本書介紹的內容

本書以下列的讀者為目標，整理機器學習與資料分析的工具該如何於職場應用，也介紹目前渾沌不明的機器學習專案如何發展。

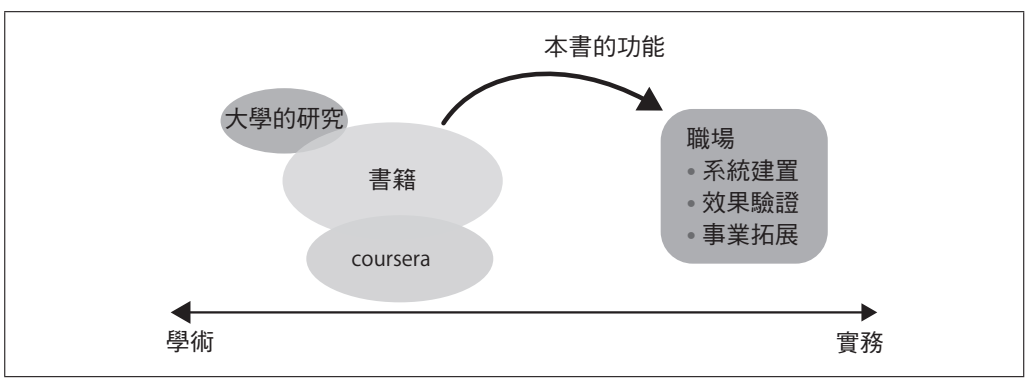
- 學完機器學習的入門教材，想於實務應用的工程師
- 想將大學學到的機器學習經驗應用於專案的年輕工程師

更具體的是介紹下列的內容：

- 該如何啟動機器學習的專案
- 該如何讓機器學習與現存的系統互動
- 該如何收集機器學習的資料
- 該如何建立假設與分析

本書一開始是以機器學習初學者為對象，最後會提到理論，適合軟體工程師實務應用的形式。

市面上已有許多書籍介紹演算法，所以本書以第一次執行專案的讀者為對象，並以收集系統建置與學習的相關資源為主題，介紹「實際該怎麼做呢？」這類讀者覺得有興趣的內容。



第一章

推動機器學習專案的方法

本章統整的是推動機器學習專案的方法。

機器學習專案比開發一般的電腦系統還要求預測的準確度，也需要反覆的試作或重作，所以先了解重點再推動是非常重要的。一開始先從機器學習的概要介紹，後續再介紹專案的流程、機器學習特有的問題以及如何組織一個成功推動專案的團隊。

1.1 機器學習都如何應用？

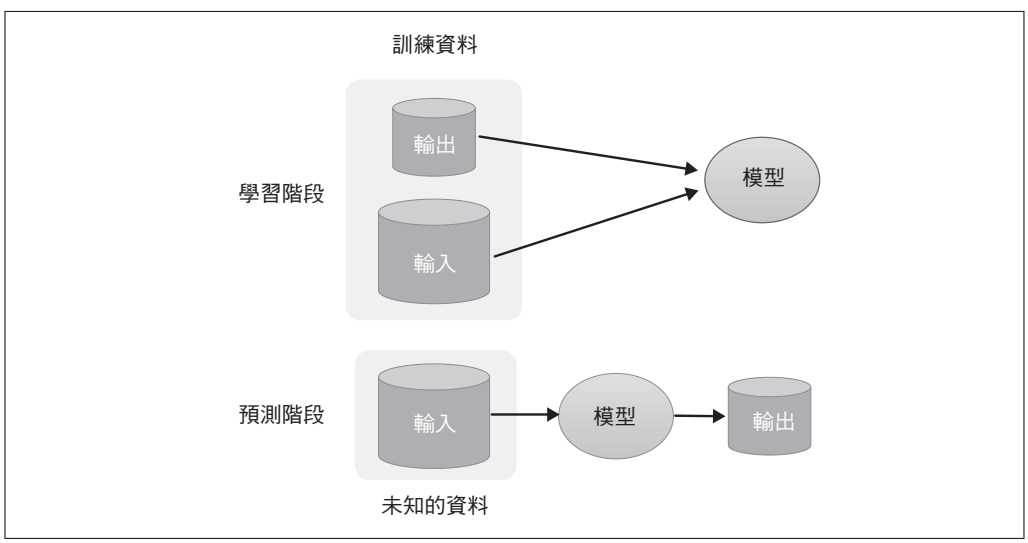


圖 1-1 機器學習（監督式學習）的概要

其他

推薦：提出使用者可能有興趣的項目或是與使用者目前正在瀏覽的項目類似的項目

異常檢測：偵測不正常的存取或是其他有別以往的動作

頻繁模式探勘：從資料篩選出常出現的模式

增強學習：在圍棋或將棋這種正確解答部分不明朗的環境下，學習應該採取的行動

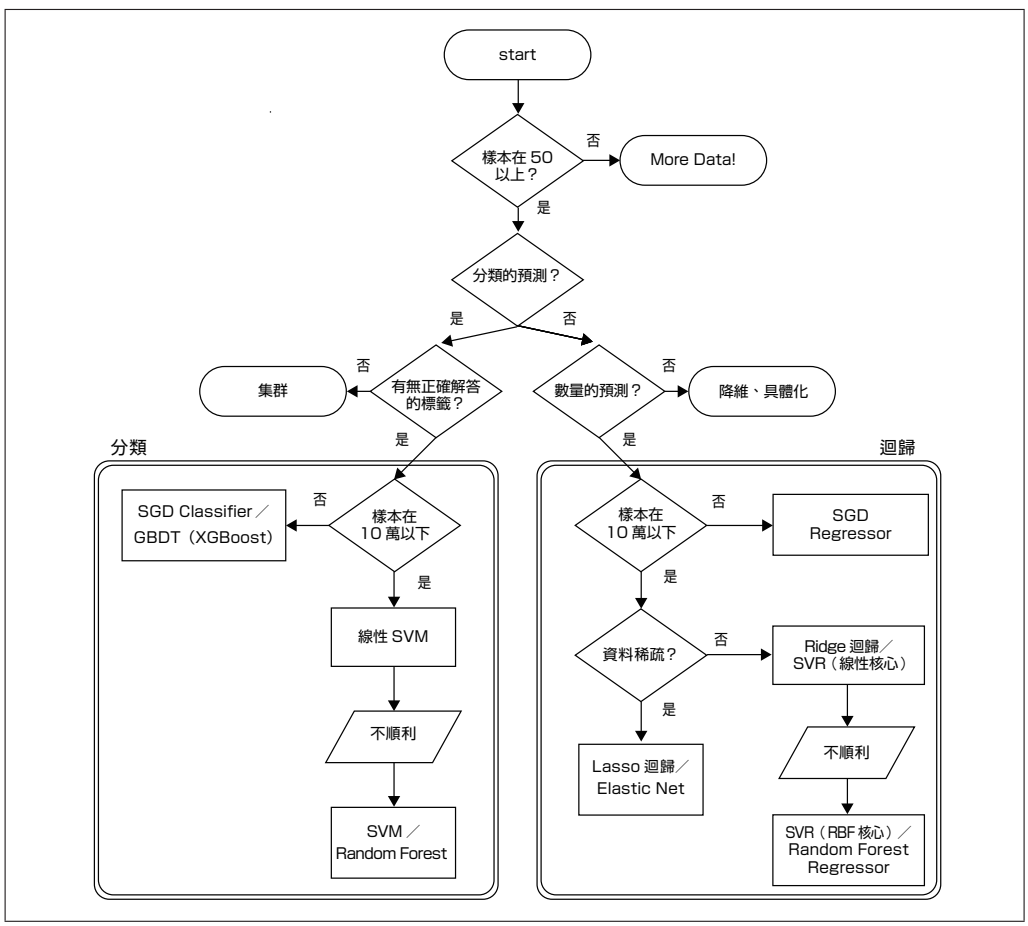


圖 2-1 該選擇何種演算法？

大部分的分類問題都可在了解後續介紹的目標函數以及決策分界線之後了解差異之處。後續的內容盡可能不寫公式，希望大家多閱讀圖表的內容。

接下來，就為大家一一介紹演算法。

2.2.1 感知學習機

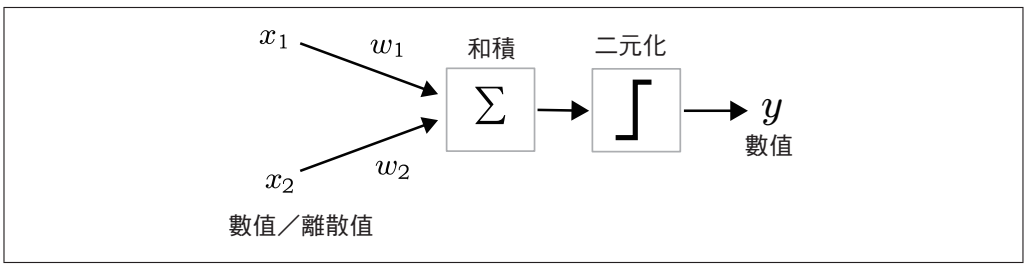


圖 2-2 感知學習機

感知學習機（perceptron）是將輸入向量與學習過的加權向量相乘之後的值加總，若結果大於 0，則分類為類別 1，若結果小於 0，則分類為類別 2 的演算法[4]。將感知學習機重疊成多層構造後，就是後續介紹的類神經網絡。接下來讓我們學習使用感知學習機的分類方式。

感知學習機的特徵

感知學習機具有下列特徵：

- 可在線學習
- 預測性能差強人意，學習速度很快
- 容易過度擬合
- 只能解決可線性分離的問題

4 感知學習機的活化函數（後述）雖然是使用步階函數，但仍可使用其他的函數。

這裡出現了好幾個不熟悉的用語吧。由於其他的演算法也會提到這些用語，所以就在此依序說明這些用語。

在線學習（Online Learning）與反義語的**批次學習**（Batch Learning）分別是逐步輸入資料再最佳化資料（在線學習）或先輸入所有資料再最佳化（批次學習）的方法。詳情請參考「4.2.1 容易混淆的「批次處理」與「批次學習」」。

過度擬合（Overfitting）的預測模型則在前一章的時候說明過，是「能正確解答訓練資料，卻完全無法處理未知資料」的模型。過度擬合是機器學習常見的現象之一，減少特徵值，採用後述的正規化項，使用更單純的演算法，可避免這個現象發生。

早期的感知學習機未採用抑制過度擬合的機制。

與過度擬合相反的現象稱為**乏適**（Underfitting），指的是模型未能反應輸入與輸出的關係。乏適會因為未包含該領域該有的特徵值或模型的合理性不足或是正規化項的影響過強而發生。

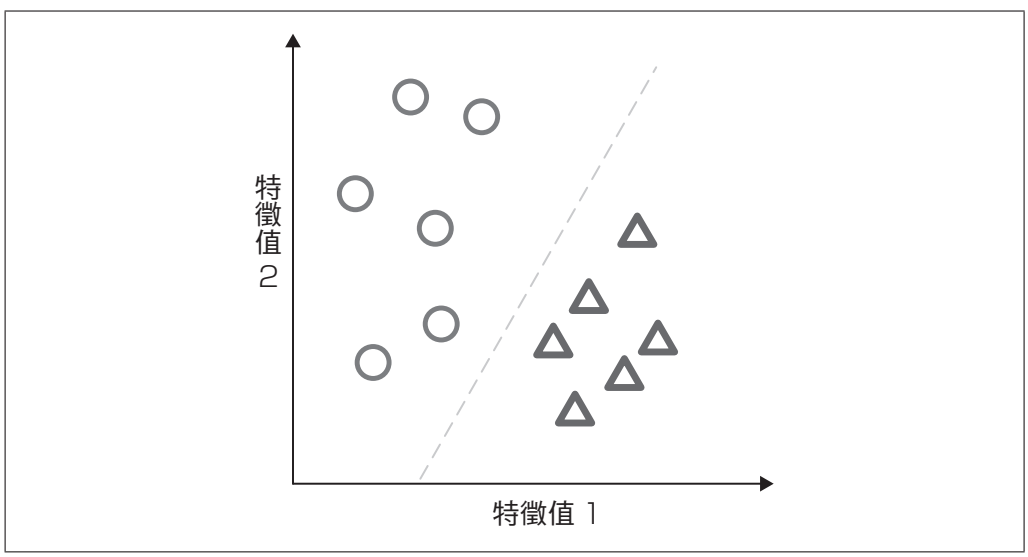


圖 2-3 可線性分離的資料

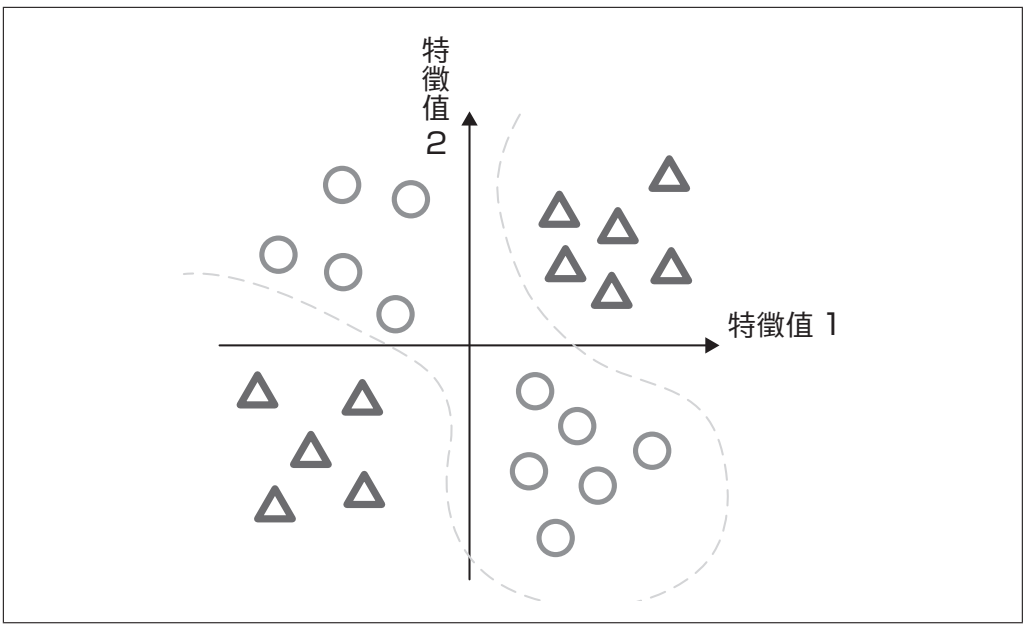


圖 2-4 不可線性分離的資料

感知學習機只能處理**可線性分離**（Linearly Separable）的問題。所謂「可線性分離」就是如**圖 2-3** 所示，能以直線將資料一分為二的意思。若以稍微專業的說法形容這條分類的直線，就稱為**超平面**（hyperplane）。二次元的時候，這條直線只是直線，但是三次元的時候，這條直線就成為平面，而在高次元的空間裡，這個平面就稱為超平面。

反之，如果是**圖 2-4** 這種無法以直線區分的資料就稱為非線性分離資料。最常見的例子就是**邏輯異或**（XOR、Exclusive or）的資料。如**圖 2-4** 所示，XOR 的資料是以原點為中心，將右上（橫軸與直軸皆為正值的區塊）、左下（橫軸與直軸皆為負值的區塊）當成一個類別，並將右下（橫軸為正，直軸為負的區塊）與左上（橫軸為負，直軸為正的區塊）當成一個類別。因此，無法以一條直線適當地分割這兩個類別。這種「無法單憑一條直線分割兩個類別」的情況就稱為非線性分離。

感知學習機的決策分界線

實際學習感知學習機的模型的決策分界線請參考圖 2-5 與圖 2-6。圖 2-5 是可線性分離的資料（之所以無法完全線性分離是因為有一些雜訊）。圖 2-5 與圖 2-4 一樣，都是以原點為中心，在右上角與左下角產生●的資料，同時在右下角與左上角產生▲的資料。感知學習機無法在非線性分離的資料使用，所以決策分界線會是直線。所以 XOR 也未呈分離的狀態。此外，繪製這條決定分界線的 notebook 放在資源庫的 chap02/Decision_boundary.ipynb。

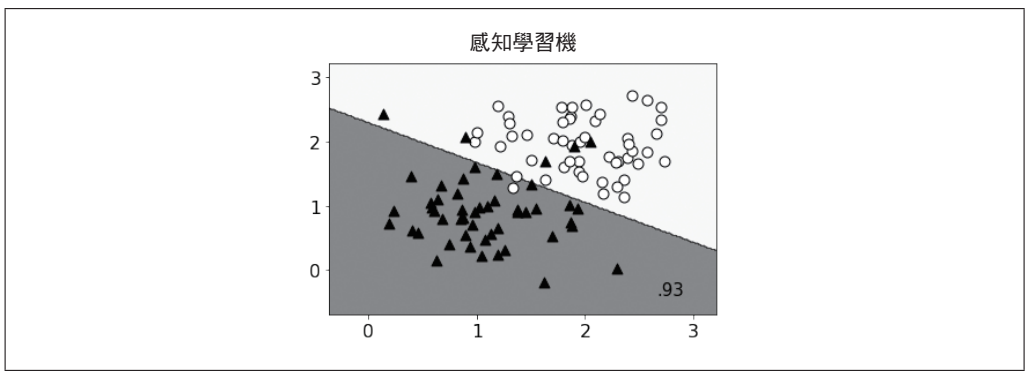


圖 2-5 感知學習機的決策分界線（可線性分離）

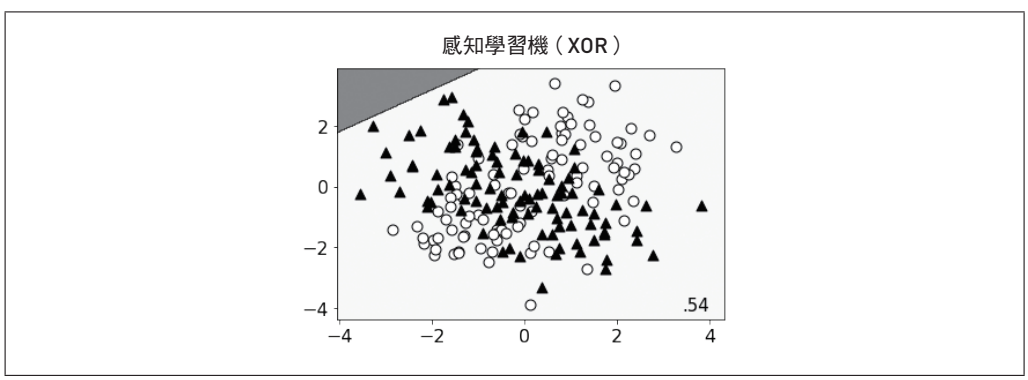


圖 2-6 感知學習機的決策分界線（不可線性分離）

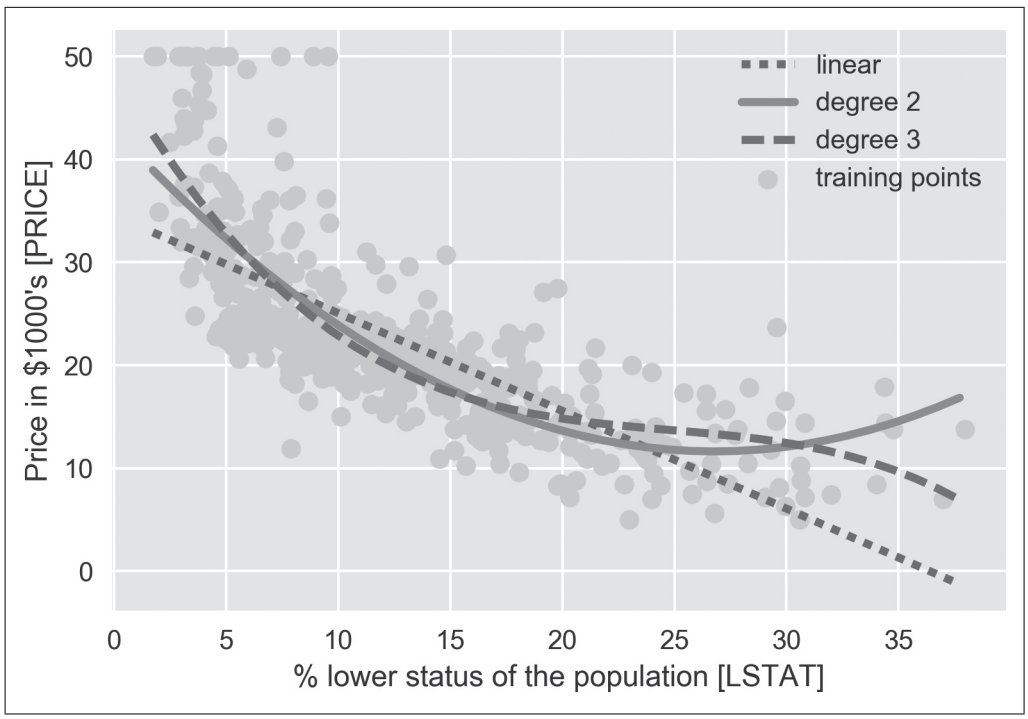


圖 2-36 線性迴歸與多項式迴歸的範例

圖 2-36 是針對美國房租資料的線性迴歸與多項式迴歸的圖表。以直線粗略呈現資料的是線性迴歸，以使用二次曲線或三次曲線的多項式的曲線趨近資料的是多項式迴歸。輸入橫軸的值之後，就傳回以直線或曲線推測的房租值。舉例來說，以線性迴歸學習房租與平均年收的關係性之際，可找出房租 = a × 平均年收 + b 這條直線公式。此時會盡可能地朝好的方向學習 a 與 b 的係數（線性迴歸的參數）。換言之，學習線性迴歸或多項式迴歸的模型可得到與各變數相乘的權重。

2.4 集群、降維

這一節要說明的是集群與降維。

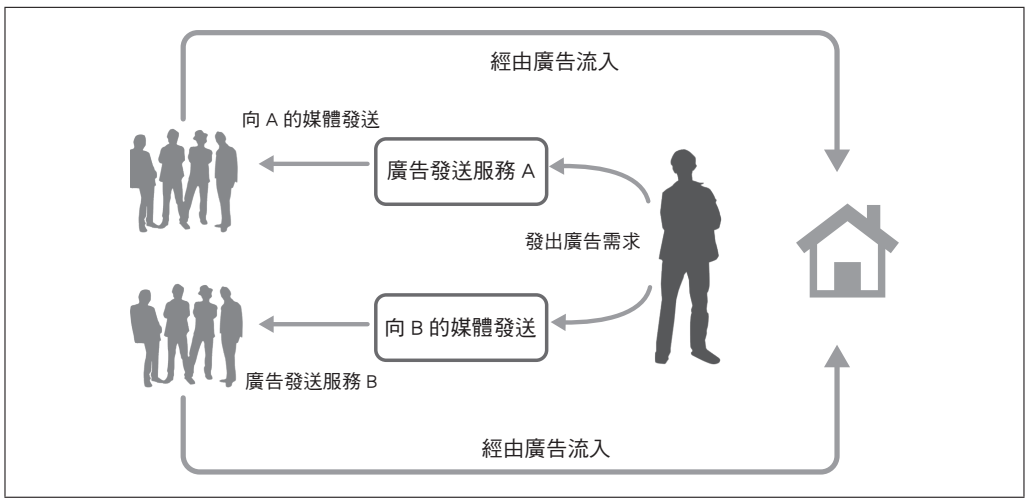


圖 6-5 雙方的使用者在品質上有落差嗎？

表 6-3 經由不同通路流入的使用者是否繼續使用服務的資料

流入通路	流入人數	持續使用人數	持續使用率
A	205	40	19.5%
B	290	62	21.4%

接著試著具體呈現持續使用率的分佈。由於樣本大小有一定的規模，所以執行二項分佈的常態近似計算。

```
# 測試資料。持續使用人數，放棄使用人數
a = [40, 165]
b = [62, 228]

print('Sample A: size={}, converted={}, mean={:.3f}'.format(sum(a), a[0],
a[0]/sum(a)))
print('Sample B: size={}, converted={}, mean={:.3f}'.format(sum(b), b[0],
b[0]/sum(b)))

Sample A: size=205, converted=40, mean=0.195
Sample B: size=290, converted=62, mean=0.214
```

6.4 因果效果的推測

假設檢定是從樣本推測母體性質的手法，接下來則是要推測對母體的效果。為了解決「Y 現象之所以發生，是受到 X 因素多少影響」這個問題，讓我們一起看看因果推論裡的因果效果。

6.4.1 Rubin 的因果模型

接下來讓我們思考網路廣告的效果。在因果推論裡，將顯示廣告這項行為稱為**介入（Cause）**，並將購買行為稱為**結果變數（Outcome）**，同時將介入的樣本稱為**處理組**或**實驗組（Treatment Group）**，將未介入的樣本稱為**對象組**或**控制組（Control Group）**。

廣告效果可根據看了廣告與沒看廣告時，有無產生購買行動來推斷。不過，以個人為實驗對象時，可觀察的結果變數只限於是否顯示廣告。若假設接觸廣告的 A 先生未接觸廣告，就是一種**反事實（Counter Factual）**的推測，此時就無法觀測（**圖 6-11**）。

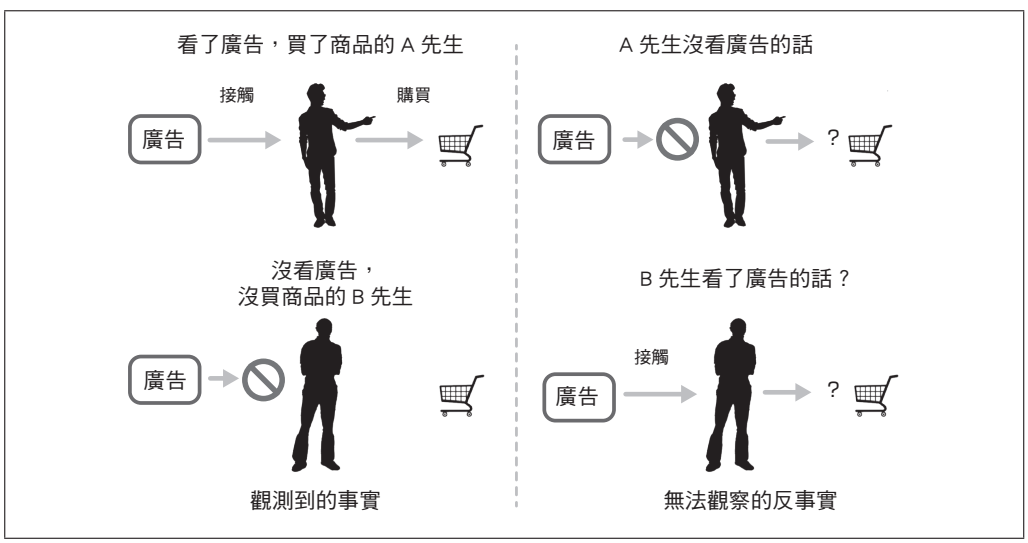


圖 6-11 以個人為觀測單位時，只能預測單側的情況

8.6 觀察目標金額達成卻取消的專案

AnyTouch Blue[4] 是插入 USB 埠，讓智慧型手機與藍牙連接，藉此將智慧型手機當成虛擬鍵盤或虛擬滑鼠的產品。目標金額為 20,000 美元，實際募集金額接近 3 萬美元，但是專案卻取消了，然後以目標金額 5,000 美元重新啟動專案（2017 年 3 月的資料）。

這個重新啟動的專案[5] 將最低 Back 金額從 18 美元降至 16 美元。撰寫本書時，目標金額為 5,000 美元，實際募集金額超過 2 萬美元，所以這個專案也確定成功了。

比較這兩個專案之後，差異最明顯的部分就是顧客單價。取消的專案的顧客單價為 29,763[美元]/235[Backer]=126[美元 /Backer]，但是新啟動的專案卻是 21,189[美元]/449[Backer]=47[美元 /Backer]。

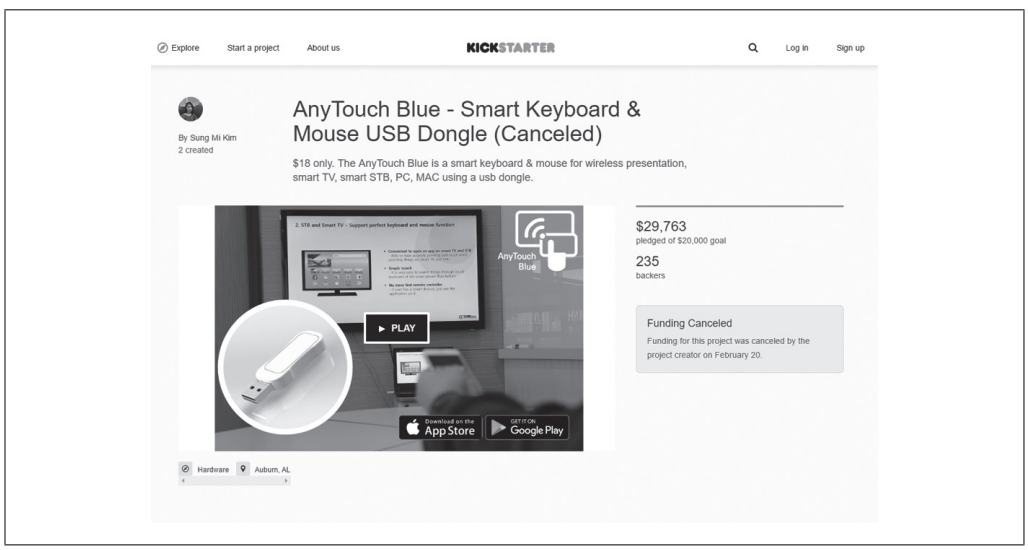


圖 8-11 取消的 AnyTouch Blue 專案的頁面

4 https://www.kickstarter.com/projects/2094324441/anytouch-blue-smart-keyboard-and-mouse-usb-dongle/

5 https://www.kickstarter.com/projects/2094324441/anytouch-blue-smart-keyboard-and-mouse-usb-dongler

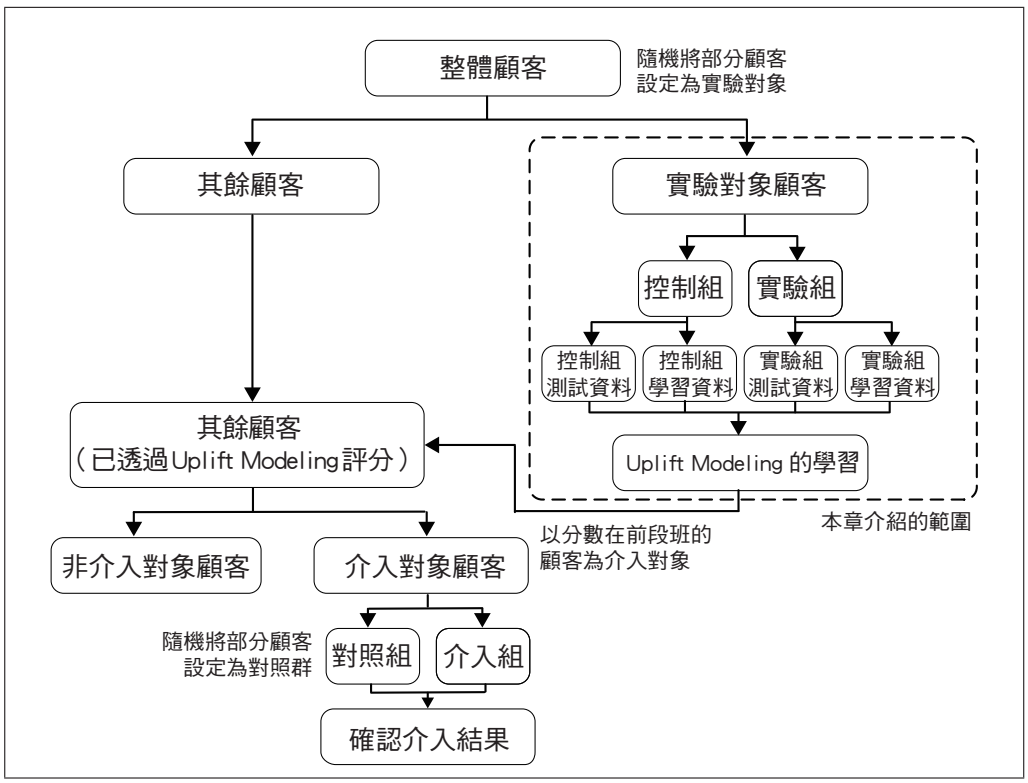


圖 9-13 於正式環境應用 Uplift Modeling 的流程

超過一定分數的顧客不需建立對照組，也能全面實施介入行為。不過此時只能透過測試資料求出的推測值了解轉換率增加多少。

透過上述流程於正式環境應用 Uplift Modeling，可驗證轉換率會在實施 Uplift Modeling 之後增加多少，也能利用 Uplift Modeling 求得的業績佐證自己的主張。

Uplift Modeling 的知名度較低，也難以主張效果，所以建議建立介入組與對照組。