迴歸函數空間，也就是只在需要的地方才配置基函數。因此，在高維度的問題中，此基函數能使得模型更有效率且精簡地使用每一個參數（迴歸係數）。倘若使用的是多項式基函數，在某些情況下不僅將導致整體函數過於複雜，同時還會產生大量的參數需被估計。其次，「**向前逐步挑選**」的階層式建模能同時考慮模型單一特徵的主效果以及多個特徵的交互作用，並從低階交互作用逐步考慮至高階（逐步乘積組合），這是基於「**若某個高階交互作用存在時，其低階必定存在**」的假設。舉例來說，當已建構模型中存在某一個二階的交互作用基函數，這時的乘積組合才會考慮到這個二階交互作用基函數與反射對中某一基函數所組合成的三階交互作用基函數。當然上述假設並不一定會總是滿足，但卻能使得「向前逐步挑選」有效率地找出主要交互作用關係，避免耗費於搜尋任意交互作用的大量計算資源。

在實務上，我們一般會增加兩項限制。第一個限制，一樣本（節點）在不同特徵上的反射對彼此間只能出現一次乘積，這是避免模型可能在邊界上產生過度配適。第二個限制，則是規範交互作用最高的階層，一般來說兩階或三階的就足夠配適複雜的函數關係了，此限制也有助於我們對於模型的解釋。而這更加凸顯了「多變量適應性迴歸樣條」在無母數迴歸方法中所扮演的重要角色，它不僅是個能配適複雜且非線性函數關係的方法外，更有機會解釋特徵與目標值之間的關係（包含了交互作用）。

紅酒品質案例：無母數迴歸

承續前章紅酒數據案例，紅酒品質作為目標變數，「酒精度」、「硫酸鹽」、「揮發酸」作為解釋變量，進行無母數迴歸的估計。首先，我們分別以單變量配適「K 最近鄰法」、「核迴歸」、「區域線性迴歸」、「區域適應性迴歸樣條」等方法進行比較，如圖 D.13 所示，由圖可知不論是「**區域線性迴歸**」或「**區域適應性迴歸樣條**」所配適的模型較能以不同的平滑程度描述函數關係。

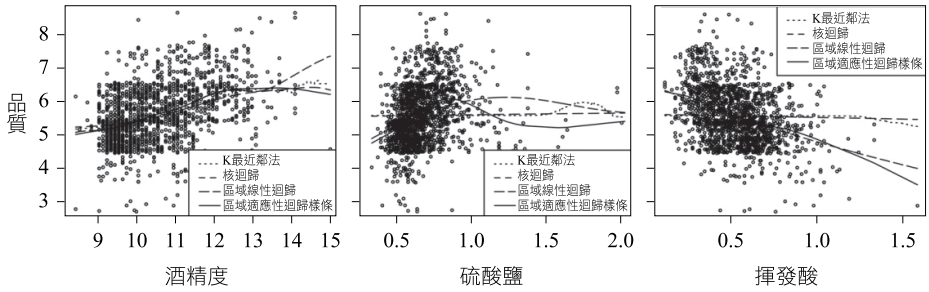

圖 D.13　單變量無母數迴歸於紅酒品質預測

另一方面，若以「多變量適應性迴歸樣條」以五折交叉驗證配適多變量的迴歸模型（也可基於上述方法以「向後配適演算法」求出加法模型），我們取其中一折的函數如公式(D.37)所示，

$$\begin{aligned}\hat{f}(X) = {} & 4.85 + .36(x_{alcohol} - 9.2)_+ - .74(x_{alcohol} - 12.4)_+ \\ & -2.72\big(.76 - x_{sulphates}\big)_+ \\ & -8.06\big(x_{volatile.acidity} - .855\big)_+ \\ & +2.01\big(.855 - x_{volatile.acidity}\big)_+ \\ & -20.5(9 - x_{alcohol})_+ \cdot \big(.76 - x_{sulphates}\big)_+ \\ & +.47(12.4 - x_{alcohol})_+ \cdot \big(.97 - x_{sulphates}\big)_+ \\ & -.37(12.4 - x_{alcohol})_+ \cdot \big(.885 - x_{volatile.acidity}\big)_+ \\ & +3.19(12.4 - x_{alcohol})_+ \cdot \big(x_{volatile.acidity} - 0.885\big)_+ \\ & +4.73(x_{alcohol} - 11)_+ \cdot \big(x_{volatile.acidity} - 0.69\big)_+\end{aligned} \tag{D.37}$$

此模型包含了 10 個基函數（5 個為單變量的分段線性函數、5 個為二階交互作用項），其中單一變量的分段線性可與圖 D.13 對應其模型係數（斜率）的關係，例如酒精度在大於 9.2 後有正相關的特性，但當其大於 12.4 時又為負相關，而實際上此三個特徵彼此間皆存在交互作用的特性，例如酒精度在小於 9 與硫酸鹽在小於 0.76 時具有負相關的特性。若以此模型與線性迴歸模型比較（同樣以五折交叉驗證），如表 D.3 所示（表中括弧內數值為標準差），多變量適應性迴歸樣條模型表現相對較好（考量了特徵的非線性與交互作用），且可由上述公式對模型進行解釋。

表 D.3　多變量適應性迴歸樣條於紅酒品質預測結果

模型＼評估指標	調整後判定係數（Adjusted R^2）	均方根平均平方誤差（RMSE）	對稱平均絕對百分比誤差（sMAPE）
多元線性迴歸	.269 (.045)	.738 (.05)	10.2% (.8%)
多變量適應性迴歸樣條	.296 (.036)	.723 (.05)	10.1% (.8%)

D.6　結語

本章節介紹了非線性模型與無母數迴歸，從「核密度估計」與「K 最近鄰法」瞭解無母數的基礎觀念，透過「核迴歸」與「區域線性迴歸」，說明對核函數進行加權平滑，呈現出無母數的非線性的預測。接著引進基函數對函數進行分段建模，包含「迴歸樣條」、「自然樣條」、「平滑樣條」、「區域適應性迴歸樣條」，以及由單變量延伸至多變量的「加法模型」跟考慮交互作用的「多變量適應性迴歸樣條」。事實上，在數據科學的建模中，無母數迴歸方法扮演的重要性也愈來愈關鍵，一方面改善了線性模型受限於線性以及給定的母體分配假設，甚至可適用於**小樣本**，另一方面提升了模型的預測能力並維持良好的解釋力，其**數學方程式**的估計也提升了參數在物理上的**解釋意義**，對於許多機構的特性與參數系統鑑別（system identification）有所助益。

參考文獻

[1] 鄭宇翔（2020），資料科學於 CNC 機台健康管理之摩擦力估測與補償，國立成功大學製造資訊與系統研究所碩士論文。

[2] Hastie, T., Tibshirani, T., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition, Springer.

[3] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. 2nd Edition, Springer.

問題與討論

1. 試說明線性模型與非線性模型在預測力與解釋力上的權衡為何。（提示：可用偏誤與變異畫圖解釋，何者較容易欠缺配適、何者又較容易過度配適）
2. (a)試解釋有母數與無母數方法的差異為何？試使用網路開放數據，分別繪製示意圖說明(b)核密度估計；(c)K 最近鄰迴歸；(d)K 最近鄰分類器。（提示：前兩者以單一特徵為例，後者以兩特徵為例）
3. 在 K 最近鄰法中，試使用網路開放數據，說明以下問題：(a)當 K 值愈大其決策邊界會發生什麼變化？試繪圖說明之。(b)試固定樣本數（試從數據中挑選適當的樣本數，例如$n = 40$），當放入資料中的特徵增加到某種程度時（i.e.維度增加），會發生維度的詛咒。試用數據建立實驗或模擬來說明維度詛咒的影響。(c)試說明該方法的優缺點。
4. 試繪製示意圖說明：(a)核迴歸；(b)區域線性迴歸（提示：以單一特徵為例，並任意選定一個樣本描繪其被賦予的權重）。(c)為何後者能修正邊界核函數的截斷並且具有一階的不偏估計？
5. 樣條是以分段多項式去建構區域與平滑的非線性模型，試說明在迴歸樣條中，(a)分段線性基函數$(x-\xi)_+$是什麼？(b)令所有小於三次方基函數相等的用意為何？（提示：試畫圖說明）(c)而迴歸樣條方法存在什麼樣潛在的問題？
6. 試說明(a)平滑樣條與區域適應性迴歸樣條在正則化方法中的關係為何？（提示：何者具有特徵挑選的功能）(b)何者更適用於具有不同平滑程度的函數？為什麼？
7. 有關多變量適應性迴歸樣條，試說明以下問題：(a)該方法如何處理高維度特

徵？(b)該方法如何考慮交互作用特性？（提示：與模型以及其配適步驟有關）

8. 在 UCI Machine Learning Repository 開放數據中包含了一個具有紅酒與白酒的酒品質數據（wine quality dataset，https://archive.ics.uci.edu/ml/datasets/Wine+Quality），其中紅酒數據一共包含了 1,599 個觀測值，而每個觀測值具有 11 個特徵以及作為目標值的紅酒品質。試著參考網路資源學習並撰寫程式，使用此數據回答下列問題：

   (a) 試以多變量適應性迴歸樣條建構模型進行預測。

   (b) 承接(a)的結果，與線性迴歸模型進行比較。（提示：試討論模型預測效果、解釋性、計算複雜度等）