人工智慧概論

CH03: 踏入 AI 領域第一步：AI 技術導論

National Taiwan Ocean University
Dept. Computer Science and Engineering

Prof. Chien-Fu Cheng

AI 技術導論

- 人工智慧就是機器模擬人類的認知能力的技術。
- 人工智慧透過資料來訓練 (Training) 進行學習，稱做機器學習(Machine Learning)。

3-1 認識資料與數據

- 資料可分成結構化資料 (Structured Data)、半結構化資料 (Semi-Structured Data) 以及非結構化資料 (Un-Structured Data)。

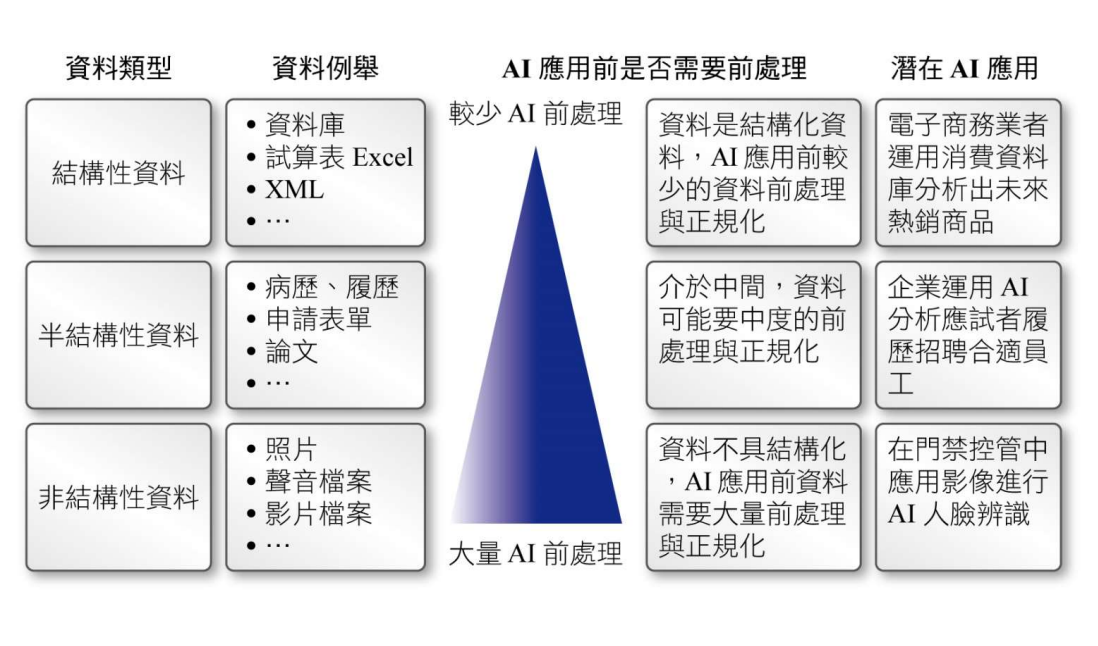


圖3-2 人工智慧上的資料結構性分類。

3-1 認識資料與數據

➢ 結構化資料

- 結構化資料意指，它擁有固定欄位、固定格式與順序等等。

表 3-1　結構化資料範例。

日期	訂票人 ID	出發車站	到達車站	票種	座位 ID
2020-10-15 16：00	李大明 （A100000000）	台北	台南	自由座	無

- 另外，結構化資料並非一定以資料庫形式儲存，也可以存在 Excel、XML。

3-1 認識資料與數據

```
- <Cell ss:Index="2">
    <Data ss:Type="String">日期</Data>
  </Cell>
- <Cell>
    <Data ss:Type="String">訂票人ID</Data>
  </Cell>
+ <Cell>
- <Cell>
    <Data ss:Type="String">到達車站</Data>
  </Cell>
- <Cell>
    <Data ss:Type="String">票種</Data>
  </Cell>
- <Cell>
    <Data ss:Type="String">座位ID</Data>
  </Cell>
</Row>
- <Row ss:Height="51">
  - <Cell ss:StyleID="s90" ss:Index="2">
      <Data ss:Type="String">2020-10-15 16:00</Data>
    </Cell>
  - <Cell ss:StyleID="s90">
      <Data ss:Type="String">李大明 (A100000000)</Data>
    </Cell>
  - <Cell>
      <Data ss:Type="String">台北</Data>
    </Cell>
  - <Cell>
      <Data ss:Type="String">台南</Data>
    </Cell>
  - <Cell>
      <Data ss:Type="String">自由座</Data>
    </Cell>
  - <Cell>
      <Data ss:Type="String">無</Data>
    </Cell>
</Row>
```

圖3-3 結構化資料XML 範例。

3-1 認識資料與數據

非結構化資料

- 非結構化的資料，重點不在內容是什麼，而是「資料格式」。非結構化資料通常是多媒體檔案內容，舉凡有圖片、聲音檔案、影片、影音串流...等等。

半結構化資料

- 半結構化資料介於結構化與非結構化之間。
- 這種檔案格式是具備有「欄位」特性的，因此可以根據欄位特性進行查找所要的資料。但是每個欄位卻有它自己的描述方式，無法確保一致性。
  - 舉例來說，履歷表中有專長描述，但每個人的描述方式卻大大不同，有人可能以條列式說明，有人以敘事方式進行說明。

3-1 認識資料與數據

資料與數據的轉換

- 原生資料 (Raw Data) 都傳遞了一些訊息 (Information)，必須將依資料特性加以轉換，轉換成數據，然後再進行分類 (Classification) 分析。
- 將圖片資料中擷取出數據來，這個數據稱之為特徵值 (Feature)，透過特徵值將圖片予以分類。

3-1 認識資料與數據

表 3-2　資料與數據的轉換。

資料	數據	分類
	2 個輪子 車長 2.2 公尺	機車
	4 個輪子 車長 4.5 公尺	汽車
	4 個輪子 車長 2 公尺	機車

資料來源：Pixabay。

3-2 機器學習

- 機器學習分成監督式學習 (Supervised Learning) 以及非監督式學習 (Unsupervised Learning)。

➢ 機器學習基礎架構

- 要能像人一樣區分圖片裡是汽車、還是機車，就要建構區分汽機車的分類器 (Classifier)：

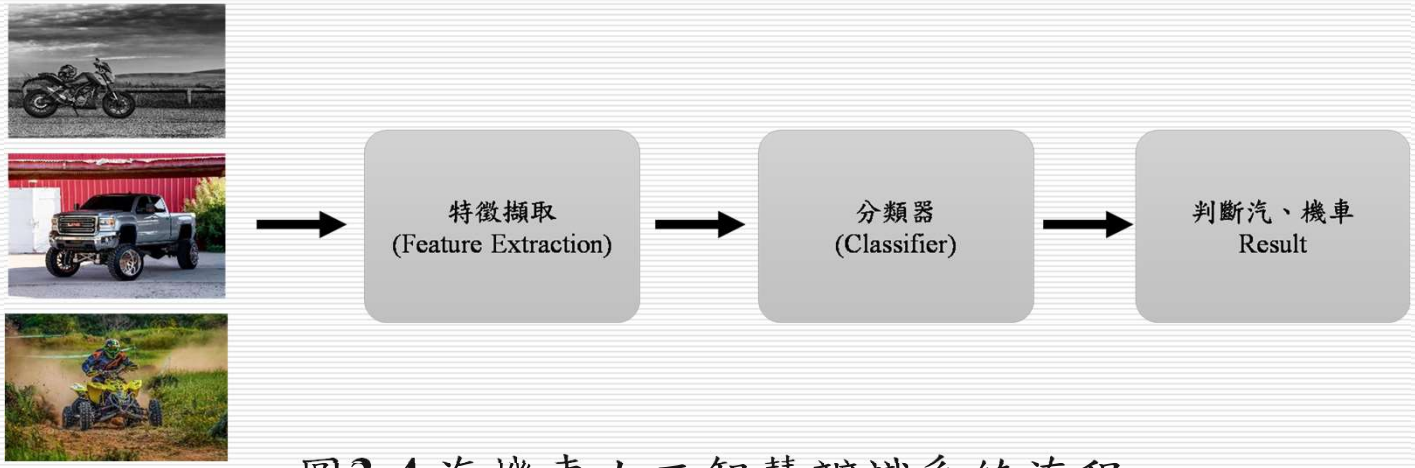


圖3-4 汽機車人工智慧辨識系統流程。

3-2 機器學習

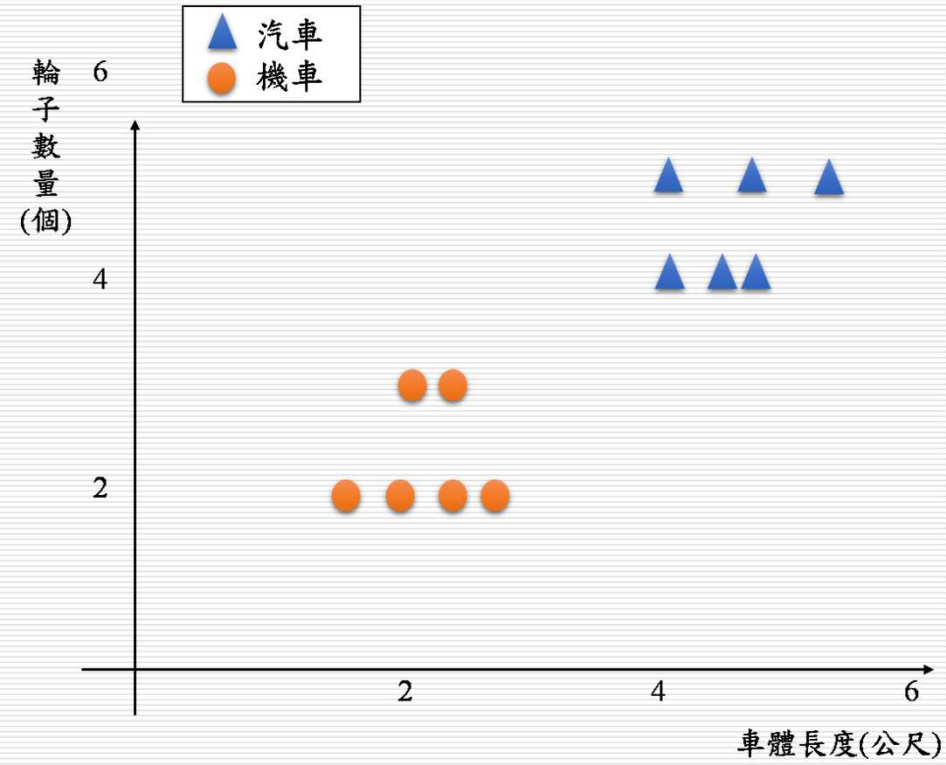
➢ 提取特徵值

- 特徵值本身對於該事物是否有代表性，將會是分類器分類準確與否的重要因素之一。
- 在汽機車圖片辨識的這個案例中，輪子數量與車體長度是我們選擇的兩個特徵。
- 用x1 表示輪子數量、x2 表示車體長度。然後用數學式表示成為 (x1, x2)，這個數學表示式稱做特徵向量 (Feature Vector)。

3-2 機器學習

- 把這些圖所代表的特徵向量放到直角坐標，看成是直角坐標中的一個點，這個點稱做是這個圖的特徵點 (Feature Point)。將更多張交通工具圖的特徵向量放到直角坐標系後，這個就稱做是如何將這些圖片進行分類的特徵值空間(Feature Space)。
- 特徵點與特徵點的距離 (Distance) 為d，d 代表的就是兩個特徵點的相似程度。

3-2 機器學習



樣本	輪子數量	車體長	分類
1	2	2.3	機車
2	4	4.5	汽車
3	5	4.3	汽車
4	2	2.5	機車
5	4	4	汽車
6	3	2.2	機車
7	4	4.6	汽車
8	3	2	機車
9	2	1.9	機車
10	5	4	汽車
11	2	2	機車
12	5	5.3	汽車
…	…	…	…

圖3-5 汽機車特徵向量的特徵值空間

3-2 機器學習

➢ 建構分類器

- 分類器函數有很多個，最簡單的就是一條直線，當特徵點帶入函數大於 0 者為一類，特徵點帶入函數小於 0 則為另一類。

$$f(x1, x2) = \begin{cases} +1, x1 + x2 - 6 \geq 0 \\ -1, x1 + x2 - 6 < 0 \end{cases}$$

3-2 機器學習

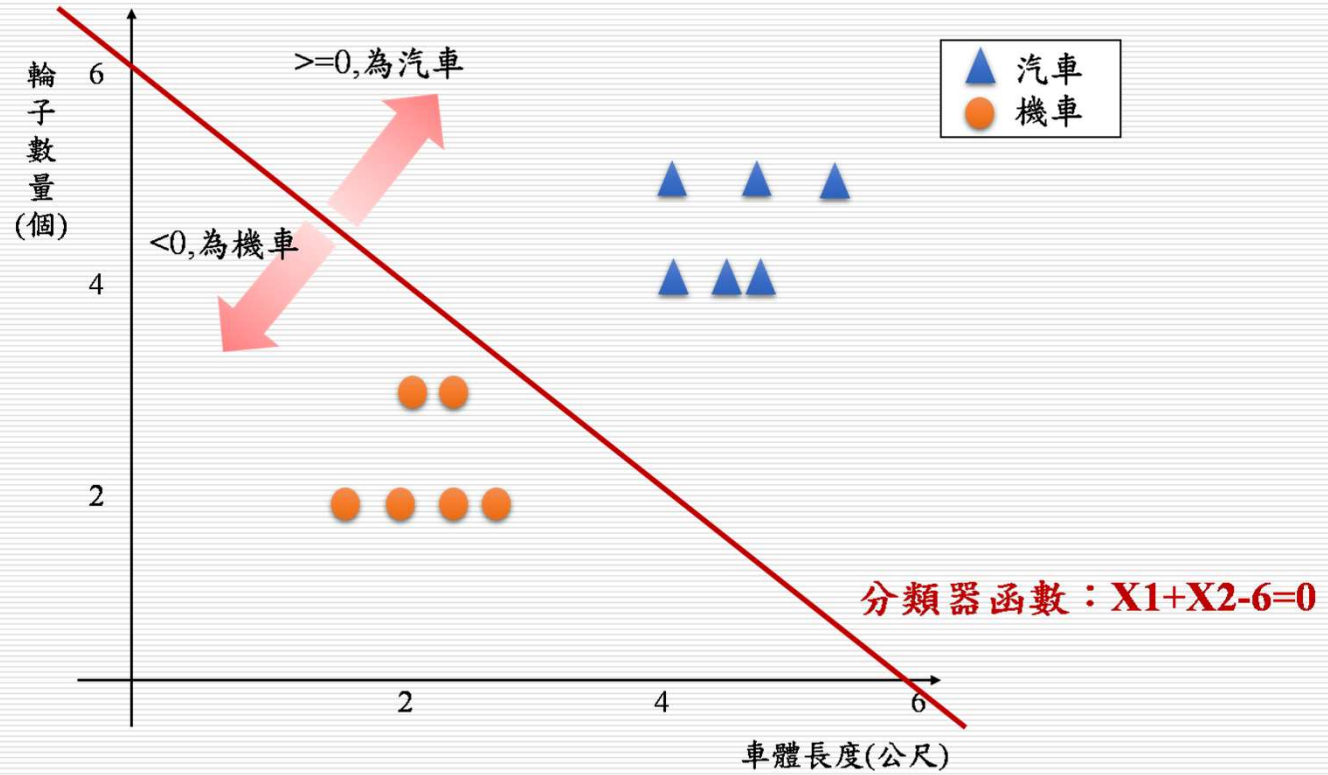


圖3-6 汽機車特徵值空間分類器。

3-2 機器學習

- 這種分類模式也稱做直線分類器 (Linear Classifier)，特徵空間通常不會只是二維平面 (Two-Dimension space)，如何讓分類器自己收斂出分類函數，稱做訓練分類器。

圖3-7 特徵向量運用分類器進行分類方法。

3-2 機器學習

➢ 訓練資料對分類器的重要性

- 機器的學習透過訓練 (Training)、然後驗證 (Validation)、最後識別 (Recognition)。

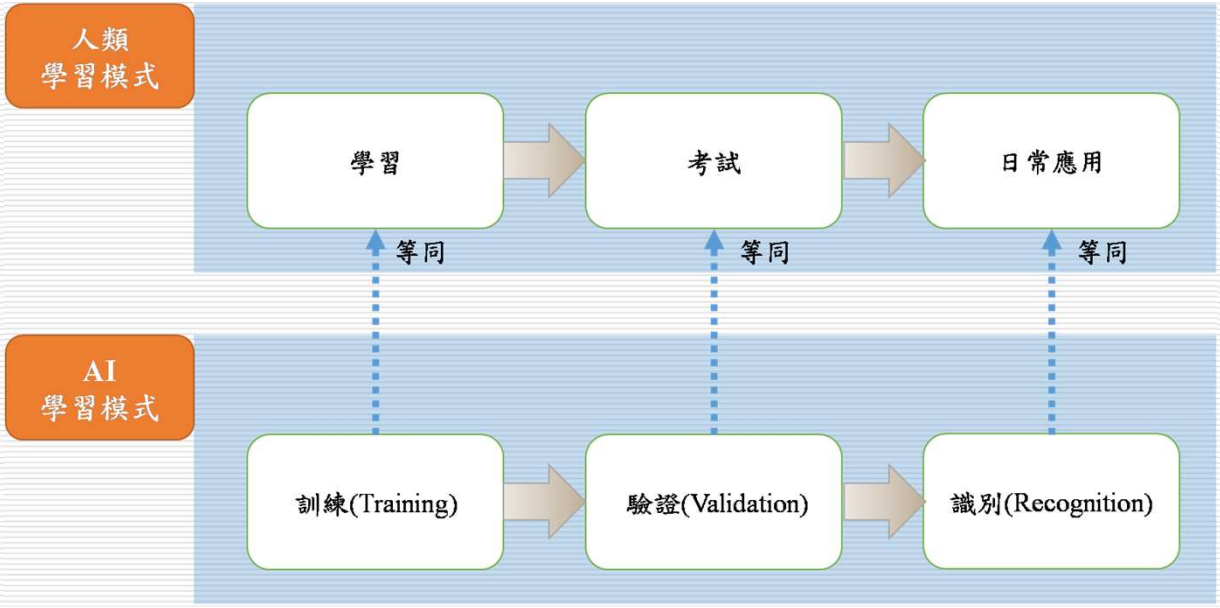


圖3-8 人工智慧學習模仿人類行為。

3-2 機器學習

表 3-3　汽機車特徵資料集。

樣本	輪子數量	車體長	分類	樣本	輪子數量	車體長	分類
1	2	2.3	機車	8	3	2	機車
2	4	4.5	汽車	9	2	1.9	機車
3	5	4.3	汽車	10	5	4.	汽車
4	2	2.5	機車	11	2	2	機車
5	4	4	汽車	12	5	5.3	汽車
6	3	2.2	機車	…	…	…	…
7	4	4.6	汽車				

3-2 機器學習

- 當我們從這個資料集抽取數個樣本出來作為分類器訓練，這被抽取出來的數個樣本則稱作訓練資料集。那1 ~ 8 個樣本，我們就稱之為訓練資料集 (Training Dataset)。9 ~ 12個樣本則稱做為測試資料集。
- 人工智慧在訓練分類器後也需要「驗證」訓練後的模型是不是符合期望。
- 「驗證」
  - 有點像是我們在一門課程學習後，就要有期中考進行考評，學生根據試卷答題、老師批改、然後給予評分。
  - 如果分數不好，代表學習狀況不好，就要再回過頭來修正一下學習的方式。
  - 如果分數在某一個標準之上，那代表學習狀況不錯，則這個學習是有成效的。

3-2 機器學習

交叉驗證 (Cross-Validation)

- 當用一個分類器函數進行分類，就要測試驗證分類器分類的精準程度。這裡有兩個重要觀念：
  - 分類器不一定需要把所有的特徵點完全分類正確。
  - 特徵點資料必須分成訓練數據 (Training Data) 和驗證數據 (Validation Data)。

分類器驗證

- 分類正確率 (Classification Accuracy, CA)。

$$\text{分類正確率 CA} = \frac{\text{分類正確樣本數}}{\text{測試樣本總數}} \times 100\%$$

3-2 機器學習

➢ 驗證模式

- 最常見的分類器驗證模式是把訓練資料分成五等分子資料集(Subset)：k，稱做是5-fold。
  - 其中4等分的子資料集用來進行訓練，1等分的子資料集用來測試。
- 為了能反映出分類器能有效分類所蒐集到的所有數據，所以必須輪流讓所有的子集合擔任測試資料集。也稱之為交叉驗證(Cross-Validation)。

3-2 機器學習

- 視情況需求分成k 等分，這個驗證模式稱做k 等分交叉驗證(k-fold Cross-Validation)。

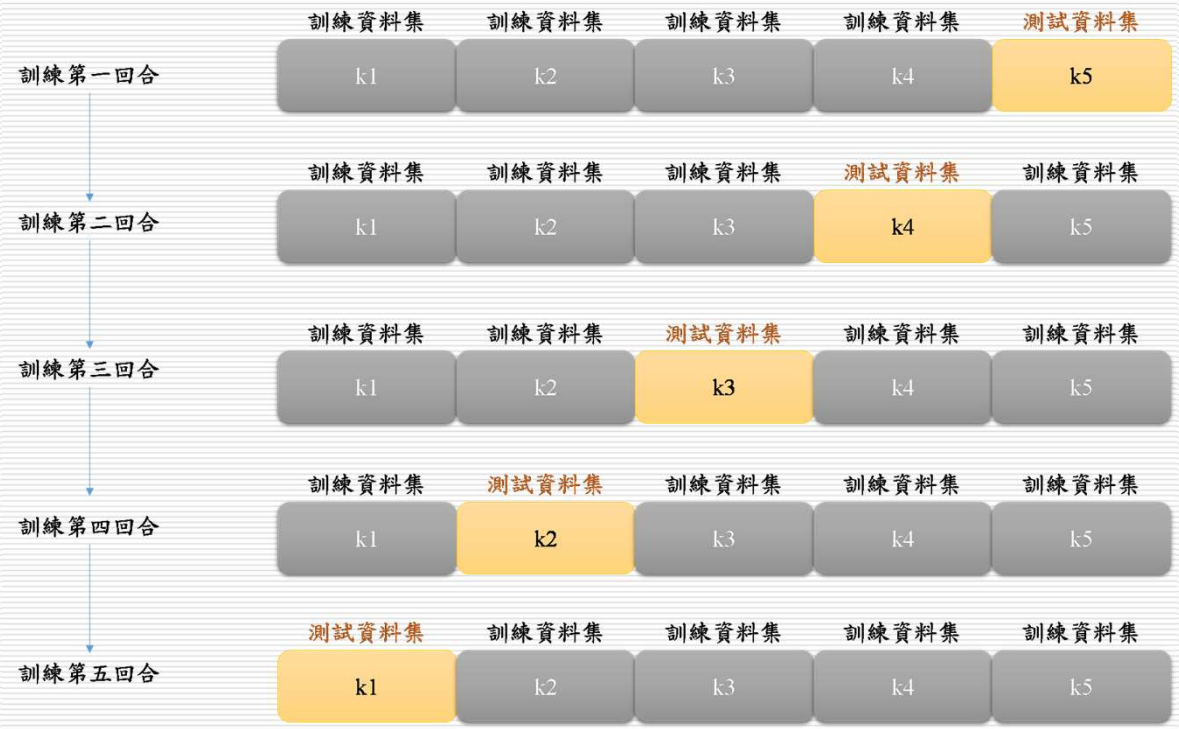


圖3-9 分類器交叉驗證模式。

3-2 機器學習

➢ 多類別分類器

- 多類別分類 (Multi-Class Classification) 是日常生活上最常遇到的問題。作法有很多種：
  - 第一種作法是建構許多個二元分類器。
  - 第二種多元分類是建構一個多元分類函數，然後將欲分析的照片特徵值帶入計算，將會得到所有類別的相似程度，也可以稱做可能是該類別的機率。

3-2 機器學習

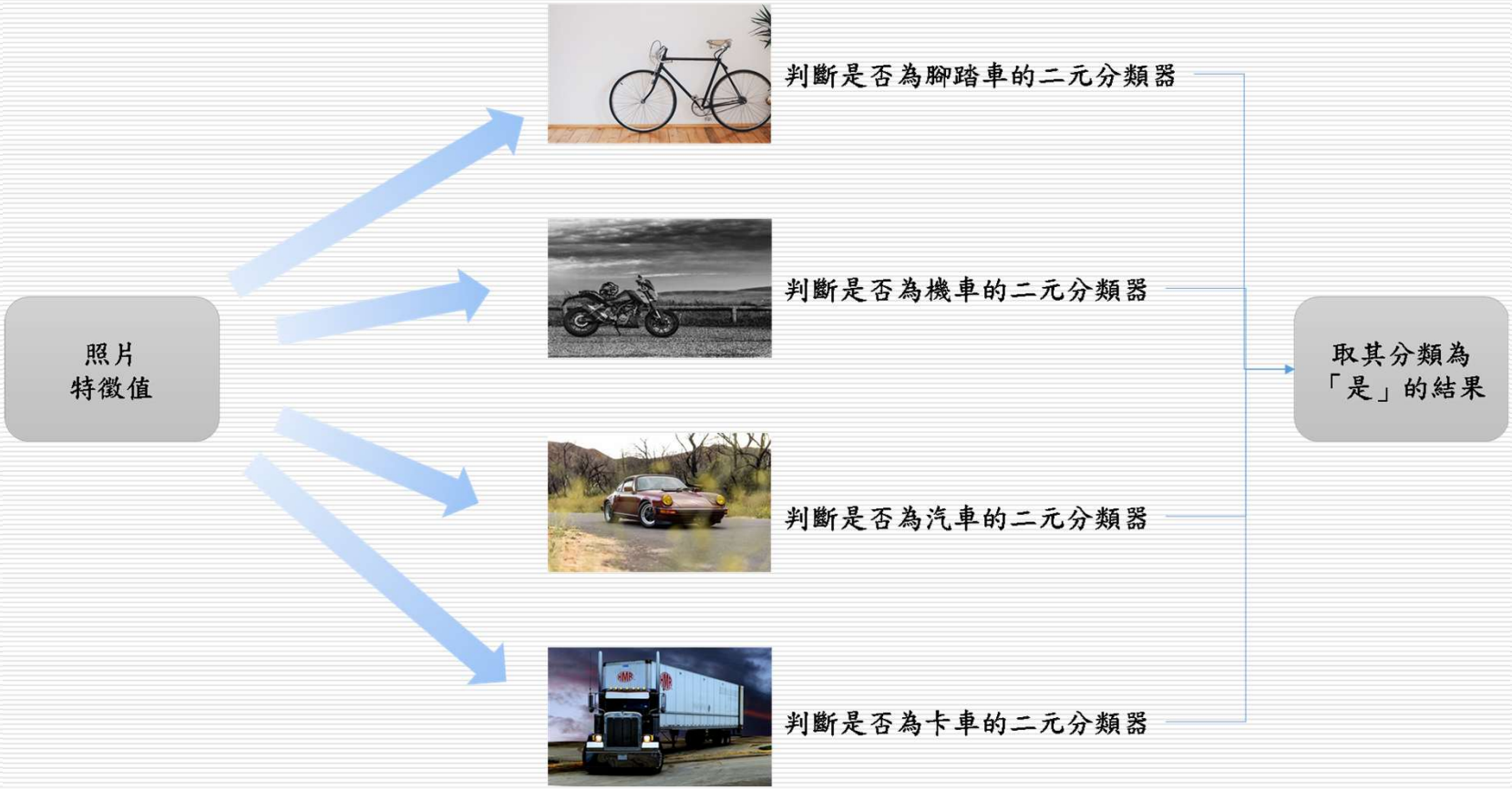


圖3-10 第一類型的多類別分類器。

3-2 機器學習

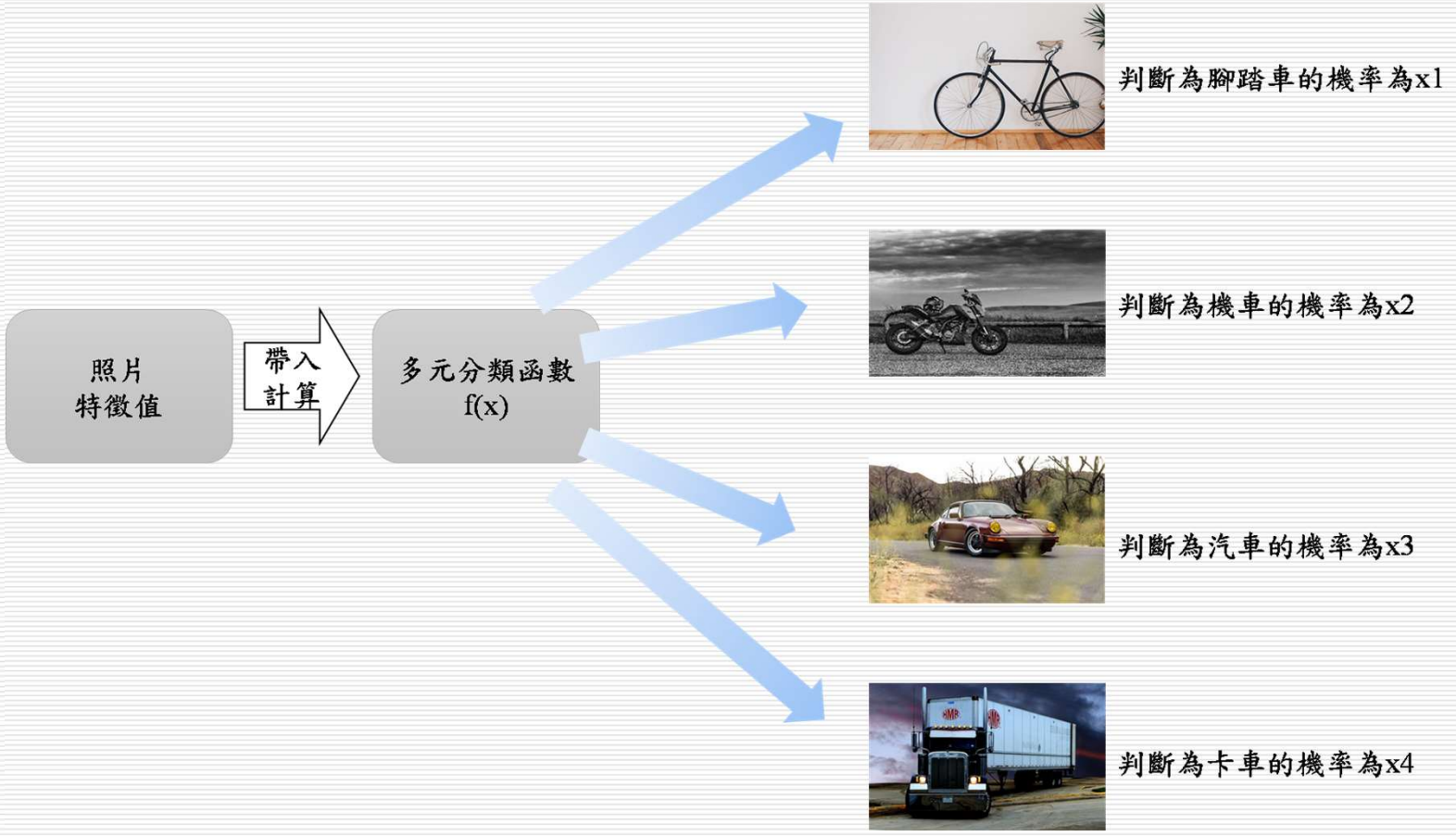


圖3-11 第二類型的多類別分類器。

3-3 人工智慧的基礎

- 本單元介紹簡單的分類概念，作為人工智慧技術導論。在機器學習下，還有分成「分類」以及「分群」不同作法、監督式學習及非監督式學習不同的概念、在深度學習下還有類神經網路等不同架構。

Sources

- 投影片資料來源說明：
  - 本投影片之內容出自於書商所提供之投影片，並根據實際授課需求進行補充及修改。