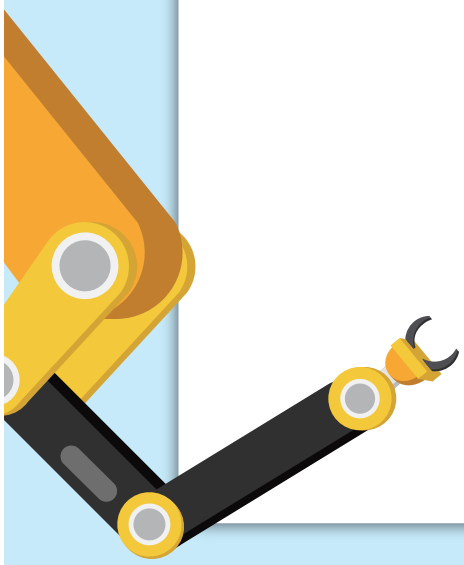
CH 2

我也能打造機器學習



2-1 機器學習的五大流程

在第一章介紹機器學習的分類與方法，以及機器學習各式各樣的應用，你是否也躍躍欲試，想要親手打造一個機器學習的應用呢？不過，機器學習看起來似乎很複雜，而且好像需要相當多的電腦知識與技能，我們該怎麼開始進入這個神秘的殿堂呢？

其實，打造機器學習是有一套準則及方式可循的，在你要打造一個機器學習應用的程式之前，透過下面這些方法，可讓自己更清楚要怎麼進行，才不會迷失方向。

機器學習的五大流程包含：定義問題 → 蒐集資料 → 處理資料集 → 訓練模型 → 推論與預測，如圖 2-1 所示。

◆ 圖 2-1　機器學習的五大流程

2-1-1 定義問題：有好問題才有好答案

進行任何機器學習的設計之前，你應該做的第一件事就是要先想想，你要藉由機器學習來做些什麼？想要用在什麼地方？幫你解決什麼問題？這個問題是否一定要用機器學習來解決才行？

機器學習可以做的事情很多，不過目前機器學習並不是萬能的，所以，什麼問題需要透過機器學習的哪些方法及技術來幫我們解決，是首先必須要思考的。這個過程，我們可以稱之為「定義問題」。

例如，我們想要用 AI 來：

- 辨認香蕉的成熟度
- 預測下一週的空氣品質
- 辨識人臉或物品

- 從歌曲辨識曲風或是歌手
- 辨識語音訊息裡的情緒是開心還是生氣

當我們把問題定義出來之後，接著判斷這個問題是不是適合，或者是否一定要用機器學習。確定這個問題可以透過機器學習的方法來解決之後，我們就可以進入下一個階段「蒐集資料」。

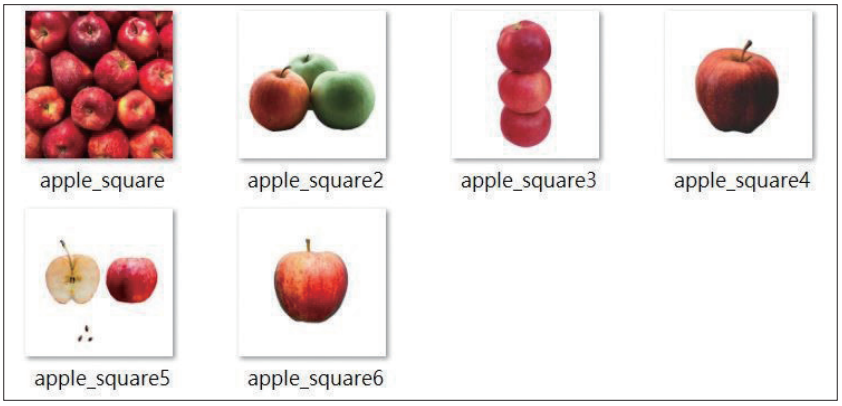
2-1-2 蒐集資料：如何蒐集大量資料讓機器學習

如果說機器學習是一枚火箭的話，資料就是讓這枚火箭推進的燃料。機器學習的一大特點，便是讓電腦看過大量的資料後，從中找出一些規律和模式，依據學習到的規律來對未知問題進行預測或推估。所以經由定義問題這個步驟，找出想要機器學習為你解決的問題後，接著準備相關的資料集，讓機器學習的演算法從這些資料的特徵找出規律與模式。

假設我們現在想要訓練一個可以辨識香蕉、草莓、蘋果、橘子及香瓜等 5 種水果的機器學習模型，那麼，需要準備的資料集大概會像是這個樣子：

1. 蒐集這 5 種水果的圖片，儘可能越多越好，而且要有各品種或是各種角度的影像等。
2. 將這 5 種水果的圖片，分門別類地放在 5 個設好標籤的資料夾裡，例如：Banana、Strawberry、Apple、Orange、Melon。

如圖 2-2，我們準備了幾張不同的蘋果照片，照片中的蘋果有數量不同或是顏色不同等特徵。



◆ 圖 2-2 蘋果的照片

要讓電腦學習蘋果的各種特徵，可能需要上千張照片才足夠，蒐集的資料集越全面，機器學習的辨識率就越高，所以蒐集資料是一點也馬虎不得的功夫。

2-1-3 處理資料集：資料前處理

當我們很辛苦地把機器學習所需的資料都準備好之後，是不是馬上就可以放到電腦裡讓機器學習了呢？先別急，由於電腦只看得懂數字資料，所以通常我們還得要將資料整理一下，讓電腦和機器學習的演算法可以理解與消化得很好，才能發揮功用。這個過程，在機器學習裡通常稱為「**資料前處理（Data Preprocessing）**」。

資料前處理是整個學習機器學習的過程中最花心力的一個地方，它不只是把資料轉換成數字這麼簡單，還必須找出各個變數間的交互關係，例如：哪些變數是對定義的問題影響程度最大、最高的？資料集處理得好不好，會直接影響訓練出來的機器學習模型的準確度。

以前面的水果辨識資料集來看，假設每種水果的照片都有上百張，其中可能有不夠清楚的，也可能有放錯的，或是解析度不夠或太高等問題時，我們就必須依照實際情況，透過各種工具或撰寫程式來處理，把資料去蕪存菁，確認做好資料標記的動作。接下來，機器學習的演算法可以根據這些資料來達成良好的學習效果。

2-1-4 訓練與測試：AI 模型的選擇

接著，是大家最期待的步驟了，我們要用準備好的資料集來進行機器學習的訓練了！不過還有個很頭疼的問題，那就是機器學習的各種演算法實在太多了，每一種演算法擅長處理的問題各有不同，有時還會同時使用多種演算法。

別擔心，還是有些方式及資源讓你有方向與準則來挑選。下面是來自 Scikit-Learn 網站建議的機器學習演算法挑選引導，我們可以依循圖 2-3，概略地協助判斷一開始應該如何挑選到適合的演算法。

非監督式學習：分群	監督式學習：分類	監督式學習：迴歸
K-means	Linear SVM 支援向量機	SGDRegressor
K-modes	Naïve Bayes 樸素貝葉斯	Elastic Net
	Decision Tree 決策樹	Gradient Boosting Tree
	Logistic Regression 羅吉斯迴歸	
	Random Forest 隨機森林	

◆ 圖 2-3 不同種類的機器學習

找到適合的演算法後，將蒐集到並整理完的資料分為兩個部份，一部份是**訓練資料集（Training dataset）**，比例大約會佔 60% ～ 80% 左右，剩下的部份便用於**測試資料集（Test dataset）**，比例佔 20% ～ 40% 左右。

上述情況類似我們在學習過程中都會經歷考試，考試前，老師會給我們一些練習題或考古題，讓我們在考試前練習，但考試當天，你所做的題目通常不會是之前老師給的練習題，應該是全新沒看過的題目。如果你通過考試，就代表你已經從先前的學習中學會了。

進行機器學習的訓練也是同樣的意思，我們利用訓練資料集，讓所選定的機器學習演算法從中學習你想要它學會的事。這個過程結束後，你會得到一個機器學習的**模型（Model）**，這時再使用剩下的測試資料集來測試這個模型的學習成果是好還是不好。如果學習成果不彰，也就是準確率不高，可能要再回頭看看資料集是否有缺失或需要再蒐集，也可以看看是否要改用其他演算法，或是調整演算法的參數等等。

經過如此反覆的嘗試後，最終會訓練出一個準確度很好的模型，這時，你的機器學習之旅，將進入最後一個步驟，我們要實際拿它來上場測試囉！

2-1-5 推論與預測：給 AI 考試

終於來到最後一個步驟。前面經過重重關卡，終於訓練出我們的 AI 模型，接著來測試看看。在機器學習裡，將資料匯入訓練好的模型進行預測的過程，我們稱為**推論（Inference）**。

機器學習基本上不會給一個絕對的答案，而是給各種候選答案的機率。以前面水果辨識的資料集為例，透過這個資料集訓練好一個機器學習的模型後，當你給它一張新的水果圖片做判斷時，機器學習的模型會給出一個最可能的答案，也許機率有 90% 是「蘋果」這個分類，也許機率有 70% 是「香蕉」這個分類。所以當發現最後模型給出的答案還是不夠好，或者是你想要它能認識更多種類的水果時，就要回到資料蒐集，再進行資料處理，接著再重新訓練出新的模型，如此反覆循環。

想想看

1. 在機器學習的過程裡，資料蒐集、資料處理與選擇演算法，哪個步驟對模型的影響較大？為什麼？
2. 下方連結為鐵達尼號生存預測的資料集，請說說看這份資料中你看到哪些特徵？有哪些資料是需要補齊的？多提供哪些資料，可能會讓預測結果更準確？

https://www.kaggle.com/c/titanic/data

2-2 總結

最後，我們總結一下本章節的重點。機器學習的五大流程，分別是：

1. 定義問題　2. 蒐集資料　3. 處理資料　4. 訓練模型　5. 推論及預測

最重要的前提是要先把問題定義清楚，針對這個問題搜集相關的資料，接著將資料整理成適合電腦處理的格式，例如：資料庫、csv 檔案、文字檔、圖片檔等。最後，選擇適合的機器學習演算法，將資料匯入演算法中進行學習訓練，再使用建立好的模型，將新的資料送給模型進行預測。如果準確率還不錯，代表訓練成功；如果準確率不如預期，必須重新審視資料集是否哪裡有缺失，或是有數量不足等問題，重新修正後，再重新訓練及預測。一直重複這個循環，直到成果滿意為止。

只要有耐心地依循這五個步驟，就可以一步步地打造出屬於自己的機器學習模型囉！