附錄B

支持向量機

B.1 最大邊際分類器
B.2 支持向量分類器
B.3 支持向量機
B.4 支持向量迴歸
B.5 支持向量數據描述
B.6 結語

李家岩、洪佑鑫（2022），製造數據科學：邁向智慧製造與數位決策，前程文化。

本章介紹的是「**支持向量機**」(support vector machine, SVM)以及支持向量家族等方法，包含了線性可分的「**最大邊際分類器**」(maximal margin classifier)，與拓展前者至可容許誤差的「**支持向量分類器**」(support vector classifier)，以及最終拓展至非線性的「支持向量機」(Chang and Lin, 2011; Hastie et al., 2009)。同時，我們也將說明「支持向量機」如何從二元分類拓展至多元分類以及迴歸問題，以及它與線性模型中的「羅吉斯迴歸」之間的關係。此外，此章節也將介紹從「支持向量機」延伸出的一分類方法，「**支持向量數據描述**」(support vector data description, SVDD)，其常見應用為製造現場裡的無母數品管圖。

回顧章節「線性分類器」，我們將分類模型分為「判別模型」與「生成模型」，而支持向量機的核心思維相當直觀，此模型期望能精準地找出分類的判別界線，因此它屬於「判別模型」的一種。更精確地說，在高維度的空間中，此判別界線會是一個「超平面」，因此藉由數學式的最佳化求解出具有「**最大邊際**」(maximum margin)的「**最佳超平面**」(optimal hyperplane)便是此模型建構的方式。如圖 B.1 所示，這是一個兩個特徵的二元分類問題，其中的最佳超平面就是我期望建構出的「支持向量機」，而在邊際上的樣本就是「**支持向量**」(support vector)。

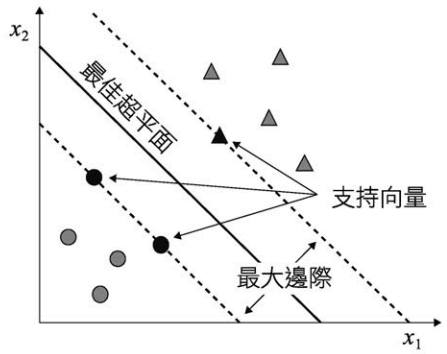

圖 B.1　最大邊際、超平面與支持向量

B.1　最大邊際分類器

首先介紹第一個支持向量機家族的方法為「**最大邊際分類器**」(maximal margin classifier)，此方法雖僅能處理「**線性可分**」(linearly

separable）的數據，但它是理解超平面以及以超平面建構分類器的重要第一步。

B.1.1 超平面與分類

在一個維度為p的空間中，超平面為一個將空間劃分成兩半且維度為$p-1$的平面。舉例而言，在維度$p=2$的平面空間中，我們可以找到任意一條維度$p=1$的線將平面空間分割；在維度$p=3$的體積空間中，我們可以找到任意一個維度$p=2$的平面將體積空間分割；而維度$p>3$的空間雖無法視覺化，但依舊能找出分割空間的超平面。若以數學式表示超平面，在維度$p=2$的空間中，一個超平面可定義如公式(B.1)所示，

$$\beta_0+\beta_1 x_1+\beta_2 x_2=0 \tag{B.1}$$

其中β_0,β_1,β_2為參數，其中滿足上式的輸入$x=(x_1,x_2)$將構成一條直線分割該二維空間，如圖 B.1 所示。而當輸入不滿足此式時有兩種情形，其中一種是「大於零」另一種為「小於零」，分別表示落於該分割線的某一邊。若我們將如公式(B.1)拓展至p個維度，可表示如公式(B.2)所示。

$$\beta_0+x_i^T\beta=0 \tag{B.2}$$

有了超平面分割空間的概念與數學式後，下一步我們接著說明如何以超平面進行分類問題。

在二元分類問題中，我們將數據定義為一個具有維度為p且樣本數為n的一個X矩陣，並且每一樣本的目標值可被定義為二元的類別分別表示為$y_i=1$與$y_i=-1$。以圖 B.2 為例，此圖為在給定數據中能完美分類的超平面，實際上有無數個超平面能完美地分類這些樣本，若給定其中一個完美分類的超平面下，對任意測試樣本均可以該超平面作為分類的判斷，而我們找出的超平面所具有的性質如公式(B.3)所示。

$$\begin{cases}\beta_0+x_i^T\beta>0 \ \text{ if } y_i=1,\\ \qquad\qquad\qquad\qquad \text{for all } i=1,\dots,n\\ \beta_0+x_i^T\beta<0 \ \text{ if } y_i=-1,\end{cases} \tag{B.3}$$

上式能更精簡地表示如公式(B.4)所示。

$$y_i\left(\beta_0 + x_i^T\beta\right) > 0 \quad \text{for all } i = 1, \ldots, n \tag{B.4}$$

也就是說，在完美分類的情形下，在超平面上方的類別應為1且下方的類別應為-1，滿足乘積大於零的條件。

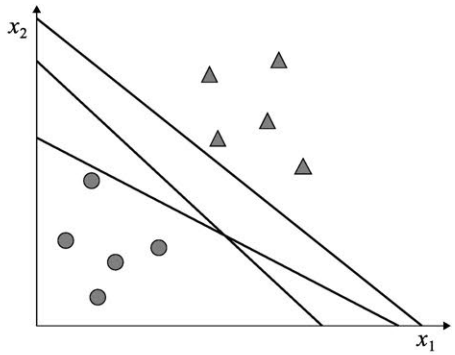

圖 B.2　超平面於二元分類問題

B.1.2　模型建構

然而，在無限多個皆能完美分類的超平面中，我們需要一個精確且合理的方式定義一個「最佳超平面」。若以數據與超平面之間的關係來看，理論上我們期望在分類邊際的樣本應與超平面的距離越遠越好，這代表著分類的結果更加準確。基於這樣的想法，樣本與超平面之間的距離定義為所有訓練樣本與超平面之間的最短距離（垂直於超平面的距離），我們稱之為「邊際」(margin)。因此，一個「最佳超平面」應具有最大的「邊際」，如圖 B.3 所示（延續圖 B.2 的數據），實線為最佳超平面，我們可以看到圖中兩圓形樣本以及一三角形樣本與超平面間的邊際（最短距離）支撐著該平面。若此超平面經任意地旋轉其邊際只會變得更小，而這個最大化邊際的超平面即是「最大邊際分類器」。接著我們從數學上定義這個最佳化問題。

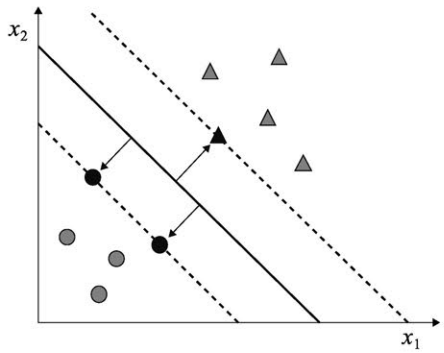

圖 B.3　最大化邊際定義最佳超平面

「最大邊際分類器」的最佳化數學式可表示如公式(B.5)所示，

$$\begin{aligned} &\max_{\beta_0,\beta_1,\ldots,\beta_p} M \\ &\text{subject to } y_i\left(\beta_0 + x_i^T\beta\right) \geq M \quad \text{for } i = 1,\ldots,n \\ &\sum_{j=1}^{p} \beta_j^2 = \|\beta\| = 1 \end{aligned} \tag{B.5}$$

其中M為邊際，也就是我們要最大化的目標，但同時要滿足兩個限制。第一個限制約束了每個樣本需落於超平面拓展出的邊際的正確分類那一側，相較於原先僅考慮超平面的限制（公式(B.4)），此模型將決策邊界拓展（平移M單位）至邊際上。其次，由於我們求出的所有係數β同乘於某個常數後均依舊能滿足第一個限制，因而第二個限制是為了避免係數的多重解產生。實際上，我們可將上述「最大邊際分類器」的最佳化問題轉換另一種更簡單的形式，如公式(B.6)所示。

$$\begin{aligned} &\min_{\beta_0,\beta_1,\ldots,\beta_p} \|\beta\| \\ &\text{subject to } y_i\left(\beta_0 + x_i^T\beta\right) \geq 1 \quad \text{for } i = 1,\ldots,n \end{aligned} \tag{B.6}$$

我們將原先公式(B.5)的第二個限制鬆綁，並將原先第一個限制的邊際M替換成1，這使得真正的邊際轉嫁到係數β中。因此，我們可將邊際等價於係數平方和倒數（$M = 1/\|\beta\|$），便將原先最大化問題轉換成最小化問題。此外，由於此問題屬於凸性最佳化問題（二次的目標式與線性的限制式），因而在求解上的方法相對完善。

然而，在可被完美分類的數據下，我們能求解出「最大邊際分類器」模型，然而，真實的大多數數據必定存在誤差使得數據無法完美分類，因而無法滿足最佳化問題（公式(B.6)）中的限制找出任何的可行解。因此，如何找出一個能容忍部分誤差的超平面是下一個關鍵〔由「硬邊際」（hard margin）轉換成「**軟邊際**」（soft margin）〕。

B.2 支持向量分類器

經由上述對「最大邊際分類器」的介紹後，我們瞭解了此模型的限制在於無法容忍任何誤差，因而對噪音與離群值樣本相當敏感，使模型有很大的變異產生過度配適的情形。為了建構一個較為穩健且一般化的模型，我們期望模型不敏感於任何單一樣本且能容忍部分樣本的分類誤差，也就是我們要介紹的「**支持向量分類器**」（support vector classifier）。

B.2.1 誤差容忍與軟邊際

相較於「最大邊際分類器」屬於一個「硬邊際」模型，「支持向量分類器」則是一個「軟邊際」模型。在求解最大邊際的問題中，前者硬邊際不僅對樣本的限制是需落於超平面分類正確的一側，還需落於正確邊際的外側；後者軟邊際則是放寬了硬邊際的兩個限制，可以容忍部分樣本落於邊際內，甚至是錯誤分類的一側。

假設數據類別在特徵空間上存在交錯重疊的部分，我們需建構一個能容忍部分誤差的「超平面」，因此我們將原先「最大邊際分類器」的「硬限制」（hard constraint）（公式(B.5)）調整為「軟限制」（soft constraint）如公式(B.7)所示。

$$y_i\left(\beta_0 + x_i^T\beta\right) \geq M\ (1-\xi_i)\quad \text{for } i = 1, \ldots, n \tag{B.7}$$

其中ξ_i為每一個樣本與邊際的誤差（寬鬆變數），表示與原先邊際相距的倍數。例如當$\xi_i = 0$時，右式等於邊際大小M（落於邊際上以及其外側）；當$\xi_i = 1$時，右式等於零（落於決策邊界上）；而當$\xi_i > 1$時，右式等於負值（落於錯誤分類的一側）。這樣的軟限制使得每一個樣本能有不同的誤差。此外，此誤差ξ_i需具備兩個限制如公式(B.8)所示。

李家岩、洪佑鑫（2022），製造數據科學：邁向智慧製造與數位決策，前程文化。

$$\xi_i \geq 0 \text{ 與 } \sum_{i=1}^{n} \xi_i \leq C \quad \text{for } i = 1, \ldots, n \tag{B.8}$$

第一個限制定義了誤差的最小值為 0，這意味著所有落於邊際上以及其外側的正確分類樣本誤差均定義為 0；而第二個限制定義了誤差的總和不能大於**軟邊際常數**C，這意味著我們所能容忍的總誤差為多少。當此常數C越小時越無法容忍誤差，因而設定過小時將導致「過度配適」；相反地，當此常數C越大時越能容忍誤差，因而設定過大時將導致「欠缺配適」。如圖 B.4 所示，左圖為給定較小的常數C，使得模型具有較小的邊際只容忍少數的誤差；相反地，右圖為給定較大的常數C，使得模型具有較大的邊際能容忍更多的誤差。因此常數C的大小決定了模型「**偏誤與變異的權衡**」，是一個需被最佳化的超參數。

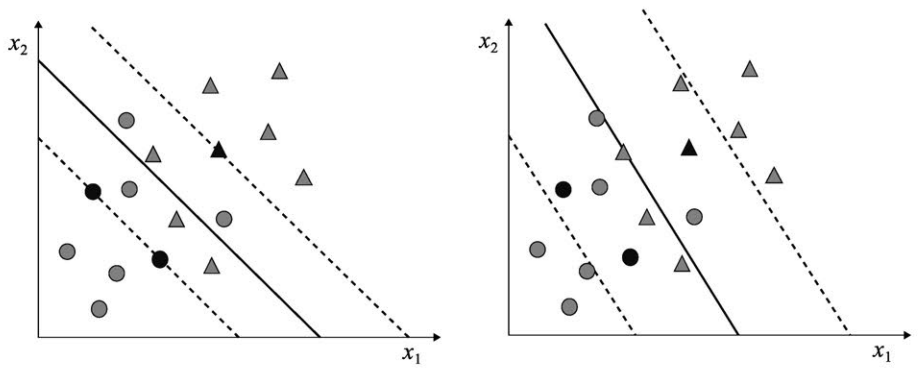

圖 B.4　支持向量分類器於不同的容忍誤差

B.2.2　模型建構

為考量模型對誤差的容忍，增加上述限制與調整的「支持向量分類器」其最佳化數學式如公式(B.9)所示（可對照「最大邊際分類器」的最佳化數學式(B.5)）。

$$\begin{aligned}
&\max_{\beta_0,\beta_1,\ldots,\beta_p} M \\
&\text{subject to } y_i\left(\beta_0 + x_i^T\beta\right) \geq M\left(1-\xi_i\right) \quad \text{for } i = 1,\ldots,n \\
&\xi_i \geq 0 \text{ and } \sum_{i=1}^{n}\xi_i \leq C \quad \text{for } i = 1,\ldots,n \\
&\sum_{j=1}^{p}\beta_j^2 = 1
\end{aligned} \tag{B.9}$$

同樣地，可將上述「支持向量分類器」的最佳化問題轉換另一種更簡單的形式表示，如公式(B.10)所示。（可對照「最大邊際分類器」簡化後的最佳化數學式(B.6)）

$$\begin{aligned}
&\min_{\beta_0,\beta_1,\ldots,\beta_p} \|\beta\| \\
&\text{subject to } y_i\left(\beta_0 + x_i^T\beta\right) \geq 1-\xi_i \quad \text{for } i = 1,\ldots,n \\
&\xi_i \geq 0 \text{ and } \sum_{i=1}^{n}\xi_i \leq C \quad \text{for } i = 1,\ldots n
\end{aligned} \tag{B.10}$$

其次，我們還能更進一步地將「誤差和」的限制式轉換到目標式中，如公式(B.11)所示，

$$\begin{aligned}
&\min_{\beta_0,\beta_1,\ldots,\beta_p} \|\beta\| + C\sum_{i=1}^{n}\xi_i \\
&\text{subject to } y_i\left(\beta_0 + x_i^T\beta\right) \geq 1-\xi_i, \xi_i \geq 0 \quad \text{for } i = 1,\ldots,n
\end{aligned} \tag{B.11}$$

由於「誤差和」必須小於某個常數C，我們可將其從限制式中加到最小化的目標式中作為一個具有懲罰權重為C的懲罰項。此模型視覺化後呈現如圖 B.5 所示，同樣地為一個凸性最佳化問題（二次的目標式與線性的限制式）。

李家岩、洪佑鑫（2022），製造數據科學：邁向智慧製造與數位決策，前程文化。

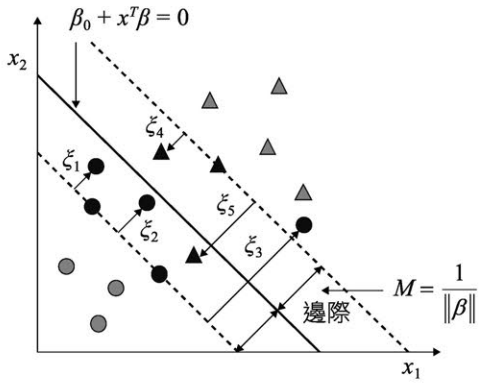

圖 B.5　支持向量分類器轉為最小化問題

B.2.3　模型最佳化求解

對於「支持向量分類器」的最佳化問題而言，相對於這樣的「**主問題**」(primal problem)，求解其「**對偶問題**」(dual problem)一般能具有更好的計算效率以及具備更合適作為特徵轉換的形式（後續會詳細說明原因）。

B.2.3.1　對偶問題推導

我們可將主問題（公式(B.12)）先改寫成「**拉格朗日函數**」(Lagrange function)的形式（將所有限制式轉換到目標式中，與前述公式(B.11)對誤差和的轉換使用的是相同的邏輯），如公式(B.12)所示，

$$L_P = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\beta_0 + x_i^T\beta\right) - \left(1-\xi_i\right)\right] - \sum_{i=1}^{n}\mu_i\xi_i \quad \text{(B.12)}$$

其中我們期望最小化的參數是其中的係數β, β_0與誤差ξ_i。若對上式的各個參數偏微分並令它們導數為零可得限制式如公式(B.13)、公式(B.14)與公式(B.15)所示，

$$\frac{\partial L_P}{\partial \beta} = \beta - \sum_{i=1}^{n}\alpha_i y_i x_i \overset{\text{set}}{=} 0 \implies \beta = \sum_{i=1}^{n}\alpha_i y_i x_i \quad \text{(B.13)}$$

$$\frac{\partial L_P}{\partial \beta_0} = \sum_{i=1}^{n}\alpha_i y_i \overset{\text{set}}{=} 0 \Longrightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \quad \text{(B.14)}$$

$$\frac{\partial L_P}{\partial \xi_i} = C - \alpha_i - \mu_i \overset{\text{set}}{=} 0 \Rightarrow \alpha_i = C - \mu_i \quad \text{for } i = 1, \ldots, n \tag{B.15}$$

以及參數的非負限制式$\alpha_i, \mu_i, \xi_i \geq 0$。因此，將上述公式代回公式(B.12)的目標式，可得到「支持向量分類器」的對偶問題如公式(B.16)所示。

$$\begin{aligned} &\max_{\alpha_i} L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} y_i y_{i'} x_i^T x_{i'} \\ &\text{subject to } \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C \end{aligned} \tag{B.16}$$

此外，除了上述限制式外，KKT **條件式**（Karush-Kuhn-Tucker condition）所包含的限制式如公式(B.17)、公式(B.18)與公式(B.19)所示。

$$\alpha_i \left[y_i \left(\beta_0 + x_i^T \beta \right) - \left(1 - \xi_i \right) \right] = 0 \text{ (complementary slackness)} \tag{B.17}$$

$$\mu_i \xi_i = 0 \tag{B.18}$$

$$y_i \left(\beta_0 + x_i^T \beta \right) - \left(1 - \xi_i \right) \geq 0 \quad \text{for } i = 1, \ldots, n \tag{B.19}$$

B.2.3.2 支持向量與對偶問題的優點

若我們看到在對偶問題中的第一個關係式公式(B.17)，可發現此限制可作為「主問題與對偶問題」在求解估計最佳參數的橋樑，此估計式可表示如公式(B.20)所示，

$$\widehat{\beta} = \sum_{i=1}^{n} \widehat{\alpha}_i y_i x_i \tag{B.20}$$

也就是說主問題所需估計的參數為$\widehat{\beta}$，其參數個數為p；而對偶問題所需估計的參數為$\widehat{\alpha}$，其參數個數為n。然而在對偶問題中，實際上我們不需估計所有的參數$\widehat{\alpha}$，唯有在觀測值落於公式(B.19)限制式的下界為零時（觀測值落於邊際或錯誤分類的區域時），為滿足公式(B.17)限制式其觀測值參數α_i才會是非零參數。而這些觀測值就是所謂的「支持向量」（support vector），也就是在超平面的兩側邊際上或裡面的重要觀測值向量，因此僅需估計支持向量的對偶問題計算效率高於原有的主問題。若回顧圖 B.5 的

案例，圖中包含了五個落於或超出邊際的支持向量。此外，對偶問題的另一項優點在於目標式中特徵的內積$x_i^T x_{i'}$，使得後續我們將介紹的非線性特徵轉換容易且合適。

然而在實務上，數據通常具有「**線性不可分割**」的特性（非線性），若以「支持向量分類器」建模將產生一個分類效果極差的模型，因而如何建構出一個非線性的超平面，也就是如何將原有的特徵拓展到一個新的非線性空間將是下一個關鍵。

B.3 支持向量機

經由上述對「支持向量分類器」的介紹後，我們瞭解此模型雖能容忍樣本部分的誤差，然而卻無法對「線性不可分割」的數據有很好的分類效果。因此，我們要介紹的「**支持向量機**」（support vector machine, SVM）是對特徵空間進行非線性的轉換，使得數據拓展至高維度後從「線性不可分割」變為「線性可分割」，從而找出合適的超平面進行分類。

B.3.1 非線性的決策邊界

依據前章節「無母數迴歸」的思維中，我們使用了大量的「基函數」對原始特徵做非線性的轉換。同理，若我們可選定M個「基函數」並將原始特徵x放入這些函數中我們將得到新特徵$h(x) = \left(h_1(x), \ldots, h_M(x)\right)$，因而在高維度空間中我們可找出合適的一個線性超平面，倘若以原始特徵的視角看高維度空間時則為一個非線性超平面。然而當我們拓展了特徵空間後，將使得維度大幅度地擴張，這將導致計算複雜度變得非常大。其次，若過度的使用基函數也將導致模型過於複雜而發生過度配適的情形。因此以下我們將說明「支持向量機」是如何將模型非線性的「特徵拓展」，以及解決「計算複雜度」與「正則化」的兩大議題。

B.3.2 模型建構

「支持向量機」拓展特徵的方式採用的是「**核技巧**」（kernel trick），也就是將原始特徵以非線性的核技巧投影至高維度的特徵空間中。如圖 B.6 所示，左圖為原始二個維度的「線性不可分割」數據，以核技巧投影至三個維度的特徵空間後呈現如右圖，即可找出一個最佳的線性超平面。

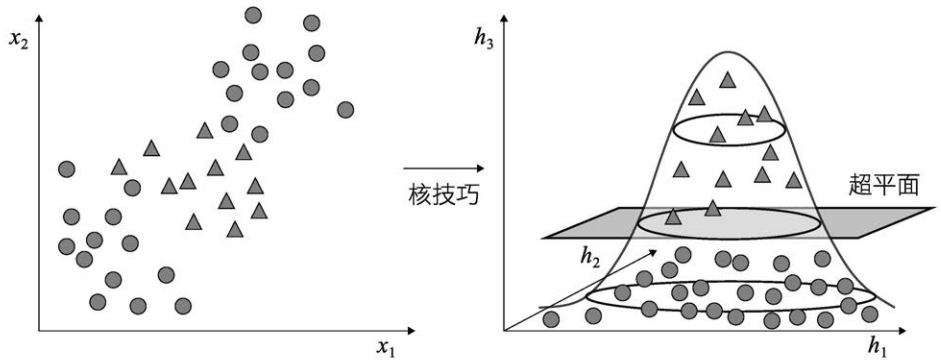

圖 B.6　核技巧的非線性轉換

B.3.2.1　特徵維度的拓展

回顧「支持向量分類器」的對偶問題（公式(B.16)），我們可發現其中目標式包含了特徵的內積$x_i^T x_{i'}$，因此，我們直接地對特徵x以基函數$h(x)$轉換後可改寫如公式(B.21)所示。

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n} \alpha_i \alpha_{i'} y_i y_{i'} \langle h(x_i), h(x_{i'}) \rangle \tag{B.21}$$

同時，這也代表原有的超平面可以改寫如公式(B.22)所示。

$$\begin{aligned} f(x) &= h(x)^T \beta + \beta_0 \\ &= \sum_{i=1}^{n} \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0 \end{aligned} \tag{B.22}$$

根據上述公式(B.21)與公式(B.22)，我們可發現基函數均涉及了內積，因而我們將「基函數的內積」以核函數K的形式表示，而不需指明各別的基函數為何，如公式(B.23)所示，

$$\langle h(x), h(x') \rangle = K(x, x') \tag{B.23}$$

而這個核函數直接將特徵的內積轉換至非線性的空間中，而這樣的技巧即是「核技巧」。常見的核函數有「**d次多項式核**」（dth-degree polynomial kernel）與「**徑向基函數核**」（radial basis function kernel, RBF）〔也被稱為「**高斯核**」（Gaussian kernel）〕，如公式(B.24)與公式(B.25)所示。

$$d\text{th} - \text{degree polynomial: } K(x, x') = (1 + \langle x, x' \rangle)^d \tag{B.24}$$

$$\text{radial basis function（Guassian）: } K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right) \tag{B.25}$$

若以兩個特徵$X = X_1, X_2$的「二次多項式核」為例，如公式(B.26)所示，

$$\begin{aligned} K(X, X') &= (1 + \langle X, X' \rangle)^2 \\ &= (1 + X_1 X_1' + X_2 X_2')^2 \\ &= 1 + 2X_1 X_1' + 2X_2 X_2' + (X_1 X_1')^2 + (X_2 X_2')^2 + 2X_1 X_1' X_2 X_2' \end{aligned} \tag{B.26}$$

經核函數轉換後，從原本的二個特徵轉換成新的六個特徵（基函數），分別為上式各項進行開根號，如公式(B.27)所示。

$$\begin{aligned} &h_1(X) = \sqrt{1}\,,\ \ h_2(X) = \sqrt{2}X_1\,,\ \ h_3(X) = \sqrt{2}X_2\,, \\ &h_4(X) = X_1^2\,,\ \ h_4(X) = X_2^2\,,\ \ h_6(X) = \sqrt{2}X_1 X_2 \end{aligned} \tag{B.27}$$

我們能將經由上述核技巧轉換後的超平面改寫如公式(B.28)所示。

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x, x') + \beta_0 \tag{B.28}$$

以一個線性不可分割的數據為例，如圖 B.7 所示，左圖為選擇「三次多項式核」的模型；右圖則為選擇「徑向基函數核」的模型，其中實線為決策邊界而虛線為邊際。而這兩種核函數的模型均能有效地分類具有非線性決策邊界的數據。

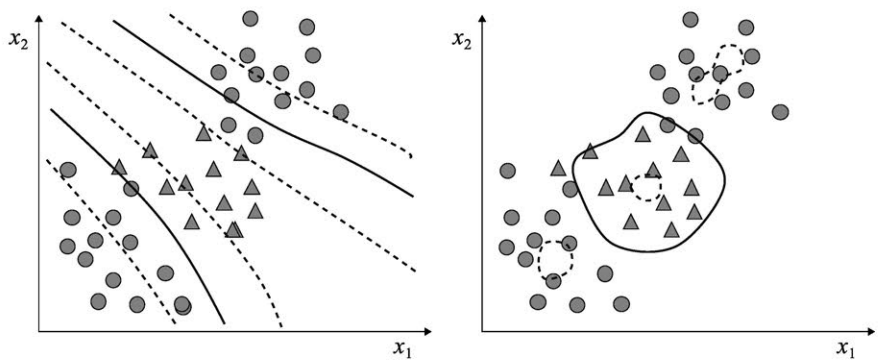

圖 B.7　三次多項式核與徑向基函數核

B.3.2.2 計算複雜度

在前述「支持向量分類器」的最佳化求解過程中，我們提到了對偶問題相較於主問題的優點在於僅需求解出少數不為零的「支持向量」。因此當我們使用「支持向量機」時，即便將特徵拓展至很高的維度對計算效率也不會有太大的影響。

B.3.2.3 正則化

在「正則化」方法中（請參閱章節「特徵挑選與維度縮減」的正則化方法）的最佳化形式，通常會表示成一個最小化損失函數加上一個懲罰項，使得模型能避免模型過於複雜而導致過度配適。對於「支持向量機」的最佳化問題，我們可將其整理如公式(B.29)所示。

$$\begin{aligned}
&\min_{\beta_0,\beta}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i, \text{subject to } y_i\left(\beta_0 + h(x_i)^T\beta\right) \geq 1-\xi_i, \xi_i \geq 0 \\
&= \min_{\beta_0,\beta}\frac{1}{C}\|\beta\|^2 + \sum_{i=1}^{n}\left[1 - y_i\left(\beta_0 + h(x_i)^T\beta\right)\right]_+ \\
&= \min_{\beta_0,\beta}\underbrace{\sum_{i=1}^{n}[1-y_i f(x_i)]_+}_{\text{損失(loss)}} + \underbrace{\lambda\|\beta\|^2}_{\text{懲罰(penalty)}}
\end{aligned} \tag{B.29}$$

可發現上式同樣是由一個損失函數與懲罰項（懲罰權重$C = 1/\lambda$）所組成，其中符號$[\,]_+$表示該函數只取大於零的值〔此損失函數又被稱為「**鉸接損失函數**」(hinge loss function)〕。因此，我們可以理解不論是「支持向量分類器」或「支持向量機」本身均具備了正則化的能力，而此兩者的二次方懲罰項就是「**脊迴歸**」的L_2懲罰，同樣地，我們也能將懲罰項改為「**套索迴歸**」的L_1懲罰，使模型具備特徵挑選的能力，如公式(B.30)所示。

$$\min_{\beta_0,\beta}\underbrace{\sum_{i=1}^{n}\left[1-y_i f(x_i)\right]_+}_{\text{loss}} + \underbrace{\lambda\|\beta\|_1}_{\text{penalty}} \tag{B.30}$$

因而即便「支持向量機」將特徵拓展至高維度空間，使用L_1懲罰也能自動化地挑選出合適的重要特徵。此外，我們也可將懲罰項替換為「**彈性網路**」(elastic net)的形式。

李家岩、洪佑鑫（2022），製造數據科學：邁向智慧製造與數位決策，前程文化。

鋼板缺陷案例：支持向量機

承續前章鋼板缺陷數據案例，我們對鋼板缺陷類型與兩特徵使用支持向量機進行分類。首先，以五摺交叉驗證的測試集平均表現如表 B.1 所示（此部分設定超參數懲罰權重$C = 10$與徑向基函數核的帶寬$\gamma = 0.5$，實際上可進一步進行超參數最佳化），其中準確率與 F1 分數均大於九成，沒有類別不平衡的問題存在。其次，若我們取其中一組測試集的混淆矩陣與其視覺化結果，分別如表 B.2 與圖 B.8 所示，圖中顯示了支持向量機的核技巧使得決策邊界具備非線性的特性。

表 B.1　支持向量機的預測結果

模型評估指標（evaluation metric）	平均數（coefficient）	標準差（standard deviation）
準確率（accuracy）	0.959	0.021
F1 分數（F1 Score）	0.961	0.021

表 B.2　支持向量機的測試集混淆矩陣

實際＼預測	刮痕	凹凸不平
刮痕	72	4
凹凸不平	6	76

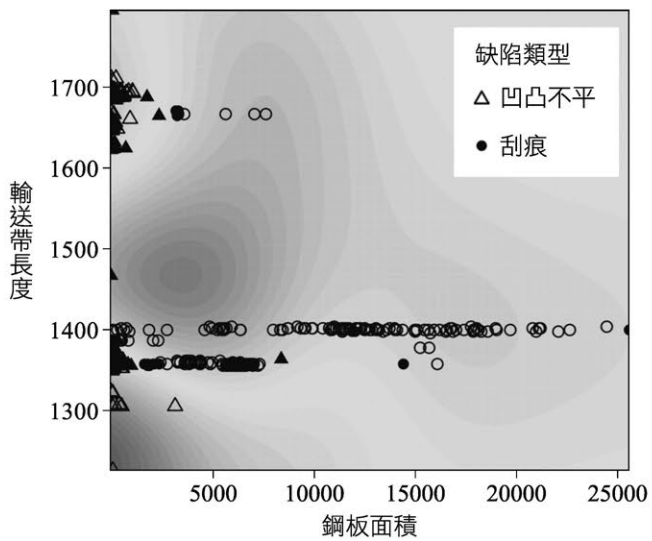

圖 B.8　支持向量機於鋼板缺陷分類

B.4 支持向量迴歸

經由上述對「支持向量機」的介紹後，我們瞭解在分類問題中，該模型是如何容忍樣本部分的誤差以及將特徵拓展至高維度空間以找出線性可分割的超平面。延續這樣的思維我們將分類問題重新定義為一個迴歸問題，建構一個「**支持向量迴歸**」(support vector regression, SVR)。

B.4.1 損失函數

將分類問題的「支持向量機」轉為迴歸問題的「支持向量迴歸」的關鍵就在於「損失函數」。在分類模型中我們期望找一個超平面能分類不同類別的樣本，而在迴歸模型中則期望找一個超平面能配適連續的目標值，因此，我們需要找出一個合適的損失函數衡量超平面與目標值的距離。若以一個經核技巧轉換的超平面$f(x)$為例，我們可根據前述「支持向量機」以L_2正則化形式將迴歸問題的目標函數表示如公式(B.31)所示。

$$\min_{\beta_0,\beta} \underbrace{\sum_{i=1}^{n} V_\varepsilon\big(y_i - f(x_i)\big)}_{loss} + \underbrace{\lambda\|\beta\|^2}_{penalty} \tag{B.31}$$

$$\text{其中 } V_\varepsilon(r)\begin{cases}0, & \text{if} |r| \le \varepsilon,\\ |r|-\varepsilon, & \text{otherwise.}\end{cases} \text{ and } f(x) = h(x)^T\beta + \beta_0$$

其中r為實際值與預測值的誤差（$y_i - f(x_i)$），V_ε則為一個「誤差密集損失函數」(ε-intensive loss function)，也就是當誤差小於可容忍的誤差ϵ時，其損失值為零；反之，誤差大於可容忍的誤差ε時，則損失值為絕對值的誤差減去容忍誤差（$|r| - \varepsilon$）。此容忍誤差與「支持向量機」邊際的概念十分相似，而「支持向量迴歸」模型呈現如圖 B.9 所示。

B.4.2 模型建構

以「誤差密集損失函數」所建構的「支持向量迴歸」可表示如公式(B.32)所示（依照上述「最佳化問題求解」中所介紹的「拉格朗日函數」進行推導）。

李家岩、洪佑鑫（2022），製造數據科學：邁向智慧製造與數位決策，前程文化。

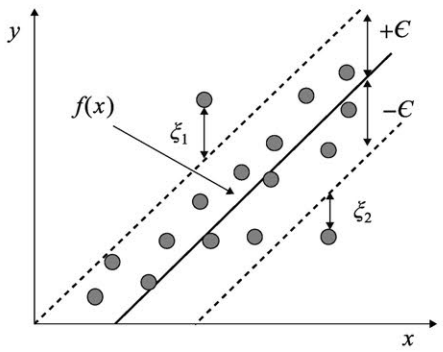

圖 B.9　支持向量迴歸

$$\begin{aligned}
&\min_{\alpha_i,\alpha_i^*} \epsilon \sum_{i=1}^{n}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) \\
&+\frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n}(\alpha_i^* - \alpha_i)\left(\alpha_{i'}^* - \alpha_{i'}\right)\langle h(x), h(x')\rangle \\
&\text{subject to } \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) = 0 \\
&0 \le \alpha_i, \alpha_i^* \le \frac{1}{\lambda}, \text{and } \alpha_i\alpha_i^* = 0 \quad \text{for } i = 1, \ldots, n
\end{aligned} \tag{B.32}$$

基於上述的限制式，僅有少數的解不為零$(\alpha_i^* - \alpha_i)$，而這些樣本也就是「支持向量」。而同樣地我們可對特徵以核技巧轉換（公式(B.23)、公式(B.24)與公式(B.25)）投影至高維度特徵，配適非線性的數據。

B.5　支持向量數據描述

「**支持向量數據描述**」（support vector data description, SVDD）是基於「支持向量機」發展，適用於屬於「**一元分類**」（one-class classification）的問題（Sun and Tsung, 2003）。此模型是透過訓練樣本尋找一個由支持向量所構成的最小體積「**超球體**」（hypersphere），用以描述這群樣本的狀態，以樣本若落在球內為同一類，在球以外則是其他類別，如圖 B.10 所示。實務上常用於「離群值偵測」（outlier detection）與「異常診斷」

（anomaly detection），舉例來說，若是在正常狀態蒐集的數據可描述成一個超球體，若落於球體之外的則為異常狀態。此外，「支持向量數據描述」也能應用在無母數管制圖的建構（請參閱章節「統計製程管制與先進製程控制」）。

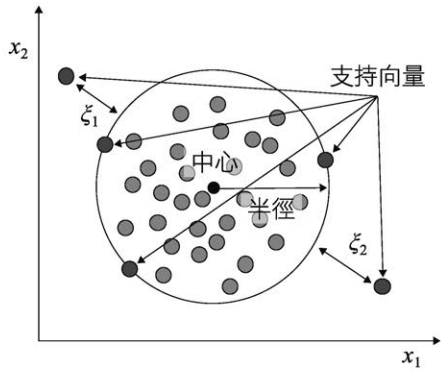

圖 B.10　超球體於一元分類問題

B.5.1　模型建構

由於「支持向量數據描述」是一元分類問題，因此需將所有訓練樣本視為同一類，並視為目標樣本把它們包絡成一個球體，期望該球體越小越好。「支持向量數據描述」與「支持向量機」思維十分相似，差異在於我們將邊際$\|\beta\|$替換成球體大小R^2，其最佳化數學式如公式(B.33)，

$$\begin{aligned} &\min_R R^2 + C\sum_{i=1}^{n}\xi_i \# \\ &\text{subject to } \|x_i - a\|^2 \le R^2 + \xi_i, \xi_i \ge 0 \quad \text{for } i = 1, \ldots, n \end{aligned} \tag{B.33}$$

其中x_i為訓練樣本，a為超球體的中心點，R為超球體的半徑，同樣地ξ_i為誤差而C為懲罰權重。當樣本x_i滿足上式限制 $\|x_i - a\|^2 < R^2 + \xi_i$時，其幾何意義表示該樣本落於超球體內；而當樣本$x_i$滿足上式限制 $\|x_i - a\|^2 = R^2 + \xi_i$時，則其幾何意義表示該樣本為建構超球體的支持向量，並且由於寬鬆變數ξ使得此模型能容忍部分較大誤差的離群值，如圖 B.9 所示。

李家岩、洪佑鑫（2022），製造數據科學：邁向智慧製造與數位決策，前程文化。

B.5.2 模型最佳化求解與距離計算

接著，如同「支持向量機」的推導方式可將上式改寫成「拉格朗日函數」的形式，如公式(B.34)，

$$L(R,a,\xi,\alpha,\mu)=R^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[R^2+\xi_i-(x_i-a)^2]-\sum_{i=1}^{n}\mu_i\xi_i \tag{B.34}$$

其中我們期望最小化的參數是其中的半徑R、中心點a與誤差ξ_i。若對上式的各個參數偏微分並令它們導數為零可得限制式如公式(B.35)、公式(B.36)與公式(B.37)所示，

$$\frac{\partial L}{\partial R}=\sum_{i=1}^{n}\alpha_i\overset{\text{set}}{=}0\Longrightarrow\sum_{i=1}^{n}\alpha_i=0 \tag{B.35}$$

$$\frac{\partial L}{\partial a}=\sum_{i=1}^{n}\alpha_i x_i-a\sum_{i=1}^{n}\alpha_i\overset{\text{set}}{=}0\Longrightarrow a=\sum_{i=1}^{n}\alpha_i x_i \tag{B.36}$$

$$\frac{\partial L}{\partial \xi_i}=\alpha_i-(C-\mu_i)\overset{\text{set}}{=}0\Longrightarrow\alpha_i=C-\mu_i\quad\text{for } i=1,\dots,n \tag{B.37}$$

以及參數的非負限制式$\alpha_i,\mu_i,\xi_i\geq 0$。因此，將上述公式代回公式(B.34)的目標式，可得到「支持向量數據描述」的對偶問題如公式(B.38)所示。

$$\max_{\alpha_i}L_D=\sum_{i=1}^{n}\alpha_i x_i^T x_{i'}-\sum_{i=1}^{n}\sum_{i'=1}^{n}\alpha_i\alpha_{i'}x_i^T x_{i'}$$
$$\text{subject to }\sum_{i=1}^{n}\alpha_i=0\text{ and }0\leq\alpha_i\leq C \tag{B.38}$$

此外，除了上述限制式外，KKT 條件式（Karush-Kuhn-Tucker condition）所包含的限制式如公式(B.39)、公式(B.40)與公式(B.41)所示。

$$\alpha_i[R^2+\xi_i-(x_i-a)^2]=0\text{ (complementary slackness)} \tag{B.39}$$

$$\mu_i\xi_i=0 \tag{B.40}$$

$$R^2+\xi_i-(x_i-a)^2\geq 0\quad\text{for } i=1,\dots,n \tag{B.41}$$

實際上，超球體的半徑R是以球體外圍支持向量到中心點距離的計算所得的（支持向量為樣本x_i的拉格朗日乘數α_i大於 0 且小於 C 時），如公式(B.42)所示。

$$R = \sqrt{(x_k - a)^2} = \sqrt{(x_k \cdot x_k) - 2\sum_{i=1}^{n} \alpha_i (x_k \cdot x_i) + \sum_{i=1}^{n}\sum_{i'=1}^{n} \alpha_i \alpha_{i'} x_i^T x_{i'}}, x_k \in \text{支持向量} \quad \text{(B.42)}$$

若是要將測試資料x_{test}代入「支持向量數據描述」計算距離時，則是將上式的支持向量替換為該測試樣本，如公式(B.43)所示。

$$D = \sqrt{(x_{\text{test}} \cdot x_{\text{test}}) - 2\sum_{i=1}^{n} \alpha_i (x_{\text{test}} \cdot x_i) + \sum_{i=1}^{n}\sum_{i'=1}^{n} \alpha_i \alpha_{i'} x_i^T x_{i'}} \quad \text{(B.43)}$$

若距離D小於等於半徑R則屬同類，反之則被視為不同類（離群值）。

B.5.3 非線性拓展

同樣地，「支持向量數據描述」可使用「支持向量機」的核技巧將特徵投影至高維度空間，拓展至非線性的模型，因此，我們可將公式(B.38)以核技巧拓展如公式(B.44)所示。

$$\max_{\alpha_i} L = \sum_{i=1}^{n} \alpha_i K\left(x_i^T x_{i'}\right) - \sum_{i=1}^{n}\sum_{i'=1}^{n} \alpha_i \alpha_{i'} K\left(x_i^T x_{i'}\right)$$
$$\text{subject to } \sum_{i=1}^{n} \alpha_i = 0 \text{ and } 0 \le \alpha_i \le C \text{ for } i = 1, \ldots, n \quad \text{(B.44)}$$

若將核函數導入公式(B.42)與公式(B.43)，則半徑R與測試樣本距離D的計算如公式(B.45)與公式(B.46)所示。

$$R = \sqrt{K(x_k \cdot x_k) - 2\sum_{i=1}^{n} \alpha_i K(x_k \cdot x_i) + \sum_{i=1}^{n}\sum_{i'=1}^{n} \alpha_i \alpha_{i'} K\left(x_i^T x_{i'}\right)}, \quad \text{(B.45)}$$
$$x_k \in \text{支持向量}$$

李家岩、洪佑鑫（2022），製造數據科學：邁向智慧製造與數位決策，前程文化。

$$D = \sqrt{K(x_{\text{test}} \cdot x_{\text{test}}) - 2\sum_{i=1}^{n} \alpha_i K(x_{\text{test}} \cdot x_i) + \sum_{i=1}^{n}\sum_{i'=1}^{n} \alpha_i \alpha_{i'} K\left(x_i^T x_{i'}\right)} \quad \text{(B.46)}$$

B.6 結語

本章介紹支持向量機，其透過最佳化方法建構最大化邊際、容忍誤差、並以核技巧解決線性不可分割的問題。支持向量機的分類效果主要取決於「**核函數選擇**」、「**核函數參數**」與「**軟邊際常數C**」，參數的最佳組合可以透過**網格搜尋法**（grid search）、**交叉驗證**、**貝氏最佳化**（Bayesian Optimization）等方法協助，以達到較佳「**偏誤與變異權衡**」。然而，支持向量機求解出的模型參數解釋不易，且目標變數需要標籤完全，因此實務上使用仍有其限制。此外，支持向量機是一個二次規劃（quadratic programming）的問題，有許多研究提出相關的演算法對問題進行分解與求解，這也說明了凸性最佳化（convex optimization）在機器學習 / 數據科學中扮演重要的學理基礎（Boyd and Vandenberghe, 2004）。

參考文獻

[1] Boyd, S., and Vandenberghe, L. (2004). *Convex Optimization*. 1st edition, Cambridge University Press.

[2] Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1-27.

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

[4] Sun, R., and Tsung, F. (2003). A kernel-distance-based multivariate control chart using support vector methods. *International Journal of Production Research*, 41(13), 2975-2989.

問題與討論

1. (a)試說明最大邊際分類器的概念，以及其最佳化問題的數學模型。(b)試說明支持向量分類器的概念，以及其最佳化問題的數學模型。（提示：可以兩特徵的二元分類問題輔助說明最大邊際與誤差容忍）
2. (a)試說明支持向量機的核技巧是如何將線性模型轉換為非線性模型；(b)如何避免支持向量機的過度配適？（提示：從數學式說明）
3. 試說明支持向量迴歸與線性迴歸的差異為何？試至少列出三點差異，並說明之。
4. 在 UCI Machine Learning Repository 開放數據中包含了一個鋼板缺陷數據（steel plates faults dataset，https://archive.ics.uci.edu/ml/datasets/steel+plates+faults），一共包含了 1,941 個觀測值，而每個觀測值具有 27 個特徵以及作為目標值的 7 種缺陷。試挑選出凹凸不平（Bumps）以及刮痕（K_Scratch）兩種缺陷進行分析，並撰寫程式從網路上找相關的套件（package），試著使用此數據回答下列問題：
   (a) 試以支持向量機建構模型進行分類。
   (b) 承接(a)的結論，與羅吉斯迴歸以及線性判別分析模型進行比較。（提示：試討論模型預測效果、解釋性、計算複雜度等）
5. 在 UCI Machine Learning Repository 開放數據中包含了一個鋼板缺陷數據（steel plates faults dataset，https://archive.ics.uci.edu/ml/datasets/steel+plates+faults），一共包含了 1,941 個觀測值，而每個觀測值具有 27 個特徵以及作為目標值的 7 種缺陷。試挑選出凹凸不平（Bumps）以及刮痕（K_Scratch）兩種缺陷進行分析。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) 試以凹凸不平類別的數據為基底，建構支持向量數據描述。接著以 K 管制圖，用以判別剩下的數據（i.e.刮痕）是否落於管制圖（半徑）外。
   (b) 試以刮痕類別的數據為基底，建構支持向量數據描述。接著以 K 管制圖，用以判別剩下的數據（i.e.凹凸不平）是否落於管制圖（半徑）外。
   (c) 承接(a)與(b)的結果，從比較中您得出什麼啟發與結論？
   (d) 承接(a)的結論，與羅吉斯迴歸以及線性判別分析模型進行比較。（提示：試討論模型預測效果、解釋性、計算複雜度等）