機器學習起步的10個技巧

文／Clint Boulton 何信達 譯／Iris.Liu

人工智慧和機器學習可以為企業提供改變遊戲規則的解決方案。而資深的 IT 主管在這個領域的應用中，需要知道的是發動和支持成功的機器學習策略的要點。

機器學習（ML）正迅速成為CIO 的前瞻性思維的決定性檢驗項目。在未來的十年內，沒有為產品開發運用機器學習技術的企業，其營運的風險將是趕不上更敏捷的競爭對手。服務於提供科學和醫學資訊的 RELX 集團，身為 Elsevier 技術長的 Dan Olley 表示，近年來他逐步帶領組織接納了ML技術的應用。

Olley日前在美國的 Colorado Springs, Colo 舉辦的 CIO100 Symposium 中，對其同事的聽眾們表示，我由衷地相信我們正處於機器學習的關鍵轉折點，並將在未來十年內改變與數位世界的互動方式。我們將越來越仰賴機器的協助以做出決策。

這是一個合理的假設。隨著運算能力的成長，演算法和訓練模型的日益複雜，以及看似無限的資料來源在在都促成了人工智慧（AI）的重大創新。AI，包括機器可以模仿人類思維行為的任何技術，ML則是其中的次領域。ML 是以統計為基礎的演算法自動化知識工程。Google、Amazon、Baidu以及其他的廠商，正傾注更多的資金在AI和ML技術上。此外，麥肯錫全球研究所（McKinsey Global Institute）認為，藉由這些技術的發展所吸引釋放出來的創業資金是 2016 年投資的三倍之多，約介於260億美元至390億美元之間。

現在是應用AI和ML的時候了！

McKinsey發佈的訊息表示，技術部門（tech sector）之外的其他部門，在採用人工智慧相關的應用部份，大多處於早期的試驗階段，很少有企業會大規模地進行部署。依據McKinsey 的表示，尚未採用 AI 技術或是將其視為營運核心部份的企業，主要的顧慮是因為不確定投資在這個部份的回收率究竟是否值得。但是，服務於 Elsevier 的 Olley，致力於 ML 技術的應用，協助製藥業顧客發現藥品，並向臨床醫生提供相關的醫療資訊。他表示，ML的應用案例遍及了人資管理、業務和行銷、以及顧客支援等領域。如果想要建立競爭優勢，或者至少保持領先，CIO們最好迎向這些新興技術的發展潮流。這是你現在要開始著手的事情！Olley 強調。

1.理解資料科學的定位在哪裡？

你有個運用資訊科學和ML的想法，想在企業中發揮其功效的意識，但是，又應該如何去實作呢？第一，不需要聚焦在資訊科學和ML作業上。事實上，將資訊科學和機器學習嵌進每個部門，包括業務、行銷、人力資源、以及財務部門等，都是合理且行得通的。Olley建議CIO們可以嘗試利用 Elsevier 中可以為他工作的東西，他將資料科學家與軟體工程師、或是腫瘤專家成對搭配在一起。這種利用敏捷小組的方式建構產品的靈感是來自Spotify模型。Olley表示，我們已經將資料科學小組融入我們的產品管理團隊和營運部門中，而且，我們將最適當的搭擋組合起來，其中，會有一個人負責主導。我們讓資訊

科學家儘可能地貼近問題的位置，因為這是一種擴大整個組織的更好的方式。

2.那就開始吧！

你不需要為發展一個資訊科學企業而規劃五點計劃，也不需要建構精鍊的ML產品的框架。Gartner表示，你應該在不同的營運領域中培育小型的實驗，為運用特定的AI技術為目標，而不是以投資回報率(Return On Investment，ROI)為評估準則。如果你的企業還沒有拿定主意，我非常地建議你就從現在開始著手吧！ Olley提出建議。你的競爭對手已經開始了！

3.對待你的資料就像是錢！

將資料視為建構AI / ML的燃料，CIO必須將資料的價值視同錢一般，並透過管理、保護、以及關注等方式對待它。Olley表示，你的財務長不會讓企業的帳戶資料遍佈整個公司。他也不會說「我認為今年的收入已經很多了」。

4.停止尋找紫色的松鼠，也就是不要再找條件好到不可能存在於真實世界的求職者

資訊科學家通常擅長於數學和統計，他們熟悉探勘資料以獲得洞察，卻不一定會撰寫演算法，或是身兼製作產品的軟體工程師。這情況說起來很簡單，因為有許多企業通常都會傾向尋找獨特的獨角獸候選人，這些萬中選一的候選人不是統計高手、就是忍者軟體工程師、或是熟悉某個產業的領域知識，像是健康照護或是金融服務等。Olley表示。我曾經聽某個人描述，我想要找一位擁有數學博士學位的軟體工程師，他最好也是一名臨床醫生，如果還具備腫瘤科專業那就真是太好了、、、Olley挖苦地補充，他知道剛才那段話說的其實是三個人。

5.建立資料科學培訓課程

不是每個實行資料科學的人都是為了成為一名資訊科學家，或者是需要同業中的黑帶高手。Olley表示，你無法找全這些人的，不如致力於培養這樣的人才。Olley 描述，他就有一位同仁負責帶領IT團隊強化資料科學的技能。Elsevier也利用發展培訓課程的方式培養所需的人才。Olley建議CIO們至少應該建立機率和統計兩門強化課程，並隨著期末考的舉行，學員們必須通過考試以證明自己的毅力。Gartner建議你識別AI知識與人才之間的落差，同時，發展培訓課程和招聘計劃以建立組織的能力。

6.認可資料科學和ML平台

企業應該加快朝向AI和ML應用的速度，或者不確定如何解決資料科學的問題，就可以將資訊轉儲到資料科學平台上，像是Kaggle。該平台聚集了資料科學家、統計學家、數學家、軟體程式設計師、以及其他喜歡解決棘手問題的團隊，可以透過聚焦在企業的營運挑戰上，以競賽方式產生最佳的模型。

7.注意「導出的資料」

機器學習起步的10個技巧

如果你要與合作夥伴分享你的演算法，那麼他們就會看到你的資料。他表示，這種分享的方式對於像Elsevier這樣的資訊公司而言並不適合。Elsevier熱衷於保護自已擁有的資料，並將其視為競爭優勢。你的資料代表的是新的貨幣，Olley說。你必須了解戰略所欲持續的目標為何，你想要分享的是什麼，以及對待它視同資金一樣。

8.不要總是試圖一次就解決掉所有的問題

健康照顧機構（health-care organization）可以嘗試建立一套演算法，應用於替代所有的一般醫療醫師（primary care physician）。如此，便能解決若沒有預先安排看診的時間，就無法看病的問題。或者，可以藉由撰寫演算法解決一個問題，至少可以辨別某個病人是否只需要阿斯匹靈（aspirin），而不是需要更慎重的後續治療。Olley表示，解決問題的一小部分，以獲取更多的資訊，並隨著時間推移的累積而建立。

9.不要把資料模型想得過於複雜

獲取正確的訓練集合（training set）比起建立完美的資訊模型更為重要。 Olley 表示，不要讓任何資料鬆散，這可能導致資訊模型的不正確。最大的挑戰是向人們展示可能的最新發展，同時，並找出這些技術的用途，然後，再擴展。

10.教導CEO和董事會這些大頭，也能了解AI 的重要性！

關於資訊科學導航員（data science pilot）的承諾。Gartner認為，身為資訊長，你應該尋求促進AI 和 ML作為影響執行長潛在破壞市場並重塑現有業務模式的策略的手段。應該記住的是，成功的機器學習的作用可能會是你所服務的組織的未來的關鍵角色。 CIO