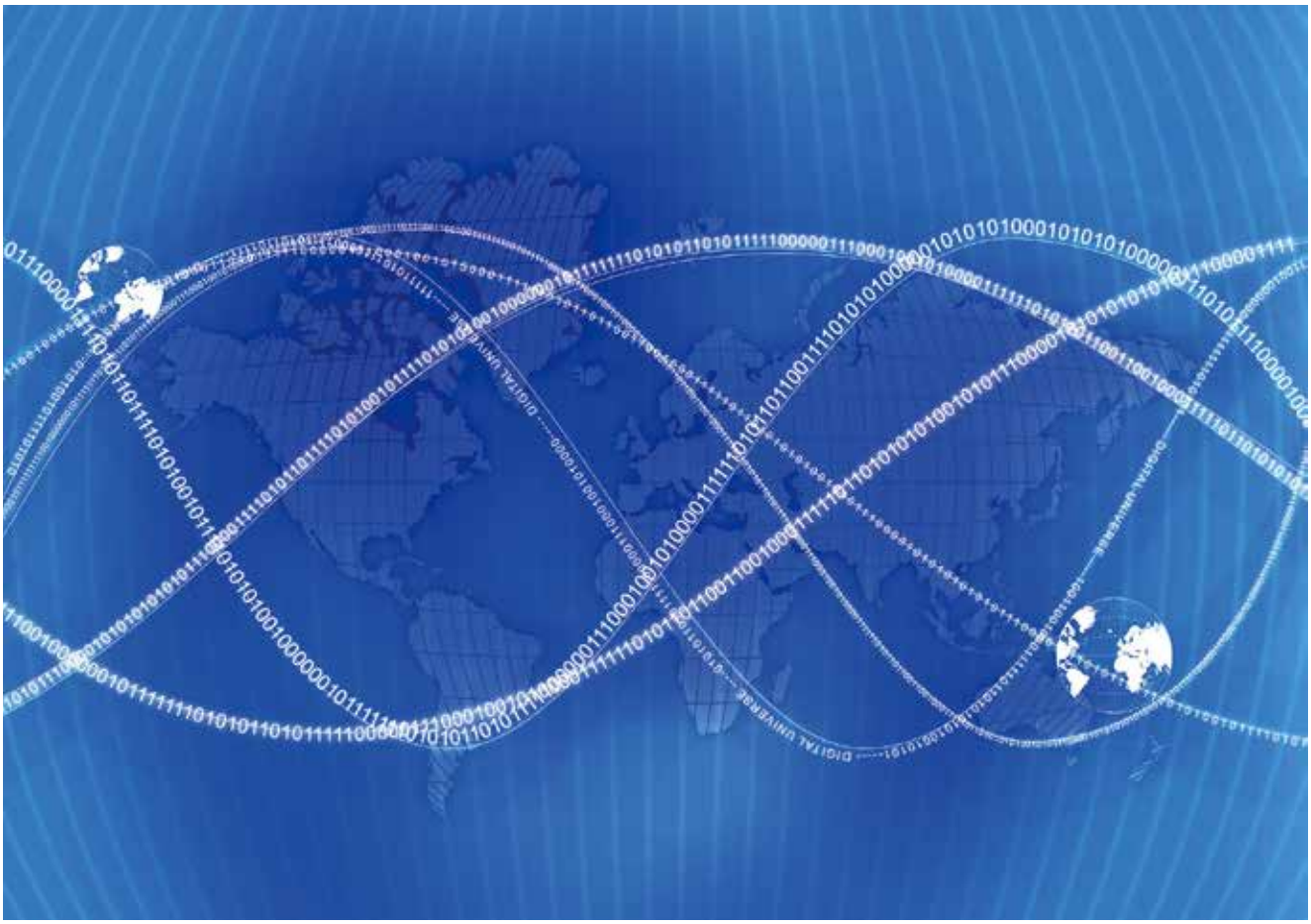
數

機器學習 數學理論與科學應用

●作者：鄂維南（Weinan E） ●譯者：葉千雅

作者簡介：鄂維南是普林斯頓大學數學系教授，應用數學與計算數學博士學程主任，同時也是運籌學與金融工程學系的合聘教授。以他在應用數學和科學計算的相關領域方面的工作而聞名，特別是在非線性隨機偏微分方程、計算流體動力學、計算化學和機器學習等方面。



（Pixabay，Galik Barseghyan 設計）

本文將主要著墨於當前結合機器學習與科學建模（例如：基於物理的建模），以解決各種學科中最具挑戰性問題的工作；以及現今機器學習數學理論建立的工作。

我想傳達兩個基本信息：首先，在理論和計算科學與工程領域中，我們過去遇到的基本障礙是處理在高維度問題上的能力有限；而現在，機器學習提供了處理此問題的新工具。機器學習和科學建模的結合，將提供空前的技術力量，並且有可能改變未來科學研究進行和工程課題探究的方式。

其次，儘管截至目前爲止，機器學習還不是應用數學中最受歡迎的領域；不過，機器學習的精神是與數值分析非常契合，尤其是機器學習能夠處理維度非常高的問題。爲了建立機器學習的理論基礎，我們需要發展高維度的數值分析。

偏微分方程與物理基本定律

數學最重要的角色之一是以它來作爲闡述物理基本定律的語言，而這些定律通常是用偏微分方程（partial differential equation，簡稱 PDE）呈現。當中最重要的兩個 PDE 是在量子力學上的薛丁格方程（Schrödinger equation）與在流體力學中的納維爾／史托克斯方程（Navier-Stokes equation）。其他的例子還有：動力學理論的波茲曼方程（Boltzmann equation）、線性和非線性的彈性方程，以及電磁學中的馬克士威方程（Maxwell equation）。

早在 1929 年量子力學剛剛建立時，狄拉克（Paul Dirac）就做出了以下的觀察 [7]：

> 在大部分的物理學和整個化學中，其數學理論所需的基本物理定律已完全清楚，而困難之處僅在於這些定律的精確應用導致方程過於複雜而無法求解。

基本上狄拉克所說的，是在實際上遇到的大多數情況，難的不再僅僅只是找出基本定律的物理問題，而是在於解出詮釋這些定律的 PDE 的數學問題。

過去的成功與失敗

在這些物理基本定律建立的隨後幾年中，數學研究工作著重於使用分析的手法來建立近似模型或找出近似解。例如：費米（Enrico Fermi）和托馬斯（Llewellyn Thomas）[13] 在同一年提出了密度泛函理論（density functional theory，簡稱 DFT），它是個簡化許多的量子力學模型。多年後，該理論發展爲計算材料科學和化學的主力軍 [13]，其中的主要貢獻者之一孔恩（Walter Kohn）於 1998 年獲頒諾貝爾化學獎。雖然建立的近似模型在實際上被大量使用，然而卻有不少的近似模型是有其特殊性而且通常隱含不受控制的可能。在應用數學圈中，找尋近似解所發展的漸近方法成敗參半。

數值方法求解微分方程

第一個求解這些 PDE 的一般方法是出現在現代計算機誕生之後的 1950 年代。在此期間以及之後的時間裡，人們開發數值演算法、分析它、以及把它應用在各式各樣的問題上，特別是有限差分法（finite difference method）和有限元素法（finite element method）。這是一項非常成功的發展進程，甚至一直持續到今天。這些數值方法相當強大，已成爲工程和許多科學領域的標準工具。包括在氣體動力學、結構分析、雷達、聲納等等相當多的問題，已經從本質上獲得解決了。

然而，仍是有相當多困難問題未能有效解決，比如：在古典多體問題或量子多體問題、基於第一原理 ❶ 的藥物和材料設計、蛋白質折疊問題、亂流（turbulence）、塑性力學與非牛頓流體（non-Newtonian fluid）等等。這些問題的共同特徵是課題本身取決於相當多的變數，例如：儘管納維爾／史托克斯方程式描述的亂流只是一個在 3 維度空間

❶ [編註] 指的是從基本的物理學定律出發，不外加假設與經驗擬合的推導與計算。

的問題，但其解卻包含許多活動尺度，也就是涉及許多的自由度。這使得求解該問題變成是一個維數詛咒（curse of dimensionality）。隨著維度（即變數或自由度的數目）增長，其複雜度（或計算量）成指數增長。這已成爲橫亙於廣泛應用前的基本阻礙。

建立多尺度模型

第二個重要的進展是多尺度（multiscale）和多物理量（multiphysics）演算法的開發。此時我們關注的是某些大尺度（macro-scale）系統的建模，但是在該尺度上卻缺乏可靠的物理模型。反之，我們有可靠的小尺度（micro-scale）系統模型，以一個涉及更多自由度的比較精緻尺度。於是，產生了一種想法：「借由小尺度模型來發展那種能夠在大尺度系統建模時的演算法。」以這種想法發展出的顯著例子有 —— 異質多尺度法（heterogeneous multiscale method，簡稱 HMM）、無方程法（equation-free approach）和準連續法（quasi-continuum method）[8]。所謂 HMM 是從既定形式的大尺度模型開始，進行計算時使用小尺度模型「即時」（on the fly）估算所涉及的未知數。當所處理的問題在大尺度和小尺度間存有尺度分離之特質時，異質多尺度法非常成功；然而，對於缺乏尺度分離的問題上，就顯得相當局限。

在發展 HMM 的早期階段，就已經掌握到有些問題在使用小尺度模型數據估算大尺度模型中的未知量是有困難性。而當該問題若是不具尺度分離特性，就更難以運用異質多尺度法。假設要運用異質多尺度法進行大尺度動力學的亂流建模，並且模擬在大渦流（large eddy simulation）的情形下。此時的未知量是倫納德應力（Leonard stress），即未解自由度對平均應力的貢獻，這必須由原有的納維爾／史托克斯方程式進行估算。難就難在倫納德應力取決於許多的自由度，因此，以納維爾／史托克斯方程式估算這種應力成爲一項非常困難的任務。

結合機器學習與基於第一原理的建模

所有這些問題的難處主要是源自於維數詛咒，也就是我們處理多變數函數的能力有限。近年來，我們已經看到機器學習（尤其是深度學習）在解決電腦視覺上問題或是人工智能的課題上的成功運用案例。很自然的會想問，機器學習能否幫助我們克服先前提到的各種應用中所遇到的障礙，或是更普遍的運用到傳統人工智能之外的領域呢。接下來，我們將描述機器學習在這方向上取得初步成功的一些案例。

在進一步討論之前，我們要先指出我們所關注的機器學習的各種應用課題與傳統人工智能之間的差異。第一個差異是在這裡我們想要運用機器學習的幫助提出可靠和實用的物理模型，而這些模型不應違反諸如對稱性和不變量之類的物理限制。第二個差異是我們經常用來訓練機器學習模型的數據是由某些（小尺度）物理模型所生成。原則上我們可以生成無窮多的數據，但是實際上生成數據所付出的代價相當大。因此，需要開發一種演算法來產生「最佳數據集」（optimal set of data）：涵蓋所有我們關注的實際情況，而除此以外的應儘量少。這

樣的演算法屬於主動學習（active learning）的框架，也可以視爲一個自適應（adaptive）完成數據採樣的自適應演算法。在開發基於機器學習的物理模型上，我們必須解決這兩個一般性的問題。

分子動力學

分子動力學（molecular dynamics）的目的是通過追踪系統中所有原子核的軌跡來進行分子或材料系統的建模。爲此，使用古典牛頓動力學（Newtonian dynamics）是合理的近似法，由所有原子核的位置與電子結構來定義系統中的原子間勢（interatomic potential）。建模這個多變數的原子間勢函數（也稱爲勢能面〔potential energy surface，簡稱 PES〕）是分子建模的主要難題。以往會用兩種方法：第一種是使用量子力學模型（通常是 DFT）進行即時計算。這類方式也被稱爲全始算分子動力學（ab initio molecular dynamics，簡稱 AIMD）[8]。第二種方法是猜測函數的形式，再搭配實驗和既有的數值數據，憑著經驗法則構建出函數。第一種方法相當準確，但也非常昂貴，能夠建模的僅限於數百個原子大小的系統；而第二種方法效率高，但是準確性不可靠。

借助機器學習，人們可以預見一種新的分子動力學典範 —— 機器學習使用由量子力學模型作爲生成器產生的數據來參數化 PES。這種新方法有潛力提供一種精確度媲美 AIDM，而複雜度比擬經驗法則的方式來呈現分子動力學。

關於前文提出的兩個一般性問題：前者中所謂重要的對稱性是指平移、旋轉和置換；後者指的是如果我們把同一種原子重新標記也不會改變系統的事實，因此 PES 也應保持不變。

爲了確保在對稱性與可伸縮性下的不變性，文獻 [5,17] 提出以下針對神經網路結構的設計原則：

1. 整個網路是由子網路的疊加而成，每個子網路對應於系統中的一個原子核。如此一來可確保伸縮性。但是，這種結構不適合用來表示遠距離的交互作用。
2. 每個子網路由一個編碼網路（encoding network）伴隨著一個擬合網路（fitting network）組成。編碼網路確保進入擬合網路的數據要滿足對稱性限制。

文獻 [18] 開發的主動學習演算法由三個主要部分組成：

1. 探索：
   - 大尺度空間：收集熱力學空間變數，例如：溫度和壓力。
   - 小尺度空間：對每組固定的熱力學變數，收集所相應的正準系集（canonical ensemble）。

   兩者都可以通過標準方法來完成。
2. 標記：
   - 用誤差估計式（error estimator）來決定是否標記某個特定的原子組態。可依如下得到一個簡單的誤差估計式：訓練一組神經網路模型（例如具有相同網路結構，但是由不同的初始化訓練），以其預測變異數（variance of prediction）作爲誤差估計式。誤差估計式的數值大表示目前的網路模型對考慮的組態而言不

夠準確。因此，需要標記該組態並放到重新訓練的數據集。

- 對於需要標記的組態，使用 DFT 計算出其原子核上的勢能和作用力，然後將結果放入數據集。

3. 訓練：借由擬合數據集中的勢能和作用力來更新網路參數。

只要特定一個有興趣的系統，我們可以從完全沒有數據開始，演算法就會持續的改進模型，直到獲得滿意的精準度爲止。文獻 [18] 中的例子發現，探索的原子組態中僅有約 0.01% 需要標記。

到目前爲止，這套方法已經應用於廣泛的各種系統，包括大小分子、水分子、半導體、表面材料、高熵合金等等。在每種情況下，人們都可以得到精確度媲美 DFT 的機器學習模型。一款名爲 DeePMD-kit 的通用軟體也已經開發出來，並且在世界各地有許多團隊都採用它 [16]。

氣體動力學建模

氣體動力系統是由描述單一氣體粒子相空間分佈函數之演化的波茲曼方程所建模的。對稠密氣體，則由描繪粒子在空間中密度、動量和能量分佈之演化的歐拉方程（Euler equation）來精確近似波茲曼方程式。控制該近似的關鍵量是努森數（Knudsen number），即氣體粒子平均自由徑（mean-free-path）與系統典型長度尺度的比值。當努森數小，歐拉方程是一個相當精確的近似。在這種情況下，單一粒子分佈函數與所謂的局部馬克士威分佈（local Maxwellian）相似，而且歐拉方程可由波茲曼方程的0階、1階與2階矩（moment）的跡（trace）投影導出 [2]。

對於更大的努森數，通過在投影方案中加入更高階的矩，人們大量致力於擴展歐拉或類歐拉（Euler-like）方程的有效性。這項工作遇到了兩個主要困難 [6]：

1. 導出的矩方程不是適定（well-posed）的。例如眾所周知的格瑞德 13 矩方程組（Grad's 13 moment equations）在某些特定的狀態空間區域上不是雙曲型。
2. 閉合問題（closure problem）：要導出閉合方程（closed equation），必須在投影系統中近似更高階的矩。對較大的努森數，我們不能再用局部的馬克士威分佈作爲假設來閉合系統。

機器學習爲建構在大範圍的努森數上有一致精準的（廣義）矩模型帶來了一絲希望，如文獻 [12] 所示，以兩個步驟完成：

1. 學習一組最能代表分佈函數的廣義矩。一個方法是透過自動編碼器最小化分佈函數的重建誤差。
2. 學習這組廣義矩的動力學。動力方程上所出現的項可由分佈函數表示，而這些項可以透過監督式學習逼近。

主要問題還是在如何保持物理上的對稱性以及如何從小尺度模型（這裡是用波茲曼方程式）獲得數據。與前例（分子動力學）相比，我們導出一個新的動力對稱性 —— 伽利略不變性（Galilean invariance）。事實上，對機器學習而言，前例是學習函數時很常見的例子；然而，目前的例子是要學習一個新的動力系統。初步測試顯示了，這種方

法前景可期 [12]。這也是對缺乏尺度分離問題的多尺度建模（multiscale model）的範例。

高維度偏微分方程式

解高維度的 PDE 是受到維數詛咒的典型例子。但有個例外是解線性拋物型 PDE；這時我們可以利用費曼／卡茨公式（Feynman-Kac formula）[14] 將解以布朗路徑（Brownian path）的泛函期望值表示，然後運用蒙地卡羅法（Monte Carlo method）求解。而求解非線性拋物型 PDE 時，一個類似費曼／卡茨的公式以向後隨機微分方程（backward stochastic differential equations，簡稱 BSDE）表示 [14]。這讓我們能制定一個解非線性拋物型偏 PDE 求解的演算法，當中以神經網路來近似在時間段離散集上的解梯度。然後再用（實際上是離散型的）BSDE 在指定的時空位置下求解，並且將近似解與給定的（起）始端條件之差異當作損失函數（loss function）拿來訓練網路參數 [9]。事實證明這個演算法非常成功，而且可以用來解相當高維度的非線性拋物型 PDE 與 BSDE。作爲副產品的 BSDE 是種可用於財務、經濟、隨機控制、與其他應用的數學利器，現已成爲一功能強大的實用工具。

這些方法的應用也出現在一個用於有違約風險的期權定價模型上的非線性布萊克／休斯方程（Black-Scholes equation）。

自然語言的數學原理

機器學習的成功引發了語言學（lingustics）和自然語言處理（natural language processing，簡稱 NLP）之間的衝突。一方面，NLP 在機器翻譯和其他實用工作中成就了令人矚目的成功，這一事實引起了人們對語言學的實用價值產生質疑。另一方面，因爲機器學習要求遠超乎人類所需的大量訓練數據，所以人們似乎也對機器學習方法的效率提出了質疑。另外，NLP 演算法的黑盒子特性也是一個問題。因此，人們需要開發自然語言的定量結構模型，以彌合語言學與 NLP 之間的鴻溝。

除此之外，自然語言的數學模型也有其自身的價值。例如：同時對於 NLP 與語言學而言的一個重要問題是語義的定義。爲了解決這種問題，最終必須建立數學模型。

語言表現有許多尺度，例如：單詞、句子、與段落。不同尺度需要不同的結構模型。在單詞與句子之類較小的尺度上，不同的語言表現出各種的多樣性和不規則性。但是，在較長的尺度上，不同的語言表現出了顯著的普遍性。這種普遍性是語言之間可以彼此相互翻譯的基礎 [11]。

「語義」的優雅數學定義是在於翻譯後的不變性。如果我們將翻譯視爲不同語言之間的運算子（operator），那麼翻譯後要保留語義意味著不同語言的生成器彼此之間是相似的：

$$\mathbf{P}_{\mathrm{A}}\mathbf{T}_{\mathrm{A}\to\mathrm{B}} = \mathbf{T}_{\mathrm{A}\to\mathrm{B}}\mathbf{P}_{\mathrm{B}}。$$

這裡 $\mathbf{P}_{\mathrm{A}}$ 和 $\mathbf{P}_{\mathrm{B}}$ 分別代表 A 語言和 B 語言的生成器，$\mathbf{T}_{\mathrm{A}\to\mathrm{B}}$ 是代表將 A 語言翻譯成 B 語言的翻譯運算子。因此，$\mathbf{P}_{\mathrm{A}}$ 和 $\mathbf{P}_{\mathrm{B}}$ 的值譜（spectrum）必須相同。

當中的挑戰是在於將這些直觀的優雅論述轉換成數學模型。文獻 [11] 中已有了初步的嘗試：除了確保上述的不變性外，還建立了一個似乎能以某種句子間的尺度捕捉動態的定量結構模型。其副產品就是人們可以使用這個結構模型開發具有最新性能、但使用的訓練數據樣本小很多的機器翻譯演算法 [11]。

機器學習的數學理論

我們將專注於探討呈現相當清晰數學圖像的監督學習（supervised learning）問題。但是我們認爲這裡討論的原理應該也適用於無監督學習（unsupervised learning）和強化學習（reinforcement learning）。

簡單來說，監督學習的問題是運用有限樣本的函數值來逼近給定的目標函數。以 $f^*:\mathbb{R}^d\to\mathbb{R}^1$ 表示目標函數，令 $\{\mathbf{x}_j\}_{j=1}^n$ 爲分佈函數 μ 在 $\mathbb{R}^d$ 獨立採樣的數據集，以及令 $y_j=f^*(\mathbf{x}_j)$，$j=1,\dots,n$。由於添加量測雜訊通常不會從本質上改變論述，爲了清楚呈現，我們將忽略它。我們的目標是以 $S=\{(\mathbf{x}_j,y_j)\}_{j=1}^n$ 逼近目標函數 f^*，這可以通過以下步驟來達成：

1. 構造某個「假設空間」（一組函數）$\mathcal{H}_m=\{f(\cdot,\theta)\}$，這裡 m 是空間大概的維度，θ 表示對應空間中特定函數的參數。
2. 最小化假設空間中的「經驗風險」（empirical risk）：

$$\begin{aligned}\hat{\mathcal{R}}_n(\theta) &= \frac{1}{n}\sum_j (f(\mathbf{x}_j,\theta)-y_j)^2\\ &= \frac{1}{n}\sum_j (f(\mathbf{x}_j,\theta)-f^*(\mathbf{x}_j))^2。\end{aligned}$$

常用的假設空間有下列幾種：

1. 線性迴歸（linear regression）：

$$f(\mathbf{x},\theta)=\beta\cdot\mathbf{x}+\beta_0，\theta=(\beta,\beta_0)。$$

2. 廣義線性模型：

$$f(\mathbf{x},\theta)=\sum_{k=1}^m c_k\phi_k(\mathbf{x})，\theta=(c_1,c_2,\dots,c_m)，$$

其中 $\{\phi_k\}$ 是一組線性獨立的函數。

3. 雙層神經網路（two-layer neural network）：

$$f(\mathbf{x},\theta)=\sum_k a_k\sigma(\boldsymbol{b}_k\cdot\mathbf{x}+c_k)，\theta=\{a_k,\boldsymbol{b}_k,c_k\}，$$

其中 σ 是某個非線性純量函數，例如：$\sigma(z)=\max\{z,0\}$。

4. 深層神經網路（deep neural networks，簡稱 DNN）：由上述形式的函數組成。

逼近函數是古典數值分析和逼近理論中研究得相當深入的課題。標準的流程如下：

1. 定義一個「適定的」數學模型。包括設定假設空間和損失函數。例如：在 1 維的 3 次樣條函數（cubic spline）情況下，假設空間是由 $\mathcal{C}^1$ 的分段 3 次多項式組成，定義損失函數爲：

$$I_n(f)=\frac{1}{n}\sum_{j=1}^n (f(x_j)-y_j)^2+\lambda\int|f''(x)|^2dx。$$

2. 確認正確的函數空間。例如：索伯列夫空間（Sobolev space）或是貝索夫空間。這些空間由正逼近定理和逆逼近定逼近定理（分別也稱

爲傑克森型定理〔Jackson type theorem〕和伯恩斯坦型定理〔Bernstein type theorem〕）都成立的函數組成；也就是說，一個函數屬於某個函數空間若且唯若可以由給定的逼近方式以特定的精確度逼近該函數。

3. 導出最佳的誤差估計。誤差估計有兩種：先驗（a priori）估計 —— 誤差範圍取決於目標函數的範數（norm）；後驗（a posteriori）估計 —— 誤差範圍取決於數值逼近的範數。例如：對於分段線性有限元（piecewise linear finite element）函數，常見的先驗和後驗估計形式爲 $\alpha = 1/d$，$s = 2$：

$$\begin{aligned} \|f_m - f^*\|_{H^1} &\leq Cm^{-\alpha}\|f^*\|_{H^s}, \\ \|f_m - f^*\|_{H^1} &\leq Cm^{-\alpha}\|f_m\|_h。 \end{aligned}$$

這裡 $\|\cdot\|_{H^s}$ 是 s 階的索伯列夫範數，而 $\|\cdot\|_h$ 是常用的網格範數 [1]。

傳統方式與機器學習之間有兩個主要區別。首先，我們在機器學習中必須處理非常高的維度。我們已經可以看到，上述的估計受到維數詛咒；而我們有興趣的是沒有該問題的機器學習模型。其次，我們在機器學習中的只有有限的數據。因此我們只能以經驗風險的方式處理；但是我們眞正感興趣是群體風險（population risk）：

$$\begin{aligned} \mathcal{R}(\theta) &= \mathbb{E}(f(\mathbf{x},\theta) - f^*(\mathbf{x}))^2 \\ &= \int_{\mathbb{R}^d} (f(\mathbf{x},\theta) - f^*(\mathbf{x}))^2 d\mu。 \end{aligned}$$

這兩者之間的差異（有時稱爲泛化差距〔generalization gap〕）是我們必須處理的另一個關鍵問題。結果證明維數詛咒的問題也在這裡出現。

爲了理解這些問題，讓我們用一個簡單的廣義線性模型來做示範。這裡的假設空間爲

$$f(\mathbf{x},\theta) = \sum_{k=1}^{m} a_k \phi_k(\mathbf{x})，\theta = (a_1, a_2, \ldots, a_m)，$$

其中 $\{\phi_k\}$ 是一組線性獨立的函數。讓我們來看看 $m > n$ 的情況：這時爲了擬合數據樣本，也就是將經驗風險降低到其全體最小值 0，我們只需選擇滿足 $G\theta^\top = \mathbf{y}$ 的參數 θ，其中 $G = (\phi_k(\mathbf{x}_j))$，$\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$。這是一個未知數比方程式多的線性方程組。除非 G 是退化的，否則該方程組有無窮多組解。假如選擇 θ 爲具有最小歐氏範數的解，則可以證明 [4]

$$\sup_{\|f\|_{\mathcal{B}_1} \leq 1} \inf_{h \in \mathcal{H}_m} \|f - h\|_{L^2(D_0)} \geq \frac{C}{dm^{1/d}}，$$

看出維度詛咒。這裡的 $\mathcal{B}_1$ 是巴隆空間（Barron space）[10]。在這情況下，即使經驗風險爲 0，仍然可以有相當大的群體風險。

在高維度的問題上，積分問題是一個重要的基準。假設我們想要估算這個積分

$$I(g) = \int g(\mathbf{x}) d\mu。$$

眾所周知，使用如辛普森法（Simpson's rule）這類的求積法會受到維度詛咒；然而使用蒙地卡羅法就沒有這個問題。令 $\{\mathbf{x}_j, j = 1, \ldots, n\}$ 是一組從 μ 採樣的獨立隨機變數，並且

$$I_n(g) = \frac{1}{n}\sum_{j=1}^{n} g(\mathbf{x}_j)。$$

則我們可以得出這個恆等式

$$\mathbb{E}(I(g) - I_n(g))^2 = \frac{1}{n}\mathrm{var}(g)， \quad (1)$$

其中 $\mathrm{var}(g) = \int_X g^2(\mathbf{x})d\mu - (\int_X g(\mathbf{x})d\mu)^2$ 而 $X = [0,1]^d$。此時 n 的指數與 d 無關。當然，在典型的應用上，例如統計物理， $\mathrm{var}(g)$ 在高維度時可以非常大。因此，在使用蒙地卡羅法這個領域的一個主要重點是變異數縮減（variance reduction）。

接著我們來看泛化差距，

$$\begin{aligned} \mathcal{R}(\hat{\theta}) - \hat{\mathcal{R}}_n(\hat{\theta}) &= I(g) - I_n(g)， \\ g(\mathbf{x}) &= (f(\mathbf{x},\hat{\theta}) - f^*(\mathbf{x}))^2， \end{aligned}$$

其中 $\hat{\theta} = \mathrm{argmin}\,\hat{\mathcal{R}}_n(\theta)$。要注意到，在這種情況下，函數 g 與數據是密切相關的，因此，等式 (1) 就不適用。界定泛化差距的一個方法是用假設空間上 $I(g) - I_n(g)$ 的最小上界（supremum）來估算，這個量的尺度自然取決於假設空間。例如，

- 對於利普希茲函數（Lipschitz function，這與瓦塞斯坦距離〔Wasserstein distance〕有關），我們就有：

$$\sup_{\|h\|_{Lip}\le 1} |I(h) - I_n(h)| \sim \frac{1}{n^{1/d}}；$$

- 對於在文獻 [10] 中定義的巴隆空間中的函數，我們就有：

$$\sup_{\|h\|_{\mathcal{B}_1}\le 1} |I(h) - I_n(h)| \sim \frac{1}{\sqrt{n}}。$$

可以看到在不同的假設空間上，維度詛咒以一種不同的形式表現出來，也就是數據集的大小。

一個估算泛化差距的重要概念是拉德馬克複雜度（Rademacher complexity）[3]。令 $\mathcal{H}$ 爲一組函數、$S = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ 爲一組數據點。則 $\mathcal{H}$ 相對於 S 的拉德馬赫複雜度是定義爲

$$\hat{R}_S(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\xi\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\xi_i h(\mathbf{x}_i)\right]，$$

其中 $\{\xi_i\}_{i=1}^n$ 是獨立同分佈（independent and identically distributed，i.i.d.）的隨機變數，以相等機率取值爲 1 或 -1。

舉例來說，

- 如果 $\mathcal{H}$ 是利普希茲空間中的單位球 (2)，則 $\hat{R}_S(\mathcal{H}) \sim O(1/n^{1/d})$；
- 如果 $\mathcal{H}$ 是 $\mathcal{C}^0$（連續函數）中的的單位球，則 $\hat{R}_S(\mathcal{H}) \sim O(1)$；
- 如果 $\mathcal{H}$ 是巴隆空間中的單位球，則 $\hat{R}_S(\mathcal{H}) \sim O(1/\sqrt{n})$。

最後的例子是我們對假設空間的拉德馬赫複雜度所能期待的最佳尺度。

拉德馬赫複雜度之所以重要在於它能對我們感興趣的最小上界給出上下界 [3]：給定一類函數 $\mathcal{H}$ 以及對任何 $\delta \in (0,1)$，在機率至少有 $1-\delta$ 的隨機樣

(2) [編註] 假設空間中的單位球指的是範數不超過 1 的函數集合。

本 $S=(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ 上，則

$$\begin{aligned}&\frac{1}{2}\hat{R}_S(\mathcal{H})-\sup_{h\in\mathcal{H}}\|h\|_\infty\sqrt{\frac{\log(2/\delta)}{2n}}\\&\le\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{\mathbf{x}}[h(\mathbf{x})]-\frac{1}{n}\sum_{i=1}^{n}h(\mathbf{x}_i)\right|\\&\le 2\hat{R}_S(\mathcal{H})+\sup_{h\in\mathcal{H}}\|h\|_\infty\sqrt{\frac{\log(2/\delta)}{2n}}\text{。}\end{aligned}$$

在這背景之下，我們現在可以將所有機器學習模型分爲兩類：

❶ 受到維度詛咒的模型：

$$\text{泛化誤差}\ge O(m^{-\alpha/d}+n^{-\beta/d})\text{。}$$

維度詛咒可能來自於近似誤差（右式的第一項的）或是泛化差距（右式的第二項）。分段多項式近似或是具有固定基底的小波（wavelet）屬於此類。

❷ 不受維度詛咒影響的模型：

$$\text{泛化誤差}\le O\left(\frac{\gamma_1(f^*)}{m}+\frac{\gamma_2(f^*)}{\sqrt{n}}\right)\text{。}$$

有三種機器學習模型已經被辨識爲這一類：

1. 隨機特徵模型（random feature model）：令 $\{\phi(\cdot,\omega),\omega\in\Omega\}$ 是一組隨機「特徵」，並用 π 表示隨機變數 ω 的機率分佈。給定任何 ω 的獨立同佈布實現 $\{\omega_j\}_{j=1}^m$，

$$\mathcal{H}_m(\{\omega_j\})=\left\{f_m(\mathbf{x},\mathbf{a})=\frac{1}{m}\sum_{j=1}^{m}a_j\phi(\mathbf{x};\omega_j)\right\}\text{。}$$

2. 雙層神經網路：

$$\mathcal{H}_m=\left\{\frac{1}{m}\sum_{j=1}^{m}a_j\sigma(\boldsymbol{b}_j^\top\mathbf{x}+c_j)\right\}\text{。}$$

3. 殘差神經網路（residual neural networks）：

$$\begin{aligned}&\mathcal{H}_m=\{f(\cdot,\theta)=\alpha\cdot\mathbf{z}_{L,L}(\cdot)\}:\\&\mathbf{z}_{l+1,L}(\mathbf{x})=\mathbf{z}_{l,L}(\mathbf{x})+\frac{1}{L}\mathbf{U}_l\sigma\circ(\mathbf{W}_l\mathbf{z}_{l,L}(\mathbf{x}))\text{，}\\&l=0,\ldots,L-1,\quad\mathbf{z}_{0,L}(\mathbf{x})=\mathbf{V}\mathbf{x}\text{。}\end{aligned}$$

這裡的 L 是網路的深度。

可執行以下步驟辨認出這三種模型 [10]：

❶ 透過某個大數法則把目標函數表示成期望值。對隨機特徵模型與雙層神經網路模型很簡單明確；而對殘差神經網路就要用文獻 [10] 建立的「複合大數法則」（compositional law of large numbers）。

❷ 藉由證明對應的中央極限定理（central limit theorem）建立近似誤差的收斂速度。同樣對於前兩種模型，這是標準的作法；但是對於殘差神經網路模型則需要費點工夫。

❸ 估計拉德馬赫複雜度。結果證明對這三種模型，其拉德馬赫複雜度的最佳尺度都可以表示爲數據集大小的函數。

在這過程中，還可以爲對應的機器學習模型辨識正確的函數空間。對應隨機特徵模型的是再生核希爾伯特空間（reproducing kernel Hilbert space），而雙層神經網路與殘差網路模型分別是文獻 [10]

定義的巴隆空間與複合函數空間（compositional function space）。

結果是，對於這三種模型而言，經過適當的正則化後，可以很容易的證明泛化誤差的先驗誤差估計都與維度無關（就像蒙地卡羅法）。

爲了給機器學習建立穩固的數學基礎，還有很多工作要做。但是，很明顯這些問題富有數值分析的精神。新的改變是維度高以及模型從參數比數據還多的意義上過參數化（overparametrized）。

結論

我所觸及的僅是迎面而來的巨大冰山之一角。我們現在正處於一場新科學革命的邊緣，它將不只衝擊科學，同時以根本的方式衝擊數學和應用數學。尤其是，

- 將機器學習（克卜勒典範爲代表）與基於第一原理的物理建模（牛頓典範爲代表）整合，開創科學研究的一個新的強大典範。應用數學正是這個整合的最佳平台。
- 爲了建立機器學習的理論基礎，我們必須發展高維度的數值分析。

如果本文中這些觀點的證據還不夠充分，還會繼續以驚人的速度發展。很難想像應用數學會有比眼前更好的機會。∞

本文出處

本文譯自 “Machine Learning: Mathematical Theory and Scientific Applications”, *Notices of the American Mathematical Society* 66 (2019) No.11, AMS。感謝 AMS 同意轉載翻譯。同文是 2019 年 7 月 15 日在西班牙瓦倫西亞舉行的第九屆國際工業和應用數學大會（ICIAM 2019）中，本文作者於彼得亨利希獎（Peter Henrici Prize）演講的筆錄。

本文參考資料請見（數理人文資料網頁）
https://yaucenter.web.nctu.edu.tw/?lang=tw

譯者簡介

葉千雅是新竹高工數學科教師。

延伸閱讀

▶基於機器學習演算法求解高維度控制問題是由作者首先提出的。參見 Jiequn Han, Weinan E,“Deep Learning Approximation for Stochastic Control Problems”, accepted, *NIPS Workshop on Deep Reinforcement Learning* (2016)。

▶作者也是首位提出基於機器學習演算法求解高維度非線性偏微分方程。參見 Weinan E, Jiequn Han, Arnulf Jentzen,“Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations”, *Communications in Mathematics and Statistics* 5 (2017)。

▶作者並架設及管理以下的網站，網站內收集許多機器學習的數學理論，以及運用機器學習在處理多尺度建模問題上。
https://web.math.princeton.edu/~weinan/

鄂維南攝於 2004 年 10 月 2 日。（維基，Weinan E）