ENTERPRISE BIG DATA FRAMEWORK®

Big Data Professional Guide to Machine Learning: Fundamentals

大數據專業人士機器學習指引：基礎

AUGUST 2023

數據素養系列
白皮書

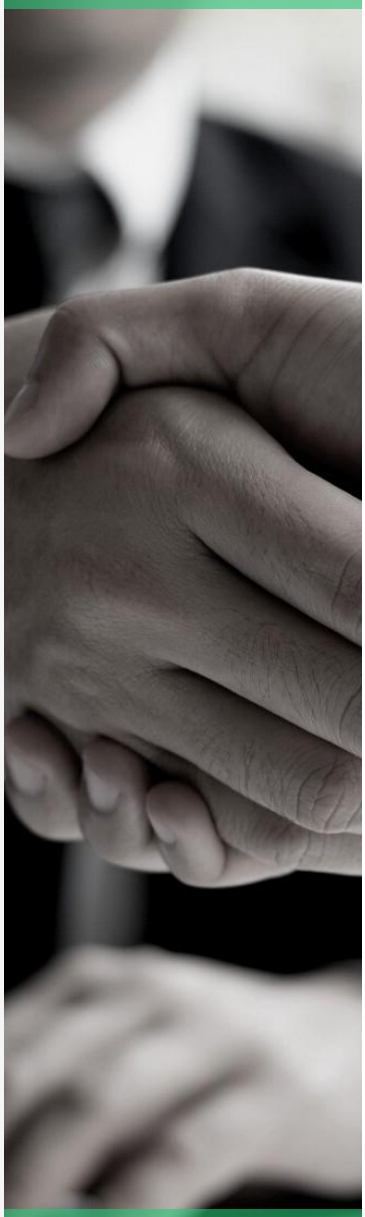
目錄

摘要

大數據和機器學習的融合，引發了革命性的協同效應，有可能重塑產業和重新定義數據分析。對於深耕大數據領域的專業人士來說，採用機器學習不再是一種選擇，而是一種策略需求。本摘要闡述了大數據專業人士急需深入研究錯綜複雜的機器學習，其背後的原因。

大數據以其數量巨大、速度快、多樣性和本質上的不確定性為其特徵，為揭示支持明智決策的見解和模式提供了前所未有的機會。然而，隨著數據繼續呈現指數級的增長，傳統的分析方法無法從資訊的龐大和複雜中收集有價值的見解。這就是機器學習 - 訓練演算法從數據中學習並做出預測的藝術 - 作為一種可以發揮作用，有效的解決方案。

透過連接數據與可操作見解之間的差距，機器學習為大數據專業人士提供了從大量數據集裡獲得出有意義結論的工具。無論是預測消費者行為、優化供應鏈、偵測異常或使用者個人化的體驗，機器學習技術都擅長辨識出人類分析無法察覺的模式。透過不斷重複的學習(iterative learning)和自適應模型的建立(adaptive modeling)，這些技術不僅提高了預測的準確性，而且適應了不斷變化的數據動態。



為什麼大數據專業人士應該了解機器學習？

以數據為中心的產業，快速發展的格局中，大數據和機器學習的融合已經成為推動創新和洞察的重要力量。本白皮書闡述了令人信服的理由，讓大數據專業人士可以踏上獲取機器學習能力的旅程。

大數據以自身龐大的規模和複雜性為特徵，為組織提供了前所未有的機會，來獲得洞察力並做出明智的決策。然而，龐大的數據規模常常讓我們使用傳統方法提取有意義的模式和知識時，帶來挑戰。機器學習是一門使電腦能夠從數據中學習並隨著時間的推移提高其性能的藝術。

這種模式的轉變，使大數據專業人士能夠以前所未有的準確度提取更深入的見解、預測趨勢並優化流程。透過吸收機器學習技術，大數據專業人士可以揭示傳統分析方法可能無法發現的隱藏趨勢和關聯性。預測模型建立、分類、聚類和異常值檢測只是機器學習如何豐富數據分析、增強跨領域決策的幾項案例。這些技術依靠數據而蓬勃發展，透過重複不斷的使理解更完善，從而形成完善的模型和見解。

此外，機器學習使數據驅動的見解民主化，使沒有廣泛統計或程式設計背景的專業人士能夠發揮其潛力。使用者友善的工具和平台，將機器學習功能帶到大數據專業人士的指尖，使他們能夠發現複雜的關係和預測模型，從而在整個組織中培養數據驅動的創新文化。

在競爭格局中，組織希望不斷的獲取策略優勢，機器學習的能力成為關鍵資產。具備這些技能的大數據專業人士可以釋放數據洪流中隱藏的價值，推動創新，並創造與客戶產生共鳴的個性化體驗。

大數據和機器學習的融合，正在改寫數據分析和決策的規則。對於大數據專業人士來說，掌握機器學習不僅是一項補充的技能，而是一種變革催化劑。憑藉挖掘預測洞察和推動數據驅動創新的能力，專業人士可以引導組織走向未來，在未來，數據不僅是一種資源，而且是一種戰略優勢，推動他們在不斷發展的環境中前進。

什麼是機器學習？

機器學習 (ML) 是人工智慧 (AI) 的一個部分，專注於演算法和模型的開發，使電腦能夠根據數據進行學習並做出預測或決策，而無需針對每一個特定的任務進行明確的程式設計。換句話說，機器學習允許電腦透過學習數據中存在的模式和關係來自動提高其表現的結果。

機器學習背後的核心思想是開發能夠識別模式、提取見解並從案例中進行歸納的演算法，以便執行任務或對新的、未見過的數據進行預測。機器學習系統不依賴明確的程式設計，而是使用數據驅動的方法，透過經驗來提高其表現的結果。

機器學習可以大致分為三種主要類型：

1.監督式機器學習 Supervised Machine Learning
2.非監督式機器學習 Unsupervised Machine Learning
3.強化式學習 Reinforcement Learning

監督式機器學習
Supervised Machine Learning

監督式學習是機器學習的一種類型，演算法從標記的訓練數據中學習，以做出預測或決策。在監督式學習中，用於訓練的數據集包括輸入特徵（通常稱為“X”）和相應的所需輸出標籤或目標值（通常稱為“y”）。該演算法的目標是學習從輸入特徵到輸出標籤的對應，使其能夠對新的、沒有見過的數據做出準確的預測。

監督式學習的流程涉及幾個關鍵步驟：

- **數據收集**：收集由輸入特徵及其對應標籤所組成的數據集。數據集通常分為兩部分：用於訓練模型的訓練數據集和用於評估模型效能的測試（或驗證）數據集。
- **特徵工程**：準備和預處理輸入特徵，以確保它們採用合適演算法的格式。這可能涉及標準化、縮放、編碼分類變數以及處理遺失的數據。
- **模型選擇**：選擇適合當前問題的機器學習演算法或模型架構。演算法的選擇取決於數據的性質以及特徵和標籤之間關係的複雜性等因素。
- **訓練**：模型使用標籤的訓練數據進行訓練。在訓練過程中，演算法會調整其內部參數，使其預測輸出與訓練數據中實際標籤之間的差異最小。這一過程涉及優化技術，以找到最佳參數。
- **驗證**：訓練後，使用測試或驗證數據集評估模型的性能。常見的評估指標包括準確度、精確度、召回率、F1-score等，這些取決於問題的性質（分類或回歸）。
- **微調**：根據驗證結果，可以透過調整超參數（訓練前所設定的參數）對模型進行微調，以優化其效能。
- **預測**：模型經過訓練和驗證後，可以透過輸入特徵，並取得預測的輸出標籤或值，來對新的、未見過的數據進行預測。

“機器學習將可使大多數人認為只能由人完成的工作自動化。”
- Andrew Ng (吳恩達)

07

監督式學習的任務，大致可分為兩大類：

1. **分類**：在分類任務中，目標是為輸入的數據指派標籤或類別。例如，將電子郵件分類為垃圾郵件或非垃圾郵件、識別動物影像或根據患者數據來診斷醫療狀況，這些都是分類的工作。

2. **迴歸**：在迴歸任務中，目標是預測連續數值。例如，根據面積和位置等特徵預測房價、估計銷售收入或預測股票價格，都是屬於迴歸的問題。

監督式學習，廣泛應用於影像辨識、自然語言處理、詐欺偵測、推薦系統等各種領域，這些領域需要根據現有的標記數據來進行準確的預測或分類。

“分類是一種讓教機器看到不可見的事物、在混亂中找到秩序，並透過隱藏在數據中的模式來指導決策的藝術。”

非監督式機器學習
Unsupervised Machine Learning

非監督式機器學習是機器學習的一種類型，其中演算法從未標記中的數據來進行學習，以發現數據內的模式、結構或關係。這與監督式學習不同！監督式學習演算法是在具有已知輸出的標記數據上進行訓練，非監督式學習涉及處理缺乏明確目標值或標籤的數據。

非監督式學習的主要目標是：探索數據的內在結構並提取有意義的見解，而無需事先了解結果應該是什麼。非監督式學習對於以下任務特別有幫助：

1. **聚類**：根據相似的數據點之固有的相似性，將它們分組在一起。聚類演算法的目標是在識別數據中的自然分組，即使這些群組沒有明確的意義。
2. **降維**：減少輸入變量（特徵）的數量，同時保留最重要的資訊。這有助於簡化複雜的數據集並避免「維數災難」(Curse of dimensionality)。
3. **異常檢測**：識別數據集裡，是否有明顯偏離正常情況的罕見和異常實例。這對於識別異常值、詐欺偵測或異常事件非常有價值。

案例：信用卡詐欺檢測

在信用卡交易的情境下，企業每天處理大量交易。這些交易大多數是合法的，但是有些交易可能是詐欺行為。異常檢測技術可用於識別表現出異常行為的交易，這可能顯示出潛在的詐欺行為。

非監督式學習中常見的使用技術包括：

- 聚類演算法：k 均值聚類、層次聚類和基於密度的聚類(DBSCAN)等演算法。根據數據點的相似性將其分群。
- 降維技術：主成分分析(PCA)和 t-隨機鄰近嵌入法(t-SNE)，都是用於在保留基本特徵的同時，降低數據維度的方法。
- 異常檢測演算法：孤立森林(Isolation Forest)、one-class SVM（一類支援向量機）和自動編碼器(Autoencoder)，可用於偵測數據中的罕見或異常值。

非監督式學習，通常用於探索型數據分析、數據預處理，以及再進一步分析之前，理解數據的底層結構。

強化式學習
Reinforcement Learning

強化式學習(RL)是機器學習的一種類型，在這種學習方式中，代理人(agent)學會著與環境互動，以便隨著時間的推移，讓獎勵累積到最大化。與監督式學習（模型是根據標記數據進行訓練）和非監督式學習（在數據中發現模式）不相同，強化式學習專注於透過反覆試驗來學習最佳決策的策略。

在強化式學習中，代理人在環境中採取行動，並根據這些行動，獲得獎勵或懲罰形式的反饋。代理人的目標是學習一種策略：一種從狀態（情況）到行動的映對。在與環境互動的過程中，使預期的獎勵，累積至最大化。

強化式學習的關鍵組成有：

- 代理人：與環境互動並採取行動，以實現其目標的學習者或決策者。
- 環境：代理人與之互動的外部系統。環境對代理人行為的反應決定了代理人的獎勵。
- 狀態：代理人在環境中所處的當前情況或內容的表示。
- 行動：代理人所做的，從環境中的一種狀態轉換到另一種狀態的決定。
- 獎勵：環境提供的純量值，用於顯示在某些特定狀態下，採取特定操作的直接收益或成本。
- 策略：代理人用來決定在特定狀態下，採取哪些操作，可以獲得最大化長期獎勵的策略。
- 價值函數：估計代理人在遵循特定策略時，從特定狀態可以獲得預期累積獎勵的函數。

強化式學習用於未明確知道最佳動作序列的場景，並且代理人必須透過互動和經驗來學習。強化式學習應用的一些例子包括：

- 訓練自動駕駛車輛在複雜的交通場景中導航。
- 教導機器人透過與環境互動來執行任務。
- 優化金融交易策略。設計遊戲式的人工智慧代理，透過遊戲方式來學習和改進。
- 調整能源管理系統中的資源分配。

強化式學習演算法包括 Q-learning、Deep Q-Network (DQN)、策略梯度(Policy Gradient methods)以及更多先進的技術，例如近端策略優化(Proximal Policy Optimization)和 Actor-Critic 架構。

機器學習的類型

各種機器學習類型的概述

類型	監督式學習	非監督式學習	強化式學習
訓練數據	標記數據（輸入輸出成對）	未標記數據（僅有輸入）	與環境的互動
目標	預測輸出標籤或值	預測輸出標籤或值	最大化累積獎勵
反饋	預測和標籤之間的錯誤	沒有可比較的標記輸出	環境的獎勵或懲罰
操作實例	分類 迴歸	聚類 降維	玩遊戲 機器人
訓練過程	模型從標記數據中學習	模型辨識數據模式	代理人透過反覆試驗來學習
輸出	預測或值	聚類、降維數據	行動或決定
評估	使用準確度、精密度等。	評估聚類品質	獎勵累積品質政策
監督方式	需要人工標記的數據	無需人工標記數據	最少的人工指導
常見的演算法	決策樹、支援向量機、神經網路	k 均值、主成分分析、t-隨機鄰近嵌入法	Q-learning、策略梯度、DQN
案例	影像分類、迴歸	客戶區隔、異常檢測	Game AI、自治系統

機器學習中的倫理考量

由於機器學習技術可以對個人、社會和更廣泛的世界產生深遠的影響，因此倫理的考量在機器學習中至關重要。倫理的考量在機器學習系統的開發和部署中至關重要的一些關鍵原因包括：

- **偏差與公平**：機器學習模型可能會無意中學習訓練數據中存在的偏差，從而導致不公平或歧視性的結果。倫理考量確保模型的設計公平且公正，避免基於種族、性別、年齡或社會經濟地位等因素的歧視。
- **透明度與問責制**：倫理的實踐提高機器學習模型決策方式的透明度。使用者和利害關係人應該理解模型為何做出特定決策，尤其是當這些決策影響個人生活的時候。
- **隱私和數據保護**：機器學習通常依賴包含個人資訊的大型數據集。倫理的考量要求數據隱私受到尊重，敏感資訊安全處理，符合相關的數據保護法規。
- **安全**：在自動駕駛汽車和機器人等領域，機器學習模型可以影響人身安全。倫理的實踐確保模型設計時採取安全措施，以防止事故發生，並最大限度地降低風險。
- **社會影響**：機器學習可以影響社會動態、經濟和就業市場。倫理的考量有助於預測和解決潛在的負面影響，促進負責任的開發和部署。
- **偏差和錯誤的問責**：當機器學習系統犯錯或產生偏差的結果時，倫理的實踐要求開發人員和組織負責糾正問題並防止未來再度發生。
- **大眾的信任**：遵守倫理的原則有助於建立大眾對機器學習技術的信任。信任對於這些技術在各個領域的廣泛採用和接受上，都是至關重要的。
- **法律法規的合規性**：許多地區都有管理數據使用、隱私和公平的法規和法律。倫理的實踐確保機器學習系統遵守這些法律要求。

機器學習中的倫理考量對於塑造符合人類價值觀、社會福祉和負責任的使用先進能力的技術方向至關重要。隨著機器學習技術越來越融入我們的生活，解決倫理問題，成為創造積極和公平未來的當務之急。

機器學習與倫理道德兩難的案例

考慮一家使用機器學習演算法來自動完成部分招募流程的企業。該演算法透過分析求職者的簡歷、資格和其他數據，來預測他們是否適合這份工作。該組織的首要目標是有效聘用最合格的候選人。

如果只為準確性而訓練演算法，那麼由於訓練數據中的歷史偏差，演算法可能會無意中偏向某些群體，並歧視其他群體。例如，如果訓練數據偏向於某個特定的人群，演算法可能會不成比例地偏向於該群體的候選人，即使他們並不是最合適的人選。

在這種倫理道德困境中，企業面臨著一個艱難的抉擇：是優先考慮準確性，但可能會因為延續偏差，而產生代價；或者是優先考慮公平性和多樣性，但可能會影響演算法的預測性能。要找到解決方案，就必須全面了解各種權衡，並採用符合組織價值觀和社會期望的倫理準則。

這個案例強調了機器學習中，倫理道德兩難的複雜性，組織必須駕馭相互競爭的優先事項，並做出明智的決策，以維護其目標和倫理原則。

推薦資源

13

由於機器學習對數據分析的變革性影響，大數據專業人士應該學習機器學習的知識。隨著數據量呈指數級增長，傳統的方法，將難以提取有意義的見解。

機器學習為專業人士提供了發現隱藏模式、預測趨勢和優化流程的工具。在數據熟練程度至關重要的時代，機器學習熟練程度對於大數據專業人士駕馭複雜環境並釋放數據資源的全部潛力，是至關重要的工作。

"每一位大數據專業人士都應該了解機器學習的基礎知識，以了解預測模型的工作原理。"

推薦的書籍

Middelburg, J.W: *The Enterprise Big Data Framework,* Kogan Page, 2023. https://www.amazon.com/Enterprise-Big-Data- Framework-Capabilities

台灣可以參考中文版的免費核心知識手冊

https://www.learnmydata.com/download.html#EBDP_Book

推薦的課程

Data Literacy Fundamentals, offered by APMG-International. https://apmg-international.com/product/enterprise-big-data- certification

台灣的相關國際認證課程與國際認證考試

大數據國際認證課程 ： EBDP 、 EBDA 、 Data Literacy
人工智慧國際認證課程 : BCS AI Essentials 、BCS AI Foundation

嵩澔國際資產股份有限公司是台灣唯一獲得 APMG 直接許可的培訓機構，直接提供 EBDFA 相關課程的組織。

http://www.learnmydata.com/course.html

關於 EBDFA

關於企業大數據框架聯盟 Enterprise Big Data Framework Alliance

透過企業大數據框架，我們向成千上萬的人和組織提供有關數據、分析、機器學習和人工智慧的知識。

企業大數據框架聯盟透過會員資格建立了一個從業者社區，透過獨立於供應商的認證計畫來推廣知識，並舉辦學習和啟發的社區活動。

免責聲明

EBDFA設計並創造了《大數據專業人士機器學習指引：基礎》，主要是作為專業人士的教育資源。EBDFA 並不聲稱使用本白皮書，將確保獲得成功的結果。

本白皮書不應被視為包含所有適當的資訊、程序和測試，或排除合理的用於獲得相同結果的其他資訊、程序和測試。在確定任何特定資訊、程序或測試的適當性時，專業人士應根據特定系統或資訊科技環境所呈現的特定情況，運用自己的專業來判斷。

ENTERPRISE BIG DATA FRAMEWORK®

Endenicher Allee 12
53115, DE Bonn
Germany

W: www.bigdataframework.org
E: info@bigdataframework.org

提供回饋:
communities.kajabi.com/ebdfacommunity

與我們聯繫:

LinkedIn:
www.linkedin.com/company/big-data-framework

YouTube:
www.youtube.com/@bigdataframework

Twitter (X):
twitter.com/bdframework

關於作者:

Jan-Willem Middelburg 是《企業大數據框架》一書的作者，也是企業大數據框架聯盟的首席審查員。在此工作上，他負責 EBDFA 計劃下所有認證的考試品質。

Jan-Willem Middelburg 是大數據和自動化技術的專家，經常在世界各地的大學和技術會議上發表主題演講。