



Manufacturing Data Science

Feature Engineering, Data Augmentation and Data Balance

(第 12 章 特徵工程、數據增強與數據平衡)

Chia-Yen Lee, Ph.D. (李家岩 博士)

Department of Information Management (資訊管理學系)
National Taiwan University (國立臺灣大學)

- 第一章 製造數據科學
- 第二章 製造系統分析與管理
- 第三章 數據科學基礎與模型評估
- 第四章 數據科學分析架構與系統運算決策
- 第五章 數據預處理與製造數據特性
- 第六章 線性分類器
- 第七章 無母數迴歸與分類
- 第八章 決策樹與集成學習
- 第九章 特徵挑選與維度縮減
- 第十章 類神經網路與深度學習
- 第十一章 集群分析
- 第十二章 特徵工程、數據增強與數據平衡
- 第十三章 故障預測與健康管理
- 第十四章 可解釋人工智慧
- 第十五章 概念漂移
- 第十六章 元啟發式演算法
- 第十七章 強化學習

藍：老師課堂講授
綠：學生自學

- 附錄A 線性迴歸
 - 附錄B 支持向量機
 - 附錄C 統計製程管制與先進製程控制
 - 附錄D 超參數最佳化
-
- 應用涵蓋
產能規劃、瑕疵檢測、製程監控與診斷、機台保養、需求預測、生產排程、電腦視覺、自動光學檢測、原料價格預測與採購等

□ 特徵工程

□ 數據增強

□ 數據平衡

□ 結語

- 在數據科學分析架構中，在數據收集後我們會進行數據預處理，為了提升資料的資訊量(質與量)我們常使用以下三種方式
 - 特徵工程 (feature engineering)
 - 數據增強 (data augmentation)
 - 數據平衡 (data balancing)
- 本章節將說明此三者在數據分析中所扮演的角色及方法。

□ 數據的質通常體現在特徵上，而量則體現在觀測值上

- 特徵工程的功能是強化、萃取特徵的特性，專注於質的部分，並從原有的特徵數 q 擴增至 p 個
- 數據增強與數據平衡的功能則是擴增具有資訊量及稀少類別的觀測值（在分類問題中），專注於量的部分，並從原有的樣本數 m 增至 n 個



圖 12.1 特徵工程、數據增強與數據平衡

□ 特徵工程目的

- 特徵工程是使用領域知識或對數據特性的理解，將原始數據轉換並淬鍊出多個新的特徵，其目的在於強化特徵的特性使得我們能轉換出更具解釋性的特徵以及訓練出預測精準的模型。

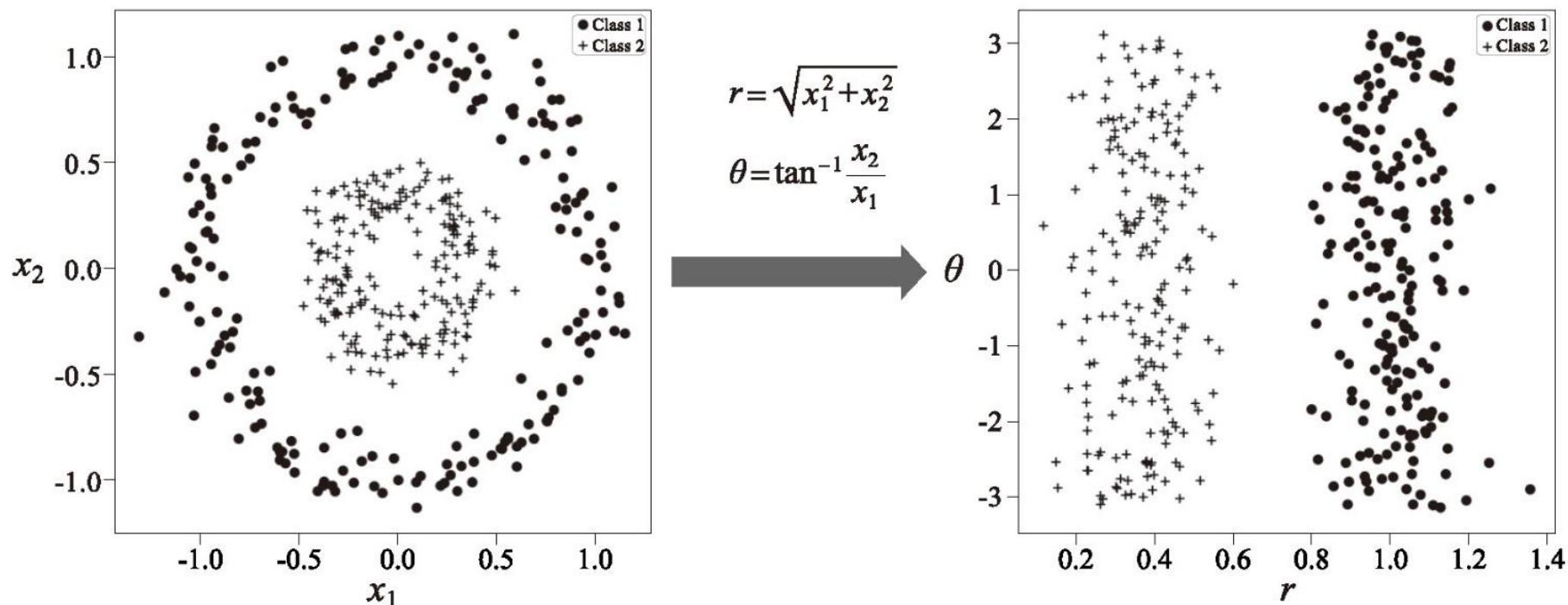


圖 12.2 原始數據與新特徵的類別分析

□ Methods

● Temporal Features

- For date / time, we can transform to timestamp and extract “day”, “month”, and “year” as new columns.
 - “hour” can be binning as “morning/afternoon/night”, or different shifts(早/晚班)
 - “day” can be binning as “weekday”, “weekend”
 - For weather, holiday, or event, we can build “is_national_holiday”, “has_xxxx_events” (eg. related to company, fab, machine, labor, ...)
- “MTBF related to the same failure type”
 - Time aggregation/average by “category(downcode)”
- “Number of tuning/MO in the last week”



● Image Features

- Colorful to Gray Sale/Black-or-white, Rotation, Convolution, Pooling, etc.

● Text Features

- Chopping, stemming, lemmatization, Word2Vec/GloVe/Doc2Vec, string similarity, TF*IDF

● Spatial Features

- Location in space, such as GPS-coordinates, cities, countries, addresses
- Latitude and longitude can build the “median_distance_within_2_miles”

□ 典型數據特徵工程

- 在典型數據中，有兩種主要的特徵工程方法
 - 「領域知識」手動地轉出特徵
 - 單一特徵的轉換
 - 連續特徵的交互作用
 - 類別特徵的組合
 - 「特徵學習」自動化地產生特徵
 - 深度學習
 - 符號迴歸

● 單一特徵的轉換

- 主要透過領域知識體察到某些特定的數據對於目標欄位（應變數、標籤）有所影響，因此對其數據做特徵轉換。
- 從製造的數據中，可以知道每一批量經過機台加工時，產品所使用的處方（recipe）其相關的工程參數。若我們根據連續變數溫度這個欄位，進行特徵工程，可以轉出二元變數1與0表示高溫與低溫作為新特徵。
- 若我們有一筆房地產的數據，目標是預測每個不動產的房價。若已知某不動產建案的經緯度（欄位），可以透過特徵工程產生其他幾項新的特徵
 - 例如產生新特徵二元變數來描述三百公尺內是否有便利商店？
 - 五百公尺內是否有捷運站點？
 - 一公里內是否有警察局 / 消防局？
 - 三公里內是否有醫院？
 - 五公里內是否有國中 / 高中？
 - 十公里內是否可上國道交流道？
- 這些新的特徵或許有助於房價的預測。

● 連續特徵的交互作用

- 藉由基於某項**領域知識**來進行兩個或兩個以上的特徵的相互加、減、乘、除的計算，這使得我們能在建模前淬煉出某些可能有潛在物理特性的新特徵，而非利用模型本身去考量交互作用。
- 這不僅減輕模型負擔（複雜度）與降低維度，更有機會提高模型準確度與解釋性
- 若在生產數據中存在兩個特徵分別為**溫度**與**壓力**，若加工產品的規格量測值會受到氣體的影響，此時我們或許可創造一個新特徵是由溫度除以壓力所得到的交互作用項，若根據**波以耳定律**與**給呂薩克體積定律**，此交互作用項會正比於氣體體積，因而可姑且命名它為「**氣體體積因子**」
- 根據前例房地產數據，欲預測不動產的房價。
 - 若在原始數據中存在兩個特徵分別為**學校總數**與**學校評鑑分數**，
 - 前者是計算五公里內的學校總數，後者是在五公里內這些學校評鑑分數的中位數
 - 基於房地產的領域知識可知，與房價有關的不僅僅是這兩個特徵各別的特性，而是同時考慮學校的**質與量**
 - 因而我們可以創造一個新的特徵是由**學校總數**乘上**學校評鑑分數**所得到的交互作用項，並且命名它為「**學區總分**」。

● 類別特徵的組合

- 針對類別特徵，重視單一變數的水準數 (level) 以及在不同水準下的樣本數，這是因為當我們將一個類別變數以「獨熱編碼」轉成多個「二元變數」時，無形中拓展了特徵維度，因而導致維度詛咒的發生，並且每個二元變數可能會成為「稀疏變數」，因此如何合併水準或重新編碼都是相當重要的特徵工程。而同樣地，類別水準的合併需基於某項領域知識。
- 延續上述房地產的案例。
- 若在原始數據中存在一個類別特徵為不動產的外牆材質，此特徵具有14個水準包含了木頭 (wood) 、磚塊 (brick) 、混凝土塊 (concreteblock) 等
- 有些水準下的樣本數相當稀少，例如混凝土塊的樣本數就小於50
- 因此，基於對材料特性的領域知識，我們能將相似的 (similar) 與稀疏的 (sparse) 的特徵水準進行合併
- 將與木頭相關的水準如木頭、木側板 (wood siding) 、木瓦 (wood shingle) 合併到木頭的水準
- 較為稀疏的水準如混凝土塊 (concrete block) 、灰泥 (stucco) 、砌石 (masonry) 合併到其他 (others) 的水準
- 因此合併後的新特徵一共僅有8個水準，且每個水準均有一定的樣本數，將有利於後續數據分析。事實上，這是處理數據不平衡的一種手法。

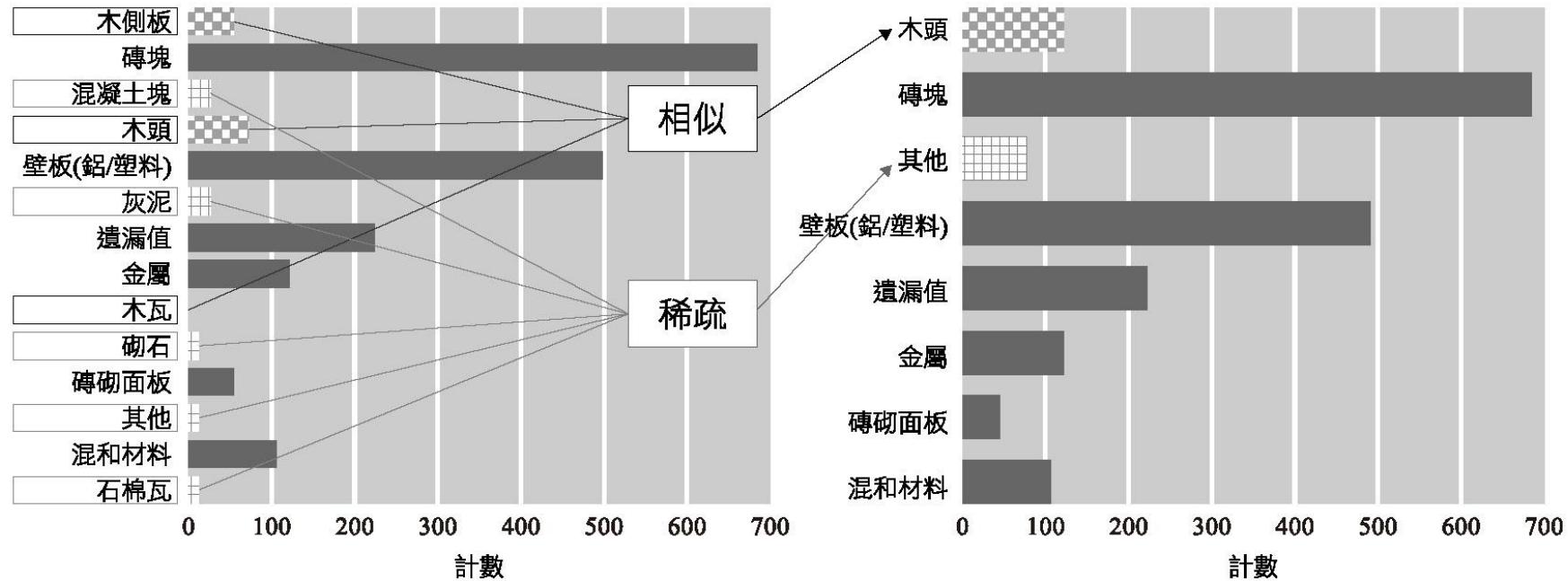


圖 12.4 不動產外牆材質的數據分布以及類別水準的合併

● 類別特徵的組合

- 在降低類別變數的水準數之後，接著進行特徵工程，考慮類別特徵的組合
- 舉例來說，若在生產數據中存在兩個類別特徵分別為溫度與材料
 - 溫度中有三水準：低溫度、中溫度與高溫度
 - 材料有兩水準：金線與銅線
- 透過特徵組合可以產生溫度×材料的新類別特徵，其水準數有 $3 \times 2 = 6$ 種
- 換言之，這新特徵考慮了原先特徵的**交互作用**項，以利後續數據分析
- 對於一個連續變數對一個類別變數特徵的組合，可以透過計算其每一類別水準的**所對應連續變數的統計量**（例如中位數、平均數、最大 / 最小值、變異數 / 標準差等），來萃取新的特徵。
- 如前例，若生產數據中存在連續變數溫度與類別變數材料，則可以產生新特徵組合材料的溫度變異數（熱膨脹係數），其計算使用**金線下的溫度變異數**及使用**銅線下的溫度變異數**，該新特徵整合了兩個特徵並提供新的資訊量以利後續數據分析。

● 特徵學習

- 先創造出新的特徵後，再試著由領域知識加以驗證與解釋
- 深度學習：特徵學習常透過深度學習方法，使用模型網路結構中生成的特徵(feature map)或是預測結果來產生新的特徵。
 - 例如使用「受限波爾茲曼機」以隨機生成神經網絡來學習輸入數據集的機率分布
 - 也可使用非監督式學習方法的「自編碼器」或「生成對抗網路」(GAN)等方法來萃取新特徵

● 符號迴歸

- 其屬於一種迴歸分析的形式，藉由搜尋由特徵任意組合成的數學式集合，求解出一組數據配適最佳的數學式。
- 對於解的搜尋採用的是元啟發式演算法，其核心思維為探索與開採（exploration and exploitation）。在產生出任意的初始解集合後，藉由任意交換組合進行解的探索，並挑選出表現最好的子集進行解的開採，重複這樣的動作進行配適（fitness）函數的優化，以找出全域最佳解或近似最佳解。

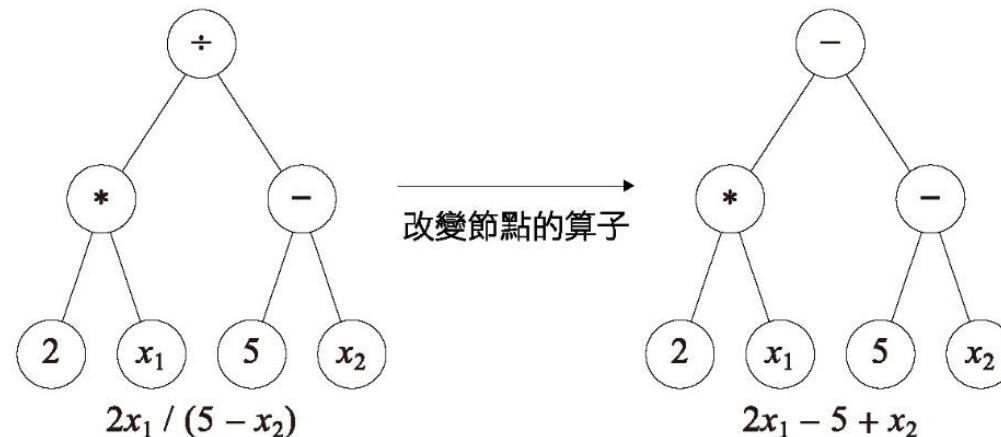


圖 12.5 符號迴歸的樹狀結構與節點改變造成數學式的變化

- 在求解出最佳數學式後，從此數學式中觀察特徵的數學函數特性以及特徵間的交互作用，從而回過頭判斷在領域知識上是否具有物理意義或能否被加以解釋，驗證後再進一步特徵工程轉換出新的特徵。

□ 在時間序列數據中，有三種主要的特徵工程方法

- 時間標記特徵 (timestamp feature)
- 移動窗口特徵 (sliding window feature)
- 時間序列分解 (time series decomposition)

● 時間標記特徵 (timestamp feature)

— 在時間序列數據中，當我們有每個觀測值完整的時間標記（年月日時分秒）時，便可從中萃取出多個重要的時間標記特徵，並且這些與時間有關的特徵萃取亦需要**領域知識**的協助。

- 從製造的數據中，可以知道每一批量經過每一個工作站的時間點，因此根據時間點這個欄位，便可以進行特徵工程
- 這時間點是哪一個月份生產的，可轉出新的類別特徵代表1 到12 月
- 這時間點是哪一個值班班別生產的，可轉出新的二元變數代表早班與晚班
- 這時間點為周間 (weekday) 或周末 (weekend)
- 這時間點為國定假日與否、假期與否 (holiday or not)
- 營業時間與否 (business hour or not)
- 每個星期的銷售量
- 每個星期的發生的機台調機次數
- 每個星期的機台當機次數
- 每個星期的MO (missing operation) 次數

- 移動窗口特徵 (sliding window feature)

- 在時間序列預測問題中，若我們想以監督式學習的模型預測而非典型的統計模型時，需事先將時間序列以**滑動時窗**的方式轉換成特徵對目標值的數據形式，才得以建構模型。
- 若以一個樣本數為8的時間序列 (t_1, \dots, t_8) 為例，並且設定時窗大小為4，也就是運用前三期 $(t_{k-3}, t_{k-2}, t_{k-1})$ 預測後一期 (t_k) ，因而最終轉換成一個具有三個**滯後特徵**(lagged feature)以及目標值的數據，因此不同的時窗大小決定了我們使用多少的滯後特徵(考慮離現在多遠的時間點)進行預測。

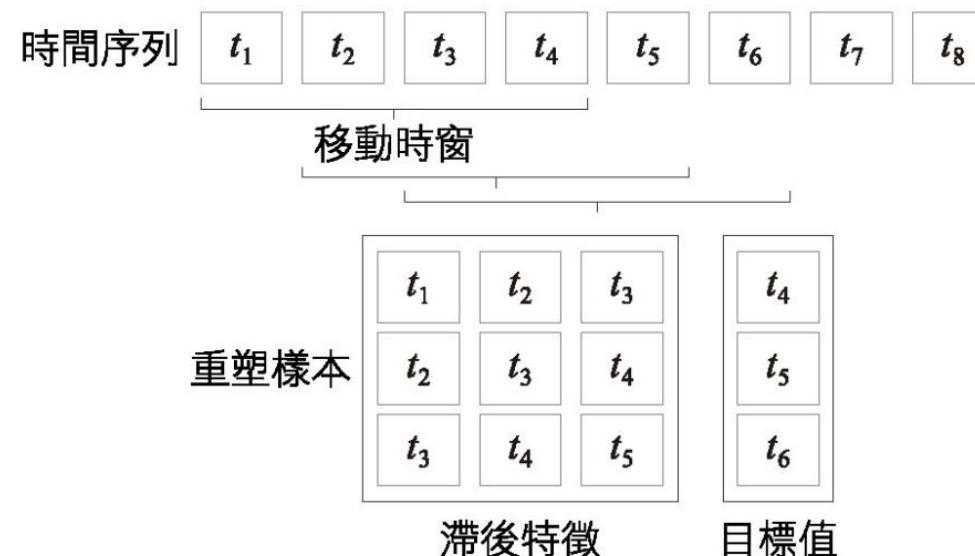


圖 12.6 滑動窗口的數據轉換

● 移動窗口特徵 (sliding window feature)

- 時間序列也可以是高頻訊號中「**時域**」(time domain)常見的弦波，因此我們可對其樣型(pattern)的統計特性進行特徵工程，例如萃取最大值(max)、最小值(min)、平均值、標準差等相關的統計量作為新特徵

滑動時窗特徵	說明	公式
最大值	時窗中的最大值	$\max(x_t)$
最小值	時窗中的最小值	$\min(x_t)$
全距 / 峰對峰值	時窗中最大值減最小值	$f_1 - f_2$
最大最小間距	時窗中最大值與最小值於時間上的間隔	$T(f_1) - T(f_2)$
平均數	時窗中分配的平均數	$\bar{X} = \frac{1}{n} \sum_{t=1}^n x_t$
中位數	時窗中分配的中位數	$median(x_t)$
分位數	時窗中分配的第一、三分位數	$1st\ quartile(x_t)$ $3rd\ quartile(x_t)$
標準差 (變異數)	時窗中分配的標準差 (變異數)	$\sqrt{\frac{1}{n} \sum_{t=1}^n (x_t - \bar{X})^2}$
三階主動差 (偏態)	時窗中分配的偏態	$\frac{1}{n} \sum_{t=1}^n (x_t - \bar{X})^3$
四階主動差 (峰態)	時窗中分配的峰態	$\frac{1}{n} \sum_{t=1}^n (x_t - \bar{X})^4$
滯後 k 期自我相關係數	時窗中與滯後 k 期的自我相關程度	$\frac{\sum_{t=1}^{n-k} [(x_t - \bar{X})(x_{t+k} - \bar{X})]}{\sum_{t=1}^n (x_t - \bar{X})^2}$
斜率	時窗中的總斜率	$\sum_{t=2}^n (x_t - x_{t-1})$
轉折個數	時窗中斜率正反轉次數	$\sum_t I_{(x_{t+1}-x_t)(x_t-x_{t-1}) \in \{negative\ sign\}}$
平均振幅	時窗中的平均振幅	$\frac{1}{n} \sum_{t=1}^n x_t $

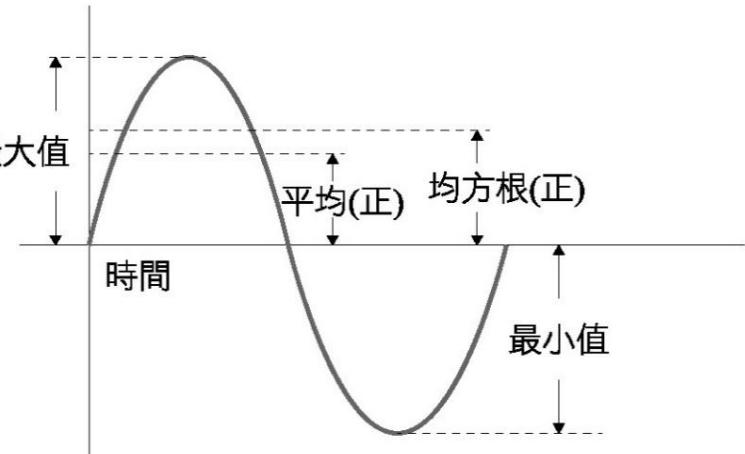


圖 12.7 時間序列與時域弦波的特徵工程

滑動時窗特徵	說明	公式
均方根值	時窗中的均方根值	$\sqrt{\frac{1}{n} \sum_{t=1}^n (x_t)^2}$
波形指標 (waveform)	時窗中的均方根值除以平均振幅	$\sqrt{\frac{1}{n} \sum_{t=1}^n (x_t)^2} / \frac{1}{n} \sum_{t=1}^n x_t $
脈衝指標 (pulse/impulse)	時窗中的最大值除以平均振幅	$\max(x_t) / \frac{1}{n} \sum_{t=1}^n x_t $
振幅均方根次方	時窗中的振幅均方根次方	$(\frac{1}{n} \sum_{t=1}^n \sqrt{ x_t })^2$
餘隙指標 (clearance) 邊際指標 (margin)	時窗中的最大值除以振幅均方根次方	$\max(x_t) / (\frac{1}{n} \sum_{t=1}^n \sqrt{ x_t })^2$
波峰指標 (crest)	時窗中的最大值除以均方根值	$\max x_t / \sqrt{\frac{1}{n} \sum_{t=1}^n (x_t)^2}$
形狀指標 (shape)	時窗中的均方根值除以平均振幅	$\sqrt{\frac{1}{n} \sum_{t=1}^n (x_t)^2} / \frac{1}{n} \sum_{t=1}^n x_t $

時間序列特徵工程

● 時間序列分解

— 將其拆解成多個成分，接著分別對每個成分進行模型的建構與預測，也就是**分而治之**最後再合併一起。

主要有三大成分，分別為

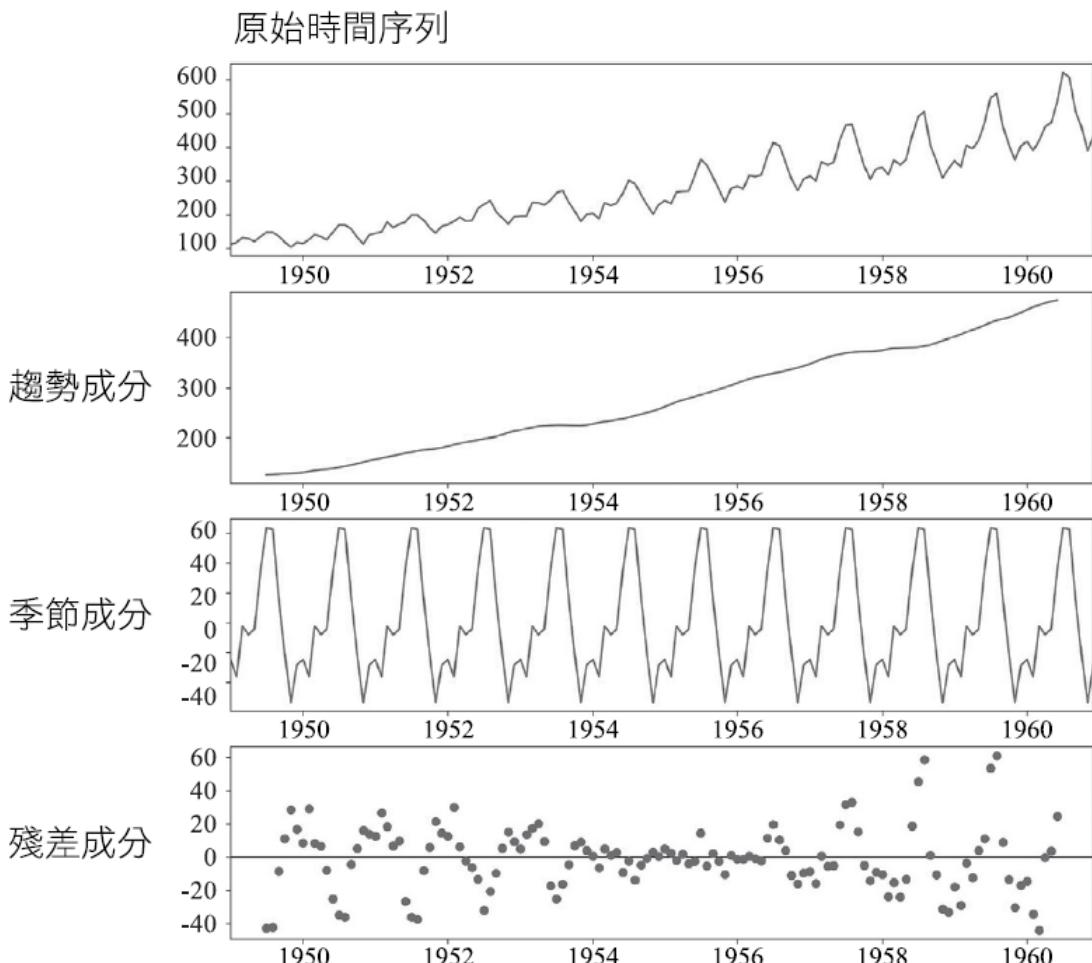
- **趨勢成分 (trend)**：代表時間序列長期趨勢的變換，可用全局且簡單的模型進行預測。
- **季節成分 (seasonal)**：代表時間序列於固定時間的週期變換，同樣可用較為全局且簡單的模型進行預測。
- **殘差成分 (residual/irregular)**：代表隨機或短週期的波動，需用較為局部且複雜的模型進行預測。

— 先瞭解原始的時間序列比較適用於「**加法模型**」或是「**乘法模型**」

➤ 加法模型(additive model) : $y_t = T_t + S_t + I_t$

➤ 乘法模型(multiplicative model) : $y_t = T_t \cdot S_t \cdot I_t$

— 模型的選擇取決於殘差是否為「**異質變異數**」 (heteroscedasticity)

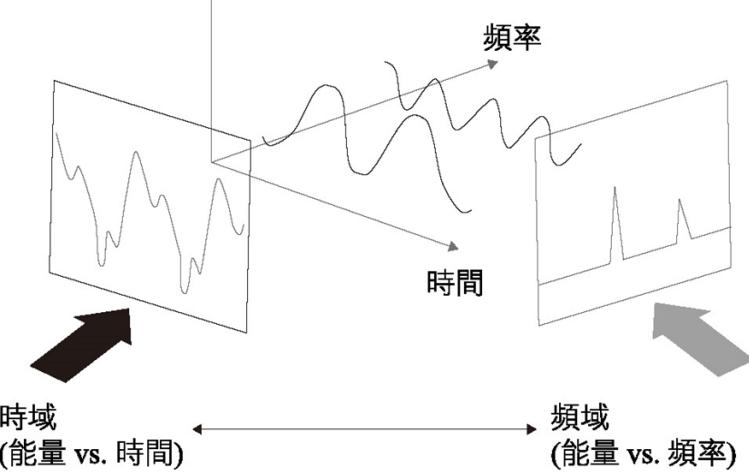


□ 高頻訊號特徵工程

- 若是與轉動有關的訊號（例如電流、振動與聲音）通常會是由多個不同週期成分的**弦波**組成，實際上轉動可呈現於**複數**空間中，然而一旦將複數空間投影到**實數**空間後，則會變為**弦波**。因此，可用「**頻域**」視角去剖析振動訊號，更能觀察出其隱含的豐富資訊

- 傅立葉轉換
- 短時距傅立葉轉換
- 小波轉換
- 時頻域轉換

能量



□ 短時距傅立葉轉換與小波轉換

- 比傅立葉轉換多了一個**時窗函數** (window function)，能隨著時間移動時窗函數以及其寬度的轉換出不同的頻率與時間解析度。
 - 當時窗函數寬度越大時，頻率的解析度較好而時間解析度較差。
 - 因此，時窗函數的大小會限制頻率的解析度，而小波轉換對不同頻率以解析度分析方式自適應地調整時窗函數的大小，能有效地解決這個問題。

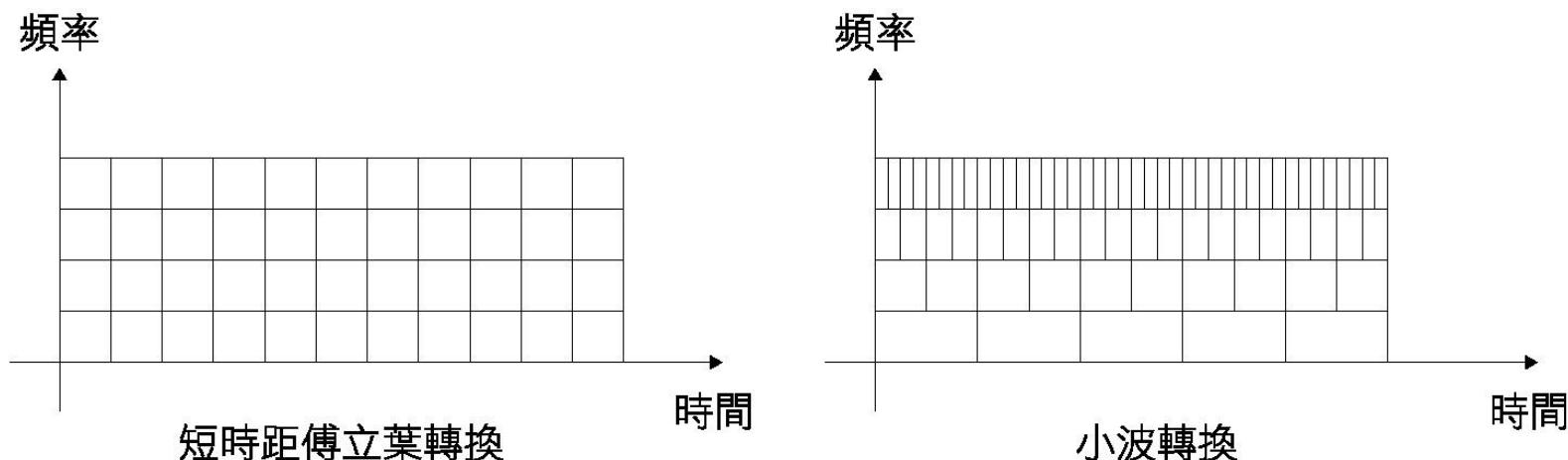


圖 12.11 短時距傅立葉轉換與小波轉換

高頻訊號特徵工程

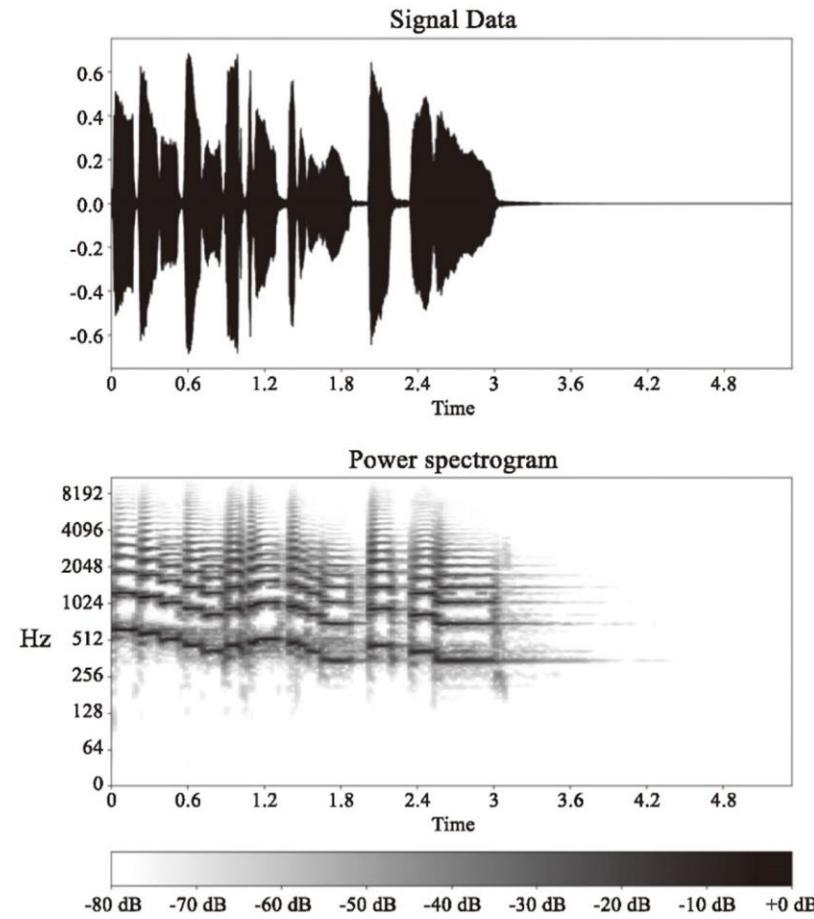
□ 短時距傅立葉轉換

- 「連續短時距傅立葉轉換」可將一個函數與時窗函數所選取的時間區間內積，進行傅立葉轉換，並隨著固定間隔的時窗函數平移，計算出一系列的傅立葉轉換的結果，展開後將得到具有時間、頻率及能量三維結果

— $c_{t,f} = \int_{-\infty}^{\infty} w(t - \tau) f(\tau) e^{-i2\pi f\tau} d\tau$

- 其中 $w(t)$ 是時窗函數，而 $f(\tau)$ 是待轉換的數據訊號。因此此二兩項的乘積 $w(t - \tau)f(\tau)$ 代表著利用時窗函數所選取區間的數據
- 時窗函數控制的是選取的時間段，可以是一個部份重疊且與中心越近權重越大的函數
- 對每個時窗進行傅立葉轉換如圖所示，其中 x 軸 y 軸對應到的是時間與頻率維度，而顏色越深則代表能量越強，從此圖中可觀察到不同頻率對應能量是如何在時間維度上改變。
- 特徵工程**

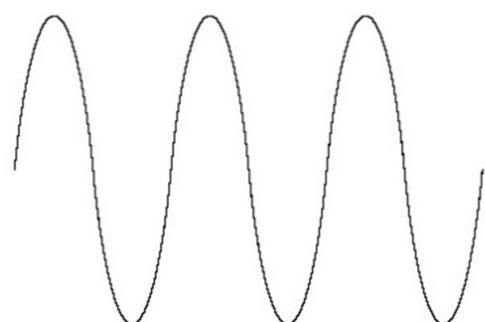
- 每個時間點挑選能量最高的頻率
- 每個頻率挑選能量最高的時間點
- 計算統計量如同前述**移動時窗特徵**的做法
- 視短時距傅立葉轉換的結果為一張圖片，進而以**圖片影像特徵工程**萃取特徵



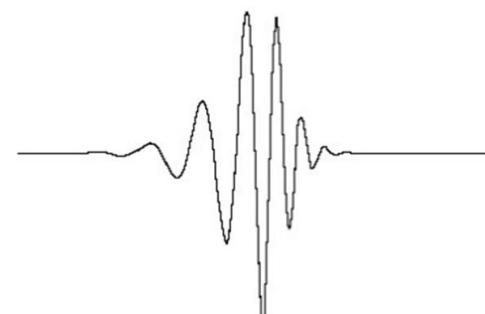
□ 小波轉換

- 「連續小波轉換」(continuous wavelet transform) 將一個函數分解成數個「小波函數」，而與短時距傅立葉轉換不同點在於小波轉換具有更能萃取局部與瞬時的特性。

— 分別為傅立葉轉換所使用的弦波與小波函數所使用的小波，前者更著重於長期穩定的狀態，後者更著重於短期瞬時的狀態。因此，在非穩定的數據訊號上，使用小波的效果更佳



弦波



小波 (db10)

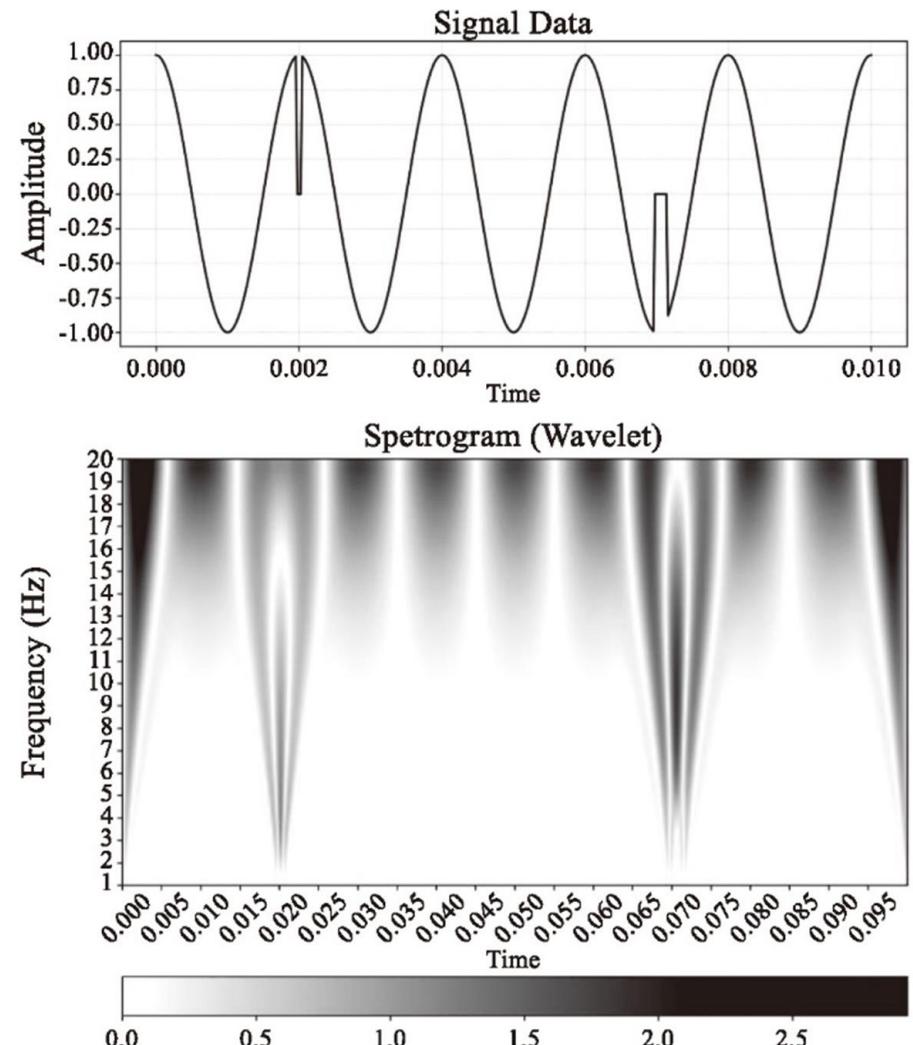
— 小波轉換能更適應性地調整時窗也在於其「小波母函數」(mother wavelet)能自由地調整平移與縮放的兩因子。連續小波轉換可表示如公式

$$c_{a,b} = \frac{1}{\sqrt{b}} \int_{-\infty}^{\infty} f(t) \psi\left(\frac{t-a}{b}\right) dt$$

— 其中 $\psi(t)$ 為小波母函數， a 與 b 分別為平移與縮放因子，進而透過此二因子產生多個「子波」(daughter wavelet)

□ 小波轉換

- 若我們以一個頻率由低至高的訊號為例，小波轉換能與短時距傅立葉轉換轉換出相似的圖像，而**小波轉換**更具備了**權衡頻率與時間解析度**的能力，且對於**瞬時變化**很大的訊號具備更好的萃取能力。



□ 圖片影像特徵工程

- 兩種主要的特徵工程方法

- **卷積神經網路** (convolution neural network, CNN)

- 對於空間（影像）特性數據 **卷積神經網路** 可直接以內建特徵萃取功能進行。
 - 換言之，在經過許多卷積與池化層層的計算之後，這些特徵的萃取最後被展平並連接到全連結層，此時除了以全連結網路進行預測或分類外，也可使用典型的方法（例如支持向量機或梯度提升機），接在後續做各式模型的預測或分類。

- **自編碼器** (autoencoder, AE)

- **自編碼器**是一個將圖像數據進行**編碼**與**解碼**的網路架構。它的思維是期望藉由這樣的網路結構實圖像數據的**壓縮**與**重建**，而**壓縮**實際上意謂著圖像**特徵萃取**，而**重建**則是由萃取後的**特徵生成數據**。

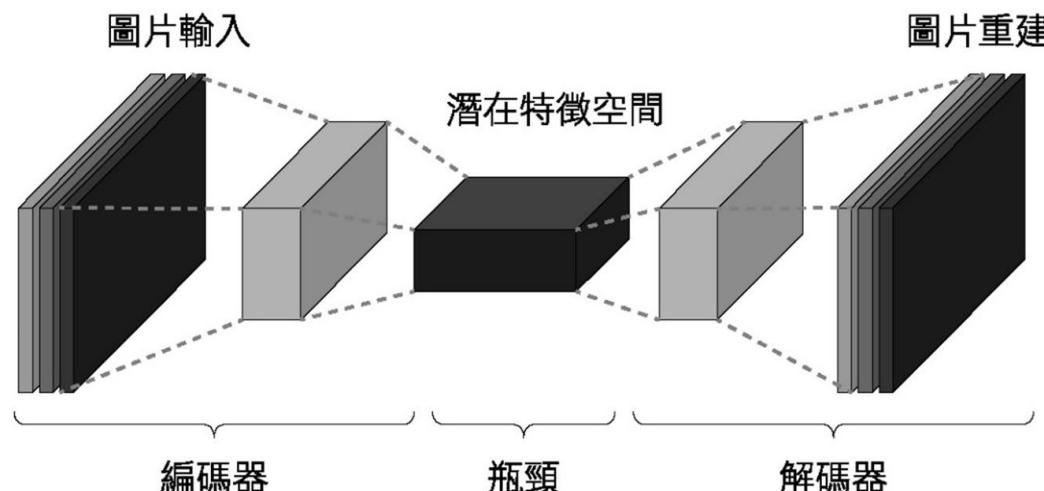


圖 12.15 自編碼器

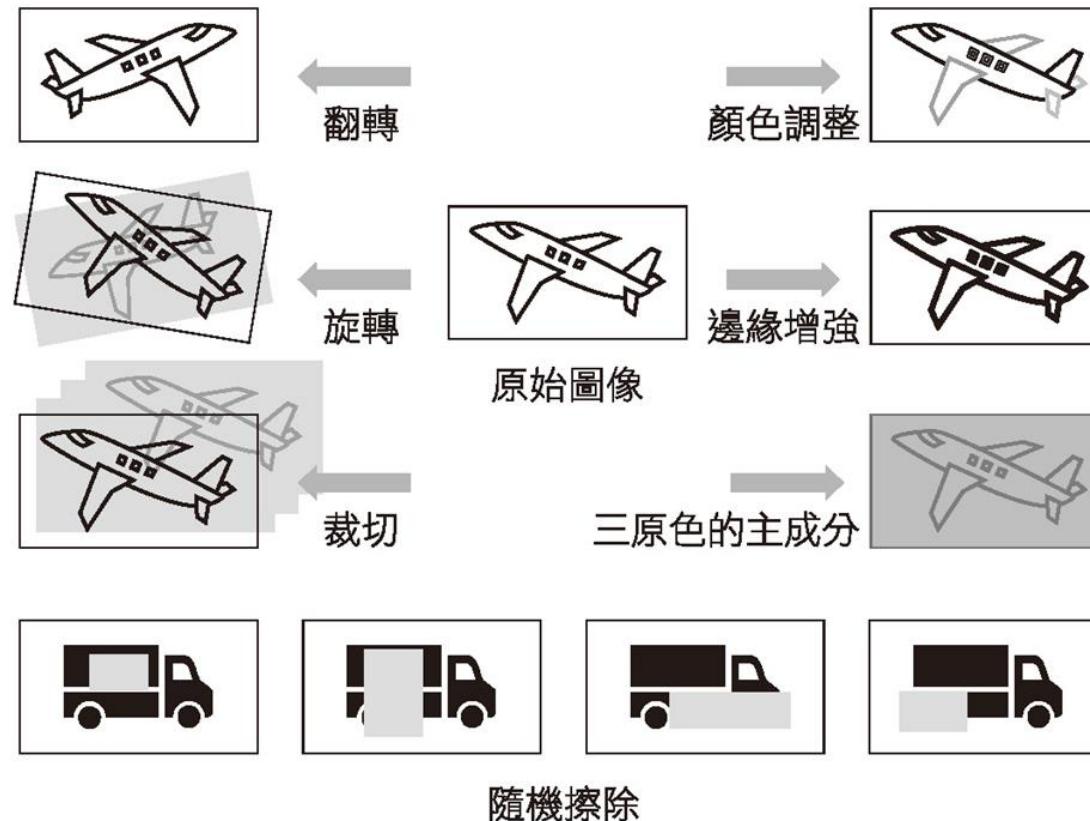
- 數據增強是利用具有領域知識的資訊將原始數據的觀測值進行擴增的過程，盡可能從現有的資訊中生成更豐富且多樣的資訊。
 - 一方面增加數據量，另一方面盡可能將數據多樣化與一般化，從而降低過度配適發生的可能

- 主要分為以下三大方法
 - 典型技巧 (classic tricks)
 - 自編碼器 (autoencoder)
 - 生成對抗網路 (generative adversarial network, GAN)

□ 數據增強- 典型技巧 (以圖像生成為例)

- 幾何方法 (geometric methods)
 - 翻轉 (flipping) 、旋轉 (rotating) 、位移 (displacement) 、縮放 (scaling) 與裁切 (cropping)
- 光度方法 (photometric methods)
 - 包括了顏色調整 (color jittering) 、邊緣增強 (edge enhancement) 以及三原色的主成分 (fancy PCA)
 - 顏色調整：對圖像的亮度(brightness)、飽和度(saturation)、對比度(contrast ratio)進行調整
 - 邊緣增強：對圖像中物件的邊緣強化(計算圖像相鄰像素的梯度)
 - 三原色的主成份：對圖像的RGB三原色進行主成分拆解後，將特徵值隨機擾動再與特徵向量相乘生成新圖像。
- 隨機擦除 (random erasing)
 - 在圖像中隨機選擇一個方框，並以隨機的值來替換圖像原有的數值。

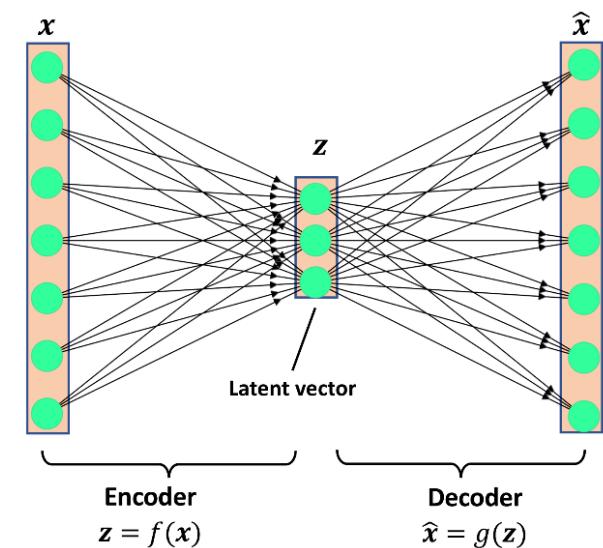
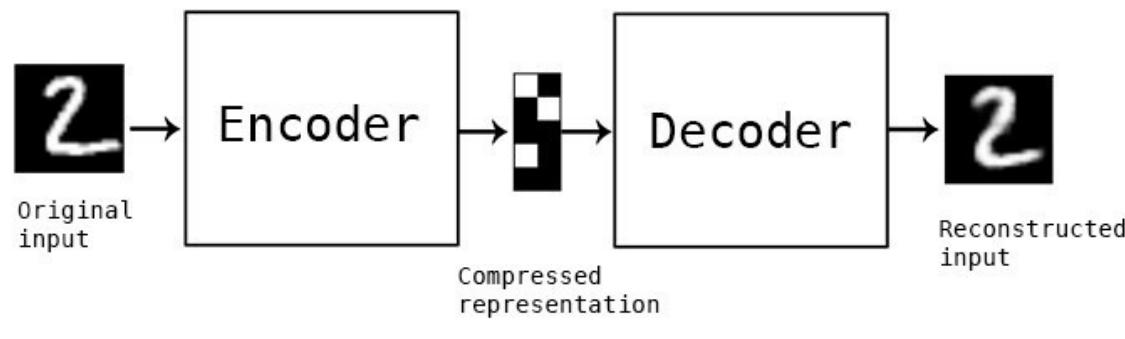
□ 數據增強- 典型技巧 (以圖像生成為例)



- 典型技巧在實作需注意，例如幾何方法在**翻轉**技巧的使用上，具有「**手性**」(chiral)的圖像被翻轉後可能會產生較不合理特徵(原物件的左右不對稱)，例如文字序列、男女鈕扣不同邊等。在**旋轉**、**位移**技巧的使用上，若在圖像背景是未知的情形下，經過這些操作後將造成部分圖像區塊產生遺漏(例如圖的飛機經旋轉後左下與右上的像素會產生遺漏)，另需填補這些遺漏值。這說明數據增強的目的是增加合理的資訊量，而合理的前提是建立在領域知識

□ 自編碼器(Unsupervised learning)

- Autoencoder (AE) (Nonlinear feature)
- 是一個將圖像數據進行編碼與解碼的架構，其主要的功能是對數據進行壓縮與重建。
- 本節主要介紹以下兩種自編碼器
 - 變分自編碼器 (variational autoencoder, VAE)
 - 條件變分自編碼器 (conditional variational autoencoder, CVAE)

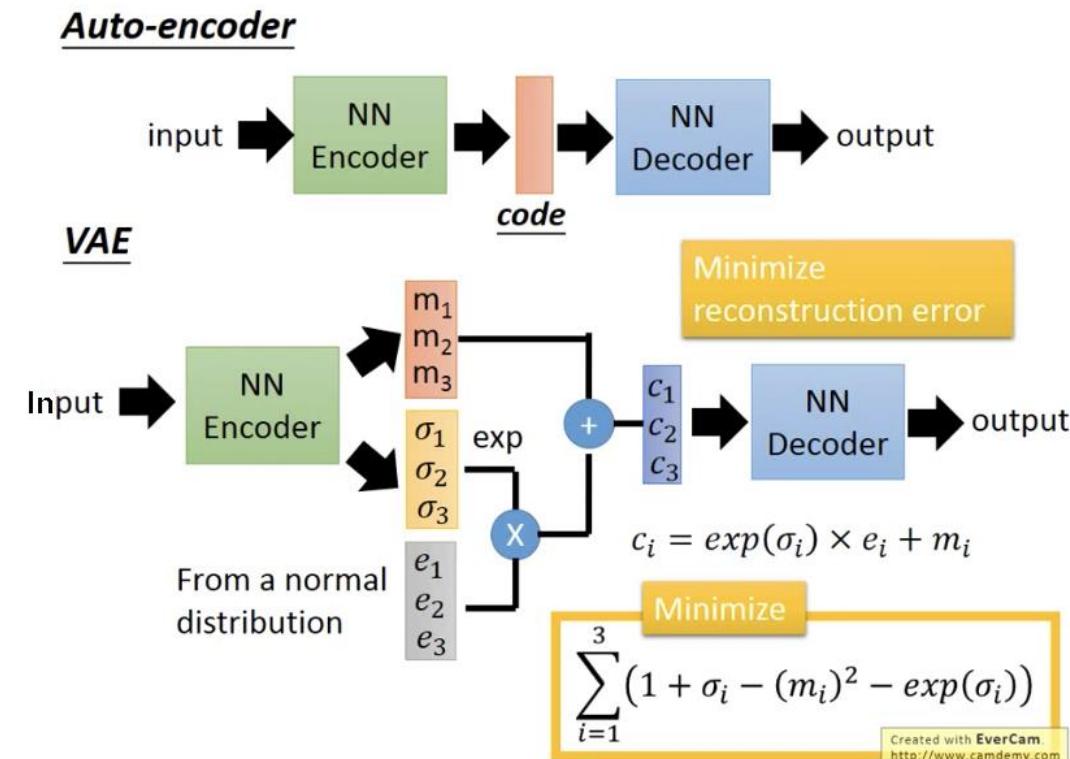
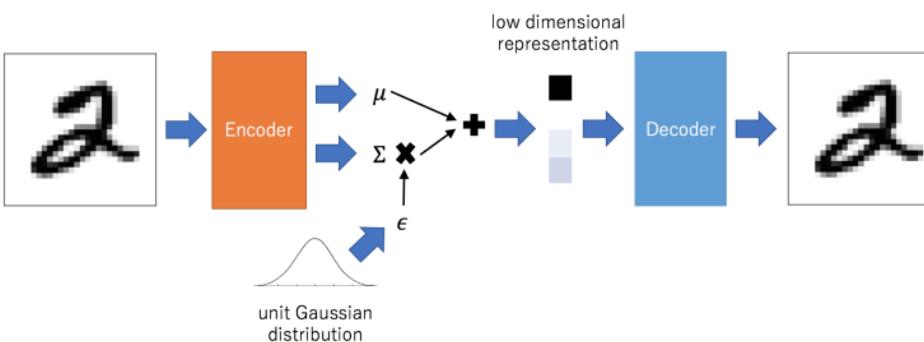


<https://medium.com/程式工作紡/autoencoder---認識與理解-725854ab25e8>

□ 變分自編碼器(Variational Autoencoder, VAE)

- 與自編碼器不同的點在於編碼器與解碼器間的隱藏層，其隱藏層是去估計這些壓縮後的**隱變數** (latent variable)，且限制他們的分配遵從高斯分布，從而估計分布參數**平均數**與**變異數**，因此控制想要生成的圖片。
- 因此，我們能對這些分配進行隨機抽樣（也可以視為加上噪音），生成不同的輸出，基於給定的數據下生成與該數據相似的樣本。

- 1. 先輸出兩個向量：平均數與變異數
- 2. 用高斯分布產生第三個向量
- 3. 把第二個向量做指數，之後跟第三個向量相乘後，把它跟第一個向量相加，即成為中間層的隱變數



<https://medium.com/程式工作紡/autoencoder---認識與理解-725854ab25e8>

□ 變分自編碼器

- 為了確保模型的生成能力，變分自編碼器在訓練上的另一個目標是讓隱變數與標準常態分配越近越好，以防止變異數退化為零，而完成上述目標最直接的方法即是將重建目標加上生成目標，也就是隱變數與標準常態分配的**KL散度**。

$$\text{— } \text{KL}\left(N(Z_\mu, Z_{\sigma^2}) \parallel N(0,1)\right) = l_{\mu,\sigma^2} = \frac{1}{2} \sum_{i=1}^d (\mu_i^2 + \sigma_i^2 - \log \sigma_i^2 - 1)$$

- 實際上，此損失函數又可被拆解為平均數 μ 與變異數 σ^2 各自的損失公式

$$\text{— } l_{\mu,\sigma^2} = l_\mu + l_{\sigma^2}$$

$$\text{— } l_\mu = \frac{1}{2} \sum_{i=1}^d \mu_i^2$$

$$\text{— } l_{\sigma^2} = \frac{1}{2} \sum_{i=1}^d (\sigma_i^2 - \log \sigma_i^2 - 1)$$

□ 變分自編碼器**本質**

- VAE為自動編碼器的一個變形(variant)或是廣義的版本
- 基於**貝氏定理**使得我們能隨意擾動代表核心特徵的隱變數(或可以理解為對隱變數加入噪音)，從而讓解碼器生成不同的圖像。
 - 例如訓練圖像為貓的數據集，我們可以想像這些經編碼器所萃取過的隱變數可能是貓的眼睛、耳朵、嘴巴、手、腳以及毛的色澤、質地等，藉由擾動這些隱變數後，最後便可經由解碼器生成出需多不同特徵組合的貓圖像。
- 模型的設計與訓練
 - 一方面期望解碼器能生成與原有數據有相似的分配特性
 - 另一方面則期望由編碼器產生出的隱變數為萃取的特徵且具有被擾動的能力
 - **特殊化與一般化對抗的過程**(為廣義的偏誤與變異權衡)
 - 因此可視損失函數中的**KL散度**為一個正則項(*regularizer*)
 - 當模型在初始訓練時，輸入與輸出誤差遠大於KL誤差，使得模型訓練會優先降低輸入與輸出的誤差，增強模型萃取特徵與還原圖像的能力
 - 當模型訓練一段時間後(此時解碼器被訓練得還不錯)，KL誤差將大於輸入與輸出誤差，使得模型訓練接著降低KL誤差，提升模型擾動與生成的能力

□ 變分自編碼器**本質**

- 簡言之，輸入與輸出誤差的目標是還原(重建)圖像，而KL誤差的目標則是擾動特徵，因此模型的訓練實際上是兩個目標相互對立與抗衡的過程，這與前述介紹的**生成對抗網路**的思維不謀而合。

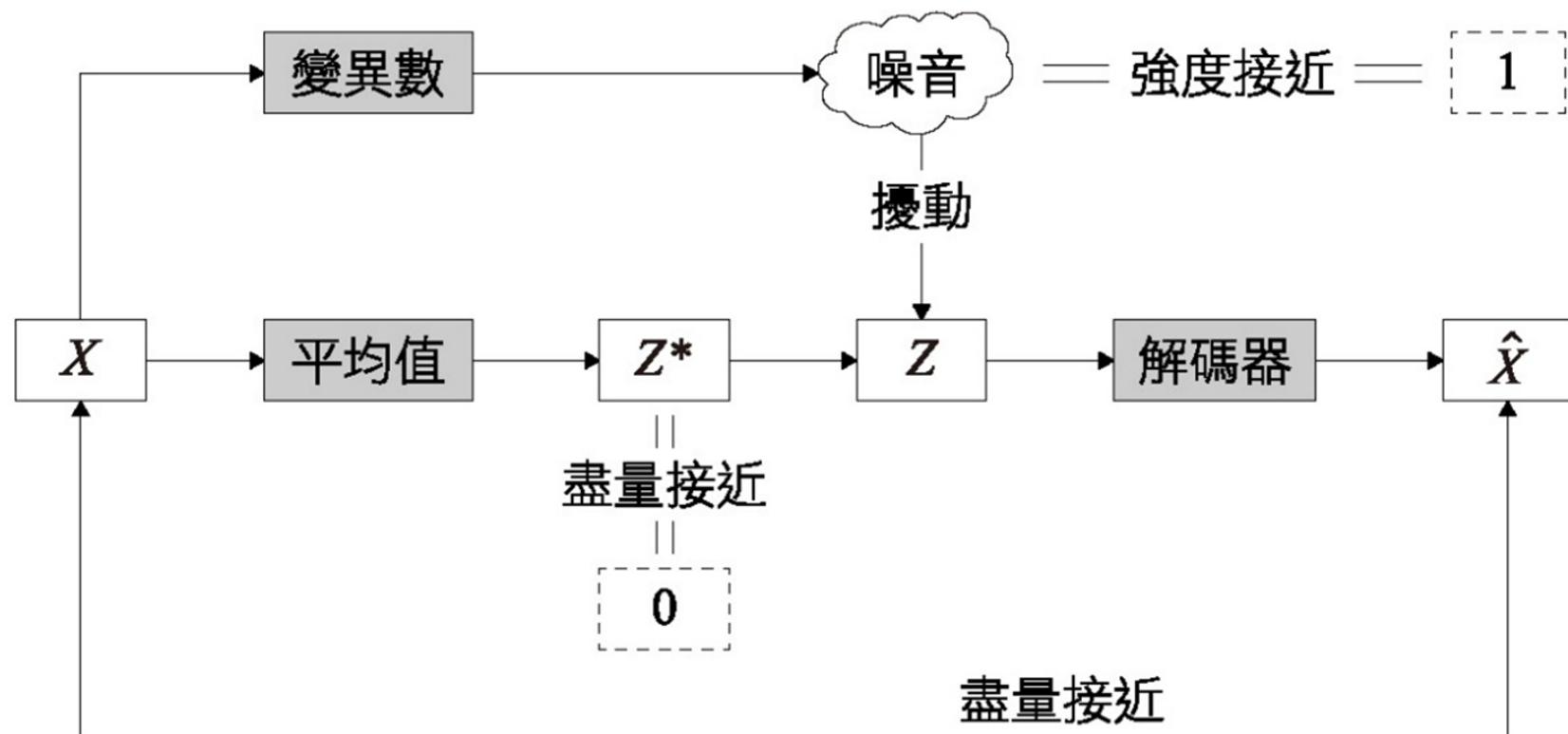


圖 12.17 變分自編碼器

□ 條件變分自編碼器(CVAE)

- 當有額外的數據與標籤時，將這些額外資訊考慮至模型中輔助生成樣本
 - 應用的價值在於可以實現控制代表額外資訊的特徵來生成特定的圖像
 - 其中Y所代表的即是額外的特徵，屬於半監督式學習的方法。

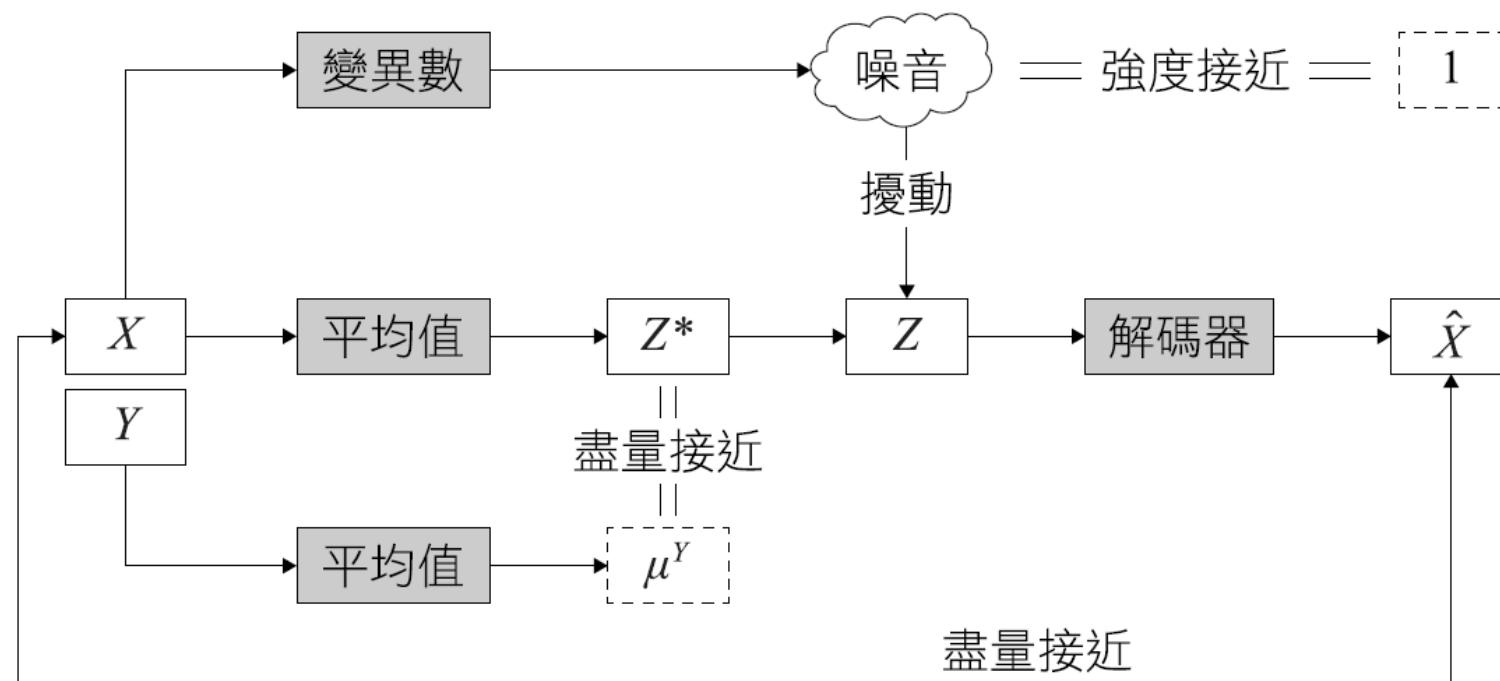
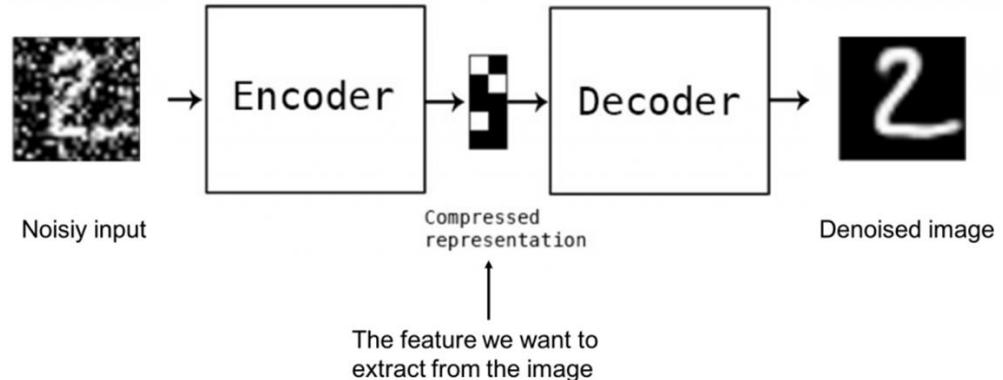


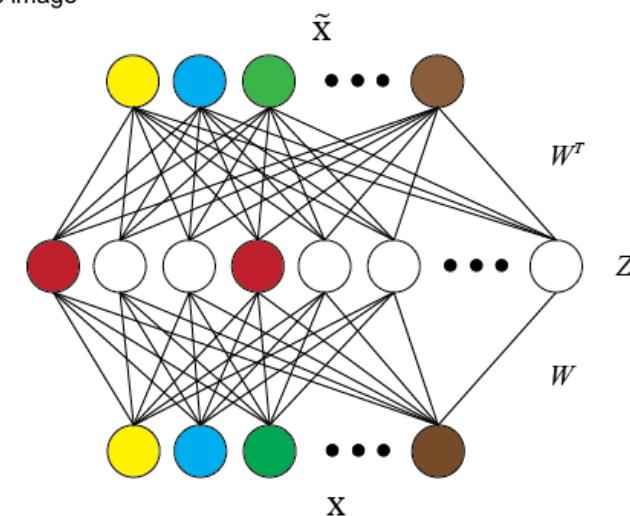
圖 12.18 條件變分自編碼器

□ 自編碼器

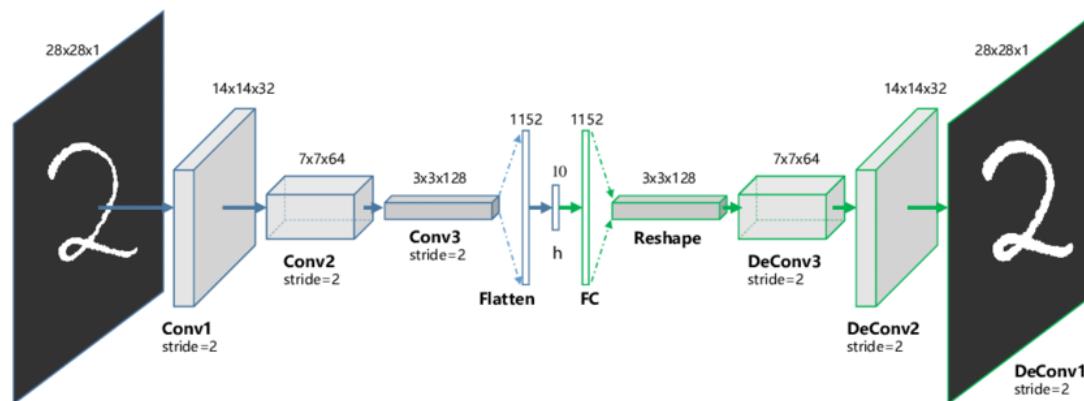
- Denoising AE (DAE)



- Sparse AE (SAE) with L1 regularization



- Convolutional Autoencoder (CAE)



<https://medium.com/程式工作紡/autoencoder---認識與理解-725854ab25e8>

□ 生成對抗網路(GAN)

- 架構是由一個生成器 (generator) 與判別器 (discriminator) 所建構
- 介紹以下三種方法
 - 非監督式學習的混合變分自編碼器與生成對抗網路 (VAE-GAN)
 - 半監督式學習的條件生成對抗網路 (conditional GAN, CGAN)
 - 生成對抗填補網路 (generative adversarial imputation network)

□ 混合變分自編碼器與生成對抗網路(VAE-GAN)

- 變分自編碼器與生成對抗網路，前者是區域單一樣本的視角，後者是全域整體數據的視角。因此，將兩種模型整合後便能充分地含有各自的優勢，而整合的架構便是混合變分自編碼器與生成對抗網路。
- 非監督式學習

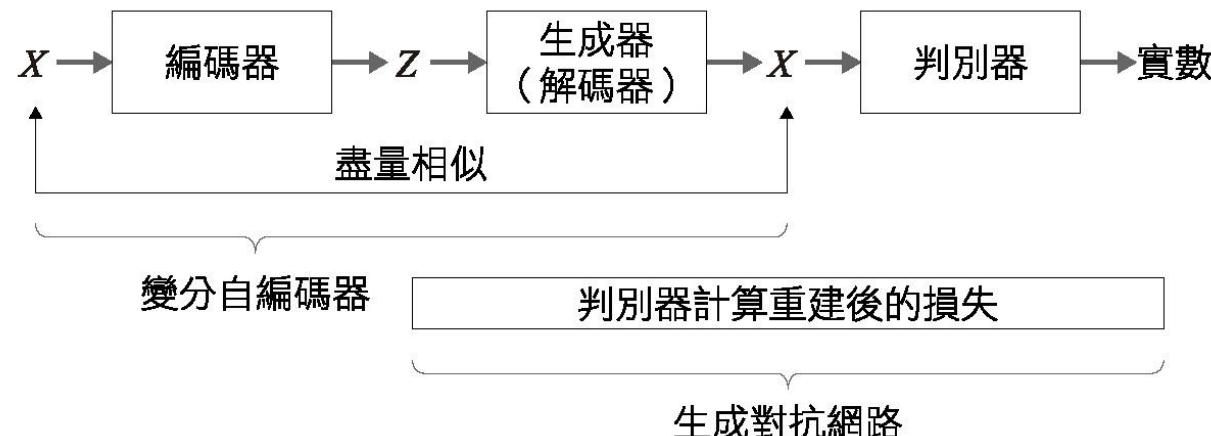


圖 12.19 混合變分自編碼器與生成對抗網路

□ 條件生成對抗網路(CGAN)

- 使得數據增強能更廣泛的應用於不同情境，能妥善地依據**領域知識**融入額外的資訊（數據與標籤）。如前述介紹了條件變分自編碼器，兩者的模型設計均使用相同的思維，它於**生成器與判別加入額外的資訊Y**
- 半監督式學習

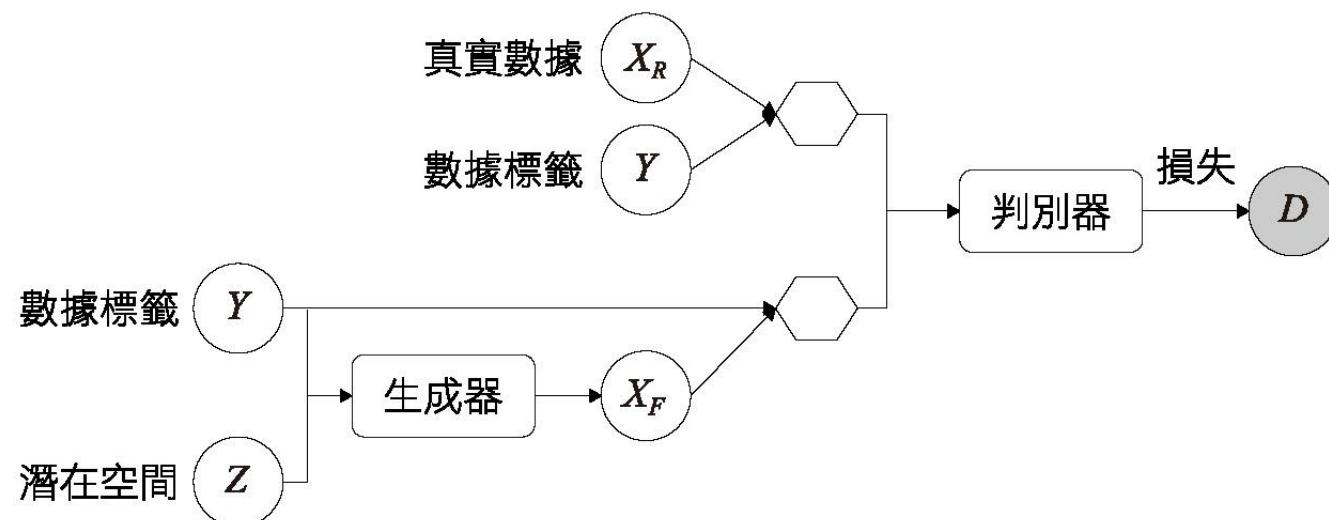


圖 12.20 條件生成對抗網路

□ 條件生成對抗網路(CGAN)

- 回顧生成對抗網路的目標函數 V 可表示如下：

$$\min_G \max_D V(D, G) = \underbrace{\mathbb{E}_{x \sim p_{data}} [\log D(x)]}_{\text{正確判別真實數據}} + \underbrace{\mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))]}_{\text{正確判別假數據}}$$

- 條件生成對抗網路的目標函數 V_c 可表示如下：

$$\min_G \max_D V_c(D, G) = \underbrace{\mathbb{E}_{x \sim p_{data}} [\log D(x|y)]}_{\text{正確判別真實數據}} + \underbrace{\mathbb{E}_{z \sim p_z} [\log(1 - D(G(z|y)))]}_{\text{正確判別假數據}}$$

- 基於**條件生成判別網路**的基礎架構下，發展出了利用判別器同時預測分類類別以輔助模型訓練的模型
 - 半監督式學習生成對抗網路 (Semi-Supervised GAN)
 - 資訊最大化生成對抗網路 (InfoGAN)
 - 輔助分類器生成對抗網路 (AC-GAN)

□ 生成對抗填補網路 (generative adversarial imputation network)

- 是一個以生成數據的方式對數據遺漏值進行填補的方法 (Yoon et al., 2018)
- 首先，生成器依據**數據矩陣** (data matrix)、**隨機矩陣** (random matrix) 以及表示遺漏的**表示矩陣** (mask matrix) 產生一個完整的填補後數據矩陣 (imputed matrix)
- 接著判別器再以此填補後的矩陣以及由**表示矩陣**產生的**暗示矩陣** (hint matrix) 來判別真偽。

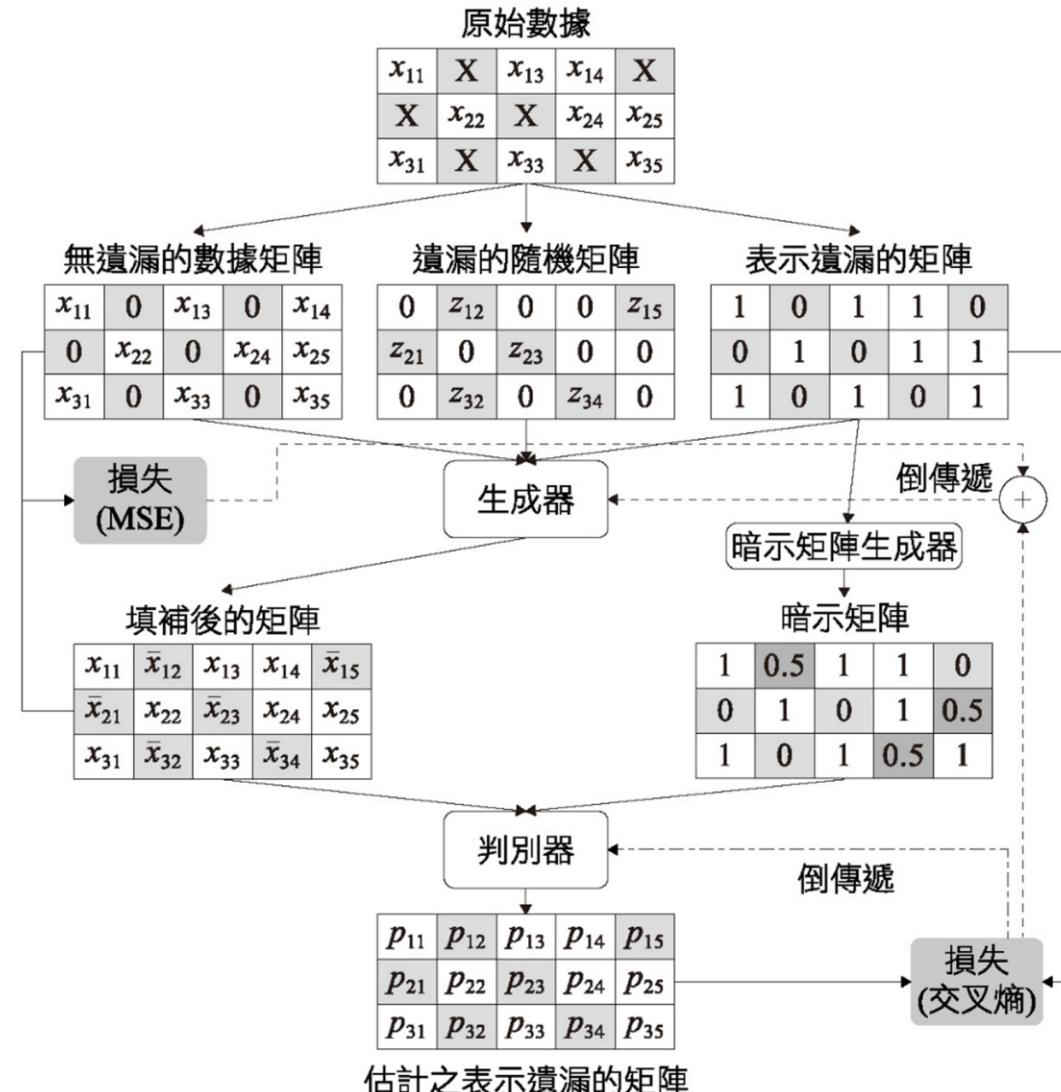


圖 12.21 生成對抗填補網路

□ 數據不平衡

- 在製造現場的穩定產線上，大多數產品品質、機台狀態以及產品良率的分布相當偏頗，在製程能力好的情況下，不良或異常狀況的發生均屬於稀少事件

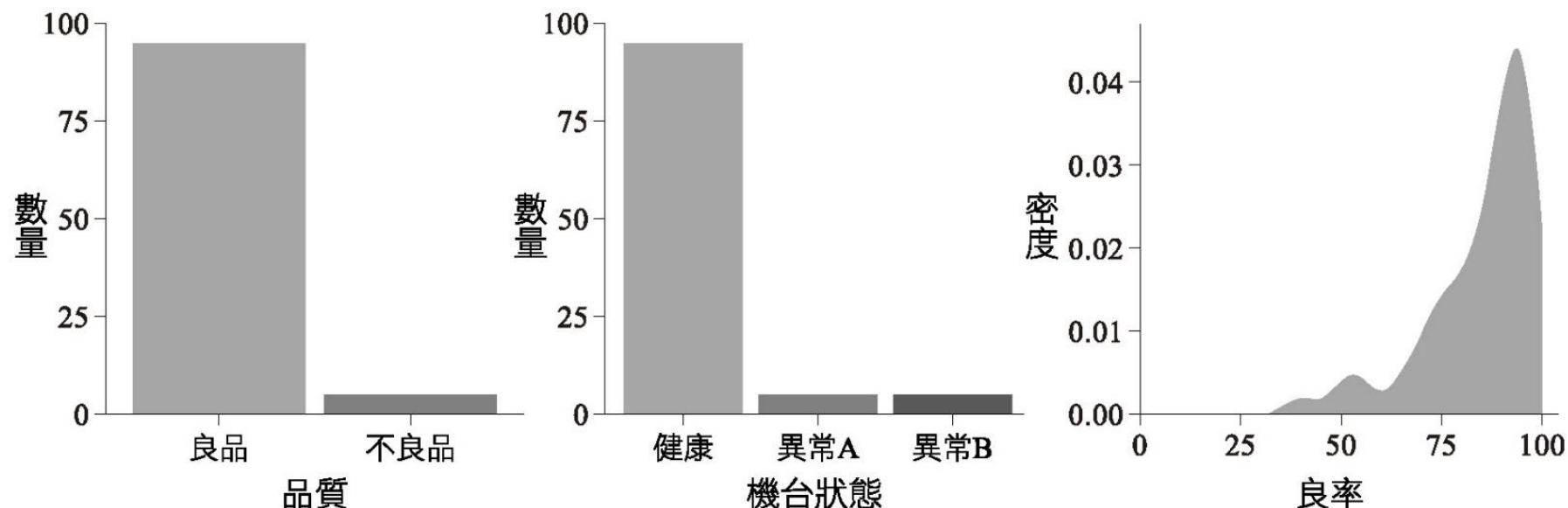


圖 12.22 數據不平衡

□ 數據不平衡

- 在這種情形下，由於多數數據科學演算法假設
 - 數據均勻分布
 - 每個樣本權重一致
- 因而**傾向往多數群預測**，導致少數群的預測效果不佳。

- 然而在實務問題中，往往是這些少數群的樣本造成產線良率不穩、利潤損失等情況，因此希望透過數據科學來協助**故障排除**（troubleshooting）或找尋**根本原因**（root cause）。

- 從學理觀之，倘若數據分布越是極端，如**不平衡率**（imbalance ratio, IR）過大、絕對值的偏態越大或峰態越大，數據不平衡對預測效果造成的影響越大。

- 本節將說明如何面對數據不平衡的問題，並有效預測出**少數群**。

數據不平衡

Lot ID	X1	...	X100	Inspection
Lot01				PASS
Lot02				PASS
Lot03				PASS
Lot04				PASS
Lot05				PASS
Lot06				PASS
Lot07				FAIL
Lot08				PASS
Lot09				PASS
Lot10				PASS
Lot11				PASS
Lot12				PASS

預測Inspection的結果

- 由於只有1筆FAIL
- 預測模型全部都猜PASS
 - 不需要分析變數X1~X100
 - 準確度可達 $11/12 = 91.7\%$

如果是1000批貨

- 995良品
- 5不良品
- 則預測準確度..99.5%!!??

□ 辨別數據不平衡

- 留意模型是否輸入數據不平衡所設置的陷阱。
- 數據視覺化

□ 不論在建立模型之前或之後，**數據視覺化**都相當重要

- 模型建立前：收集到數據後，對數據應變數的視覺化，可初步瞭解數據不平衡的程度。
- 模型建立後：比較實際與預測值，可剖析模型對數據不平衡處理的效果。

表 12.2 數據不平衡於分類問題

	預測為良品 $\hat{y} = 0$	預測為不良品 $\hat{y} = 1$
實際為良品 $y = 0$	995	0
實際為不良品 $y = 1$	5	0

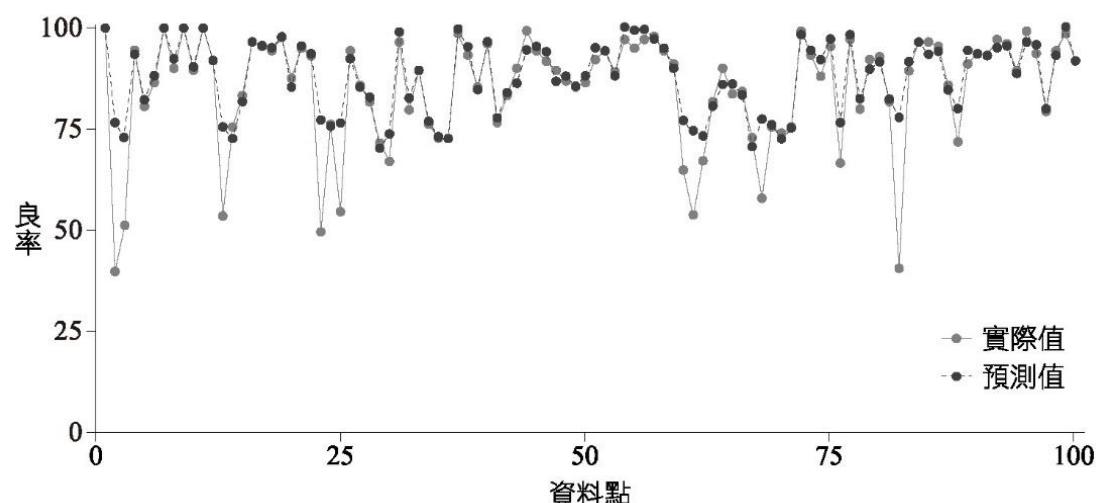


圖 12.23 數據不平衡於迴歸問題

□ 如何處理數據不平衡？

- 數據層面：對數據採取特定的**抽樣策略**使得數據平衡
- 演算法層面：對現有的**演算法**進行改進，直接考慮少數群的重要性。

□ 抽樣方法

- 對數據進行抽樣的方法，是**直接改變數據分布**，使得多數群與少數群的樣本數達到平衡，某個角度也是對數據的進行預處理
- 主要分為
 - **上抽樣 (over-sampling)**：對少數群進行取後放回的重複隨機抽樣，以增加少數群的樣本數。
 - **下抽樣 (under-sampling)**：對多數群進行隨機抽樣，以減少多數群的樣本數。
 - **混合抽樣 (over- and under-sampling)**：分別對多數群與少數群進行上述兩種隨機抽樣方法。
 - **數據生成 (data generation)**：利用少數群既有的樣本生成有變異的新樣本，以增加少數群的樣本數，屬於上抽樣的延伸發展。
 - **數據清理 (data cleaning)**：移除多數群與少數群重疊的樣本，以減少數據原有的噪音，屬於下抽樣的延伸發展。
 - **具資訊抽樣 (informed under-sampling)**：基於下抽樣會造成資訊上的損失，此方法進行多次下抽樣，並運用集成學習的概念進行建模，屬於下抽樣的延伸發展。

- 上抽樣與下抽樣的概念如下圖所示

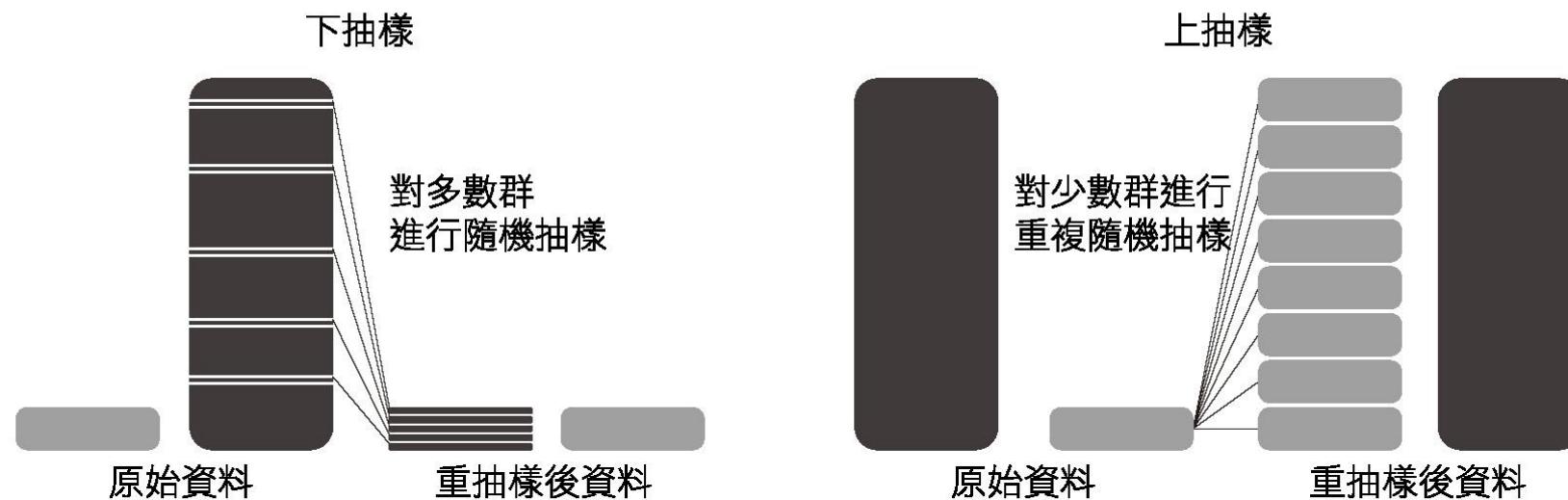


圖 12.24 上抽樣與下抽樣

□ 改善數據不平衡的抽樣方法

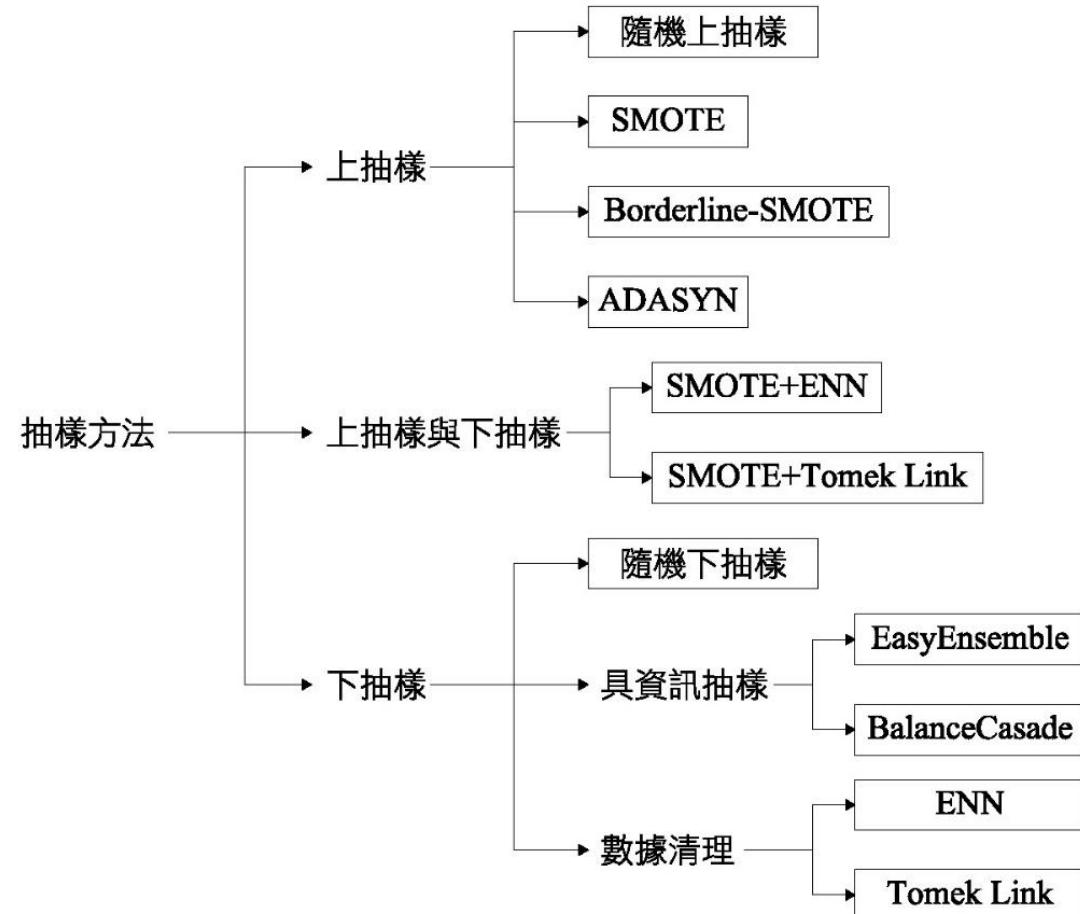


圖 12.25 處理數據不平衡的抽樣方法

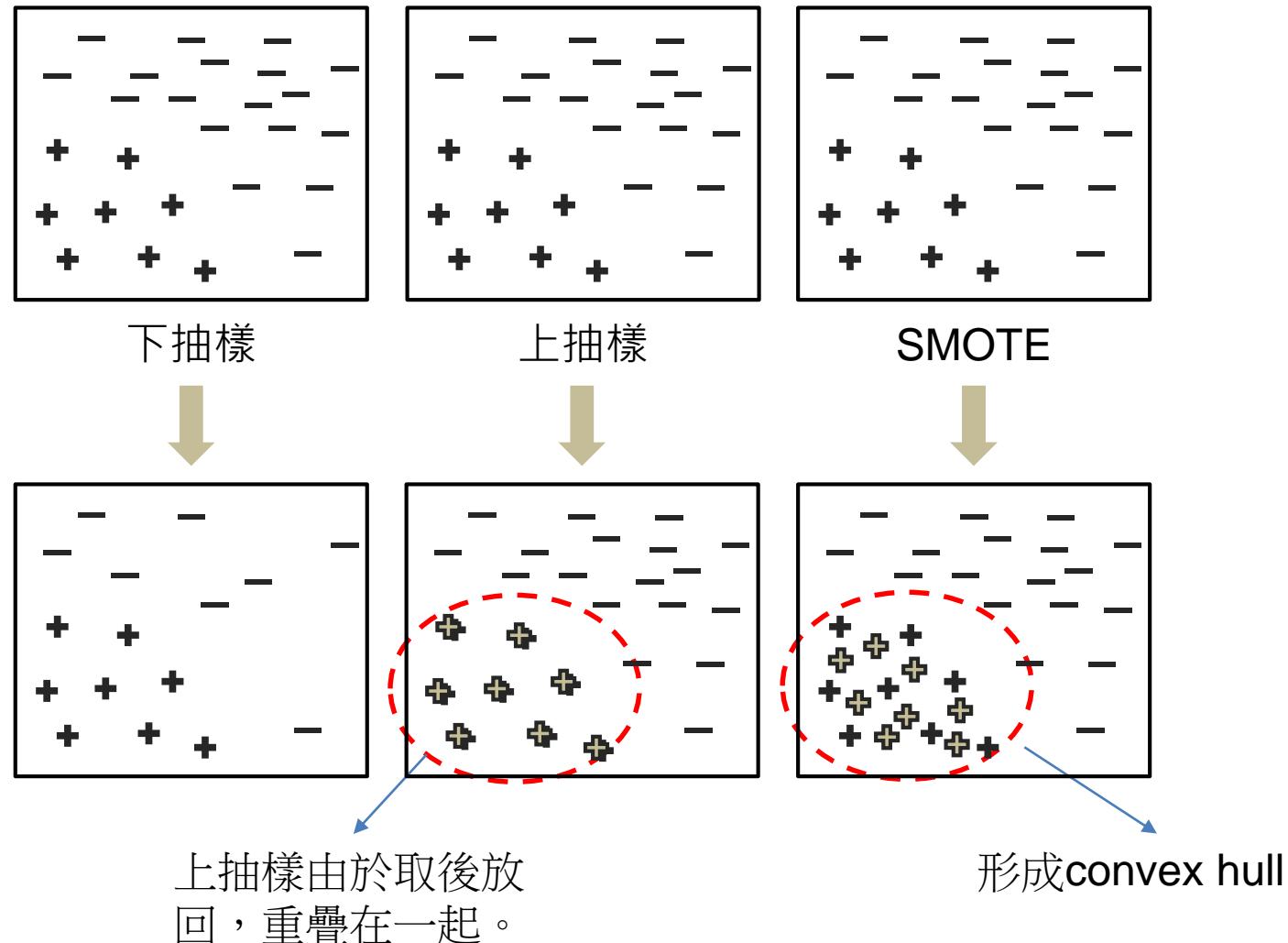
□ 合成少數群上抽樣技術 (SMOTE)

- **合成少數群上抽樣技術** (synthetic minority over-sampling technique, SMOTE) (Chawla et al., 2002) 是**數據生成**的方法，其假設少數群的樣本間在特徵空間上具備高相似度，因而對少數群的樣本間進行隨機線性內插，生成新的少數群樣本。

- SMOTE 的流程如下

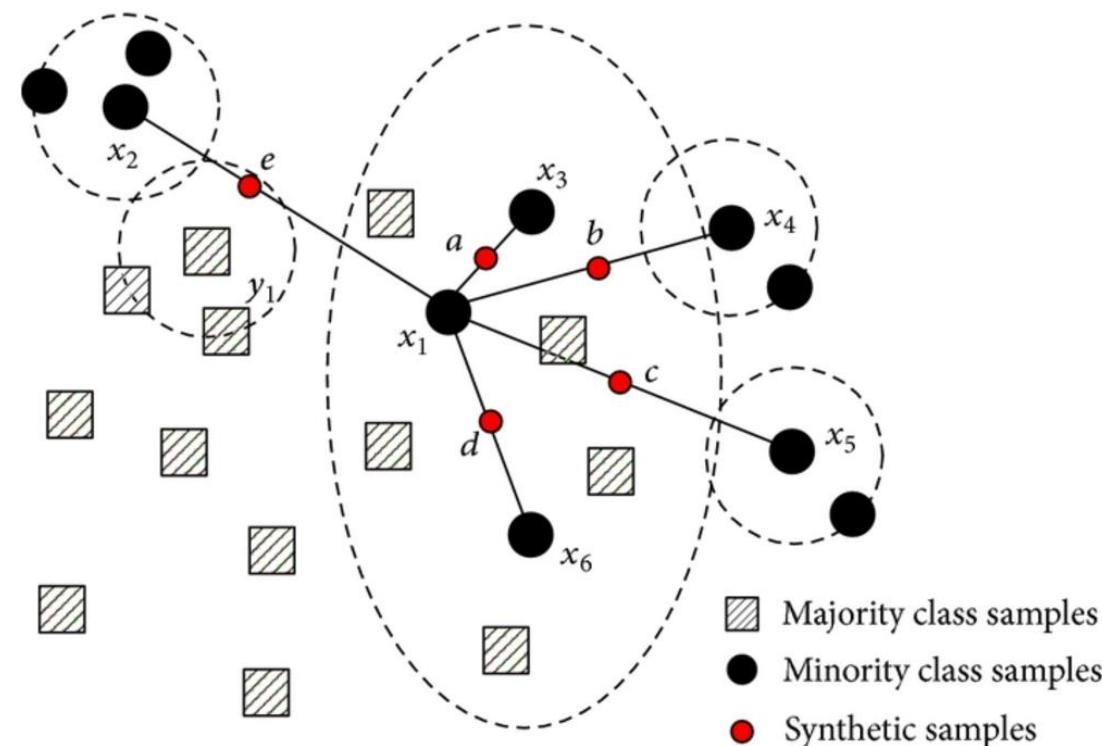
1. 設定每個少數群樣本要生成的樣本數 N 。
2. 針對每個少數群的樣本 x_i 計算並找出 k 個最鄰近的少數群樣本。
3. 從 k 個最鄰近樣本中隨機挑選出一個樣本 $x_{knn(i)}$ 。
4. 透過內插法生成新的樣本 x_{new} ，如以下公式所示，其中 γ 為介於零到一的隨機變數。
$$x_{new} = x_i + \gamma(x_{knn(i)} - x_i), \quad \gamma \sim U(0,1)$$

5. 跳到2.直到重複 N 次。

上抽樣、下抽樣與SMOTE生成結果的比較

□ 自適應合成抽樣 (ADASYN)

- **自適應合成抽樣 (adaptive synthetic sampling approach, ADASYN)** (He et al., 2008) 根據少數群與多數群樣本特徵分布的情形，計算出每個少數群樣本需生成的樣本數，使得分布於多數群密集的少數群樣本增加更多樣本，而分布於多數群稀疏的樣本增加較少樣本。



Bhattacharyya, I., 2018. SMOTE and ADASYN (Handling Imbalanced Data Set).

<https://medium.com/coinmonks/smote-and-adasyn-handling-imbalanced-data-set-34f5223e167>

□ 自適應合成抽樣 (ADASYN)

- ADASYN 流程如下

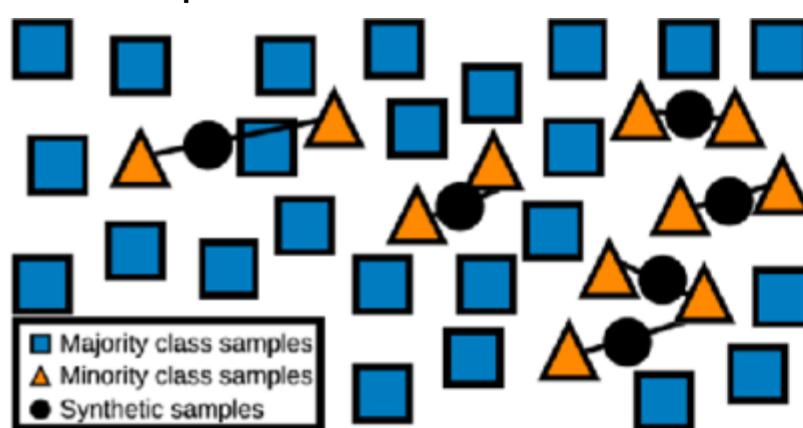
1. 定義少數群樣本數為 m_s ，多數群樣本數為 m_l ，期望平衡的比例為 β ，並設定少數群樣本要生成的樣本總數 $G = (m_l - m_s) \times \beta$ 。
 2. 針對每個少數群的樣本 x_i 計算並找出 k 個最鄰近的樣本，接著計算多數群佔的比例 $r_i = l_i/k_i$ ，其中 k_i 為鄰近樣本數， l_i 為鄰近多數群的樣本數。
 3. 將所有 r_i 正規化，使得所有正規化後的 $\hat{r}_i = r_i / \sum r_i$ 加總為1。
 4. 計算出每個少數群樣本需生成的樣本數 $g_i = r_i \times G$ 。
 5. 透過SMOTE生成新樣本。
-
- ADASYN所生成的少數群樣本具備較多的資訊量，然而卻對離群值較為敏感，當**離群值**與**類別重疊**的問題發生時，反而容易生成更多具有噪音的資訊。

□ Borderline-SMOTE

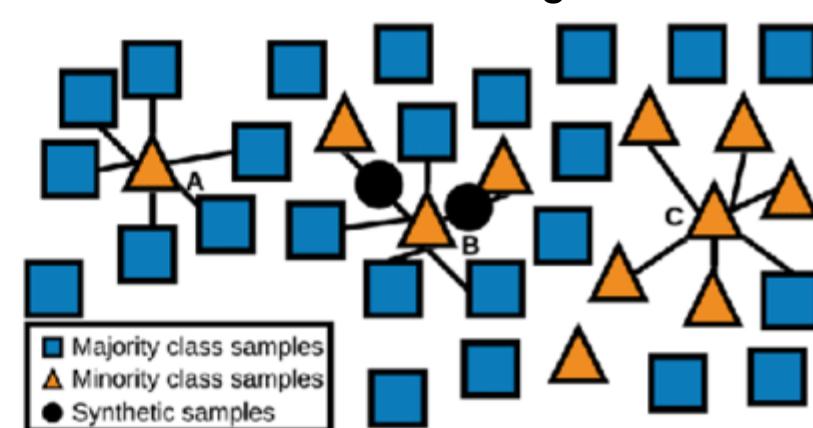
- Borderline-SMOTE (Han et al., 2005) 與ADASYN 概念十分接近，同樣是根據少數群與多數群樣本的特徵分布生成樣本，然而此方法對離群值進一步的修正。當鄰近的樣本全部皆為多數群時，視該樣本為離群值不生成樣本，而當少數群鄰近的樣本中多數群佔的比例高於一半時，則該樣本才符合生成樣本的條件。
- Borderline-SMOTE的流程如下
 1. 針對每個少數群的樣本 x_i 計算並找出 k 個最鄰近的樣本，接著計算出鄰近多數群的樣本數 l_i 。
 2. 若 $l_i = k$ ，則表示鄰近皆為多數群，可將該少數群樣本視為離群值或噪音；若 $k/2 \leq l_i < k$ ，則表示該少數群樣本落於邊界上容易被錯分，將其放入危險的集合； $0 \leq l_i < k/2$ ，則表示該少數群樣本在安全的地帶。
 3. 針對危險區的少數群樣本透過SMOTE生成新樣本。
- Borderline-SMOTE區分了少數群樣本的特性（離群值、危險樣本、安全樣本），因此對於離群值有了更好的處理，然而如何給定區分特性的界線是此方法潛在的問題，這隱含了多個參數需要設定。

□ Borderline-SMOTE

- For each minority class sample (orange triangle), Borderline-Smote finds its K-nearest neighbours among the whole samples.
 - Case A: If all of them are majority class samples (case A), the sample is considered **noise**.
 - Case B: If there are more majority class neighbours than minority class ones (case B), the sample is considered easily misclassified and put into a **danger** set. Only the samples in the danger set are the borderline ones and they are used by SMOTE to generate new instances (black circles).
 - Case C: if there are more minority class neighbours than majority class ones (case C), the sample is considered **safe**.
- Comparison between SMOTE and BORDERLINE-SMOTE using K=6



(a) SMOTE example



(b) BORDERLINE-SMOTE example

Bernardo, A., and Della Valle, E., 2021. VFC-SMOTE: very fast continuous synthetic minority oversampling for evolving data streams. Data Mining and Knowledge Discovery, 35, 2679–2713.

□ 演算法改良

- 相較於數據層面，演算法層面的改進是針對少數群的**錯分代價與數據權重**進行調整，部分方法也結合了數據層面的預處理手法。
- 以下我們分為兩個層面進行討論
 - 「**代價敏感學習**」 (cost sensitive learning)
 - 「**一分類學習**」 (one-class learning)

□ 代價敏感學習

- 代價敏感學習 (Elkan, 2001) 考慮在錯分樣本時，對錯分的這些少數群樣本給予不同的懲罰代價。

表 12.3 代價矩陣

	預測為良品 $\hat{y} = 0$	預測為不良品 $\hat{y} = 1$
實際為良品 $y = 0$	正確分類 TN $C_{(0,0)} = C_{TN} = 0$	誤判為不良品 FP $C_{(1,0)} = C_{FP} = 1$
實際為不良品 $y = 1$	誤判為良品 FN $C_{(0,1)} = C_{FN} = 10$	正確分類 TP $C_{(1,1)} = C_{TP} = 0$

□ 代價敏感學習

- 代價敏感學習的方法主要有三大類

— **元學習** (meta-learning) : 此類方法是不改變演算法本身，而是透過懲罰代價在建模前對數據的進行預處理，或是建模後預測結果的進行調整。

➤ 抽樣法：此方法與前述數據層面的概念是一致的，差異僅在於平衡的比例是由定義出的代價作為參考依據。

➤ 閾值調整法：二元分類模型的閾值 p 一般為 0.5，當預測值(後驗機率 $P(j|x)$)大於 0.5 時預測為 1，小於 0.5 時預測為 0。此方法是對後驗機率的閾值進行事後調整如公式所示。調整後的閾值為 p^* (若將代價均換成 1 時等同原始閾值 0.5)，但此分法僅限於分類結果為機率的模型。
$$p^* = \frac{C_{(1,0)}}{C_{(1,0)} + C_{(0,1)}} = \frac{\text{FP}}{\text{FP} + \text{FN}}$$

— **直接法** (direct method) : 此類方法是將代價敏感學習直接引入不同的演算法當中，最簡單的形式為引入懲罰函數例如**分類錯誤率**、**交叉熵**或是**最小平方差**中，在決策樹、支持向量機、類神經網路等都有更進階的嵌入應用。

➤ 例如在決策樹中，長樹策略是將代價的資訊用於：(1) 選擇最優的特徵來劃分數據；(2) 決定某棵子樹是否需要被剪枝。

— **集成學習**：將代價用於權重的調整，使得模型滿足代價敏感的特性，例如 AdaCost (Fan et al., 1999) 與 Cost-Sensitive Random Forest 。

➤ AdaCost 加強了 AdaBoost 更新權重的策略。AdaBoost 是基於對錯誤分類的樣本權重的加大，AdaCost 則進一步對代價高的錯誤分類樣本加大權重。

➤ Cost-Sensitive Random Forests (Krawczyk et al., 2014) 將代價應用於弱決策樹模型的分支策略，並在投票階段同時引入考慮代價的權重。

□ 一分類學習 (One-Class Learning)

- 在數據極度不平衡的情況下，只由多數群建立一個單一分類的模型。

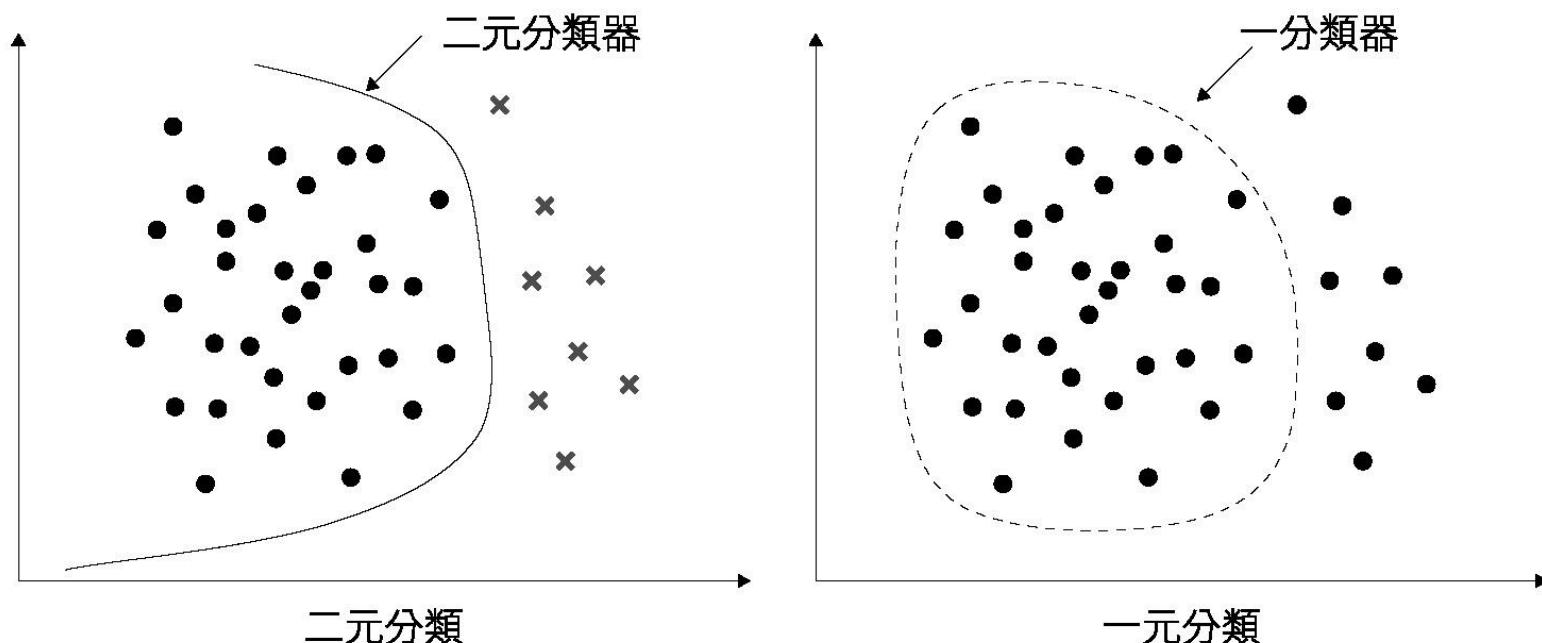
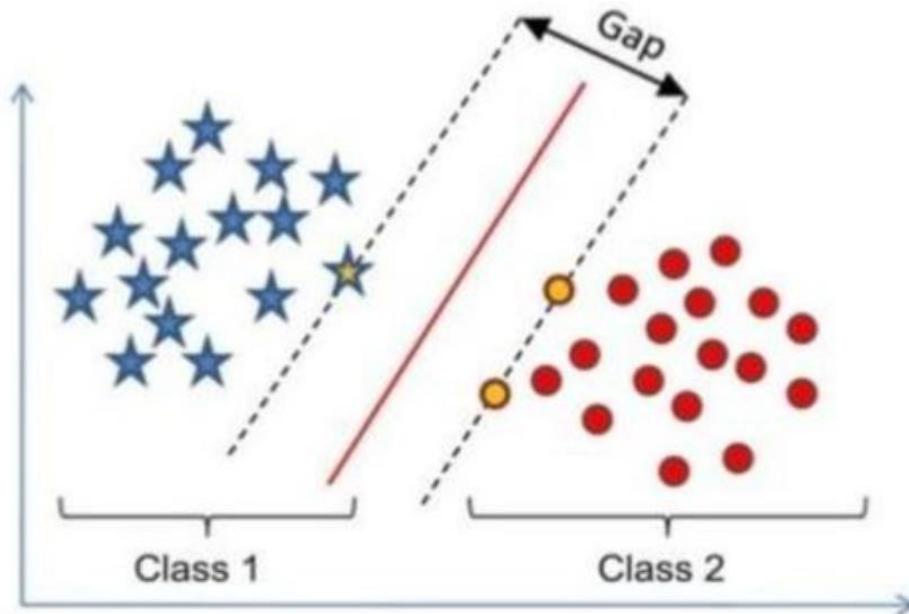


圖 12.29 二元分類與一分類比較

□ 一分類學習方法

- 異常偵測 (anomaly detection) (Liu et al., 2012) 的方法包含了
 - 一分類支持向量機 (one-class SVM) (Schölkopf et al., 2000)
 - 支持向量數據描述 (SVDD) (Sun and Tsung, 2003)
 - 孤立森林 (Isolation Forest) (Liu et al., 2008)
 - 區域離群因子 (Local Outlier Factor, LOF) (Breunig et al., 2000)
 - 卷積自編碼器 (convolutional autoencoder) (Kwak and Kim, 2021)
- Anomaly Detection
 - Novelty detection (奇異值檢測) : 訓練集沒有異常樣本
 - One-class SVM
 - Local Outlier Factor
 - Outlier detection (離群值檢測) : 訓練集有異常樣本
 - Robust Covariance
 - Isolation Forest

□ 支持向量機(support vector machine, SVM)



- Supervised, binary and linear classifier
- Decision surface is a hyperplane
- Maximizes the margin between classes (hard-margin SVM)
- If data non-separable, maximizes the margin and minimizes misclassifications (soft-margin SVM)
- If data non-separable, maps the input space to a higher dimensional feature space (kernels)

$$\max \frac{2}{\|w\|}$$

s.t. $y_i(w \cdot x_i + b) \geq 1, \forall x_i$

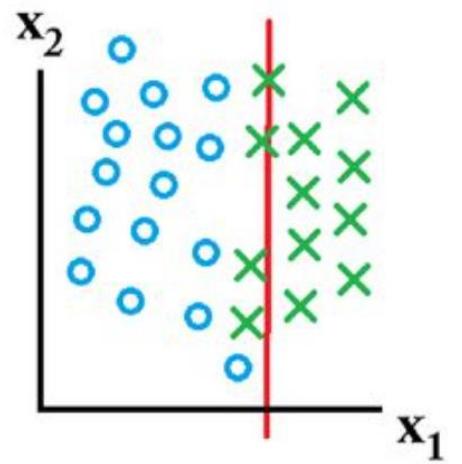
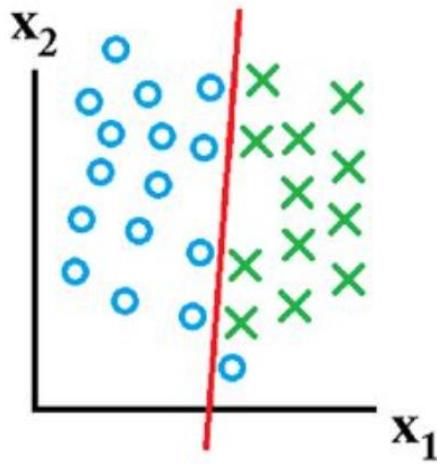
or

$$\min \frac{1}{2} \|w\|^2$$

s.t. $y_i(w \cdot x_i + b) \geq 1, \forall x_i$

Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*, 58, 121-134. <https://slideplayer.com/slide/15169258/>

□ 支持向量機(support vector machine, SVM)



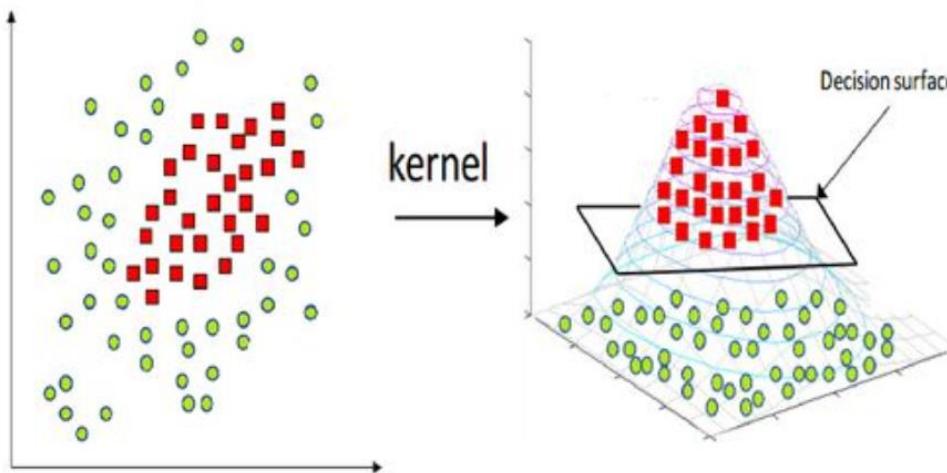
- Supervised, binary and linear classifier
- Decision surface is a hyperplane
- Maximizes the margin between classes (hard-margin SVM)
- If data non-separable, maximizes the margin and minimizes misclassifications (soft-margin SVM)
- If data non-separable, maps the input space, to a higher dimensional feature space (kernels)

$$\min \frac{1}{2} \|w\|^2 + C \sum_i \xi_i$$

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall x_i \\ \xi_i \geq 0$$

Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*, 58, 121-134. <https://slideplayer.com/slide/15169258/>

□ 支持向量機(support vector machine, SVM)

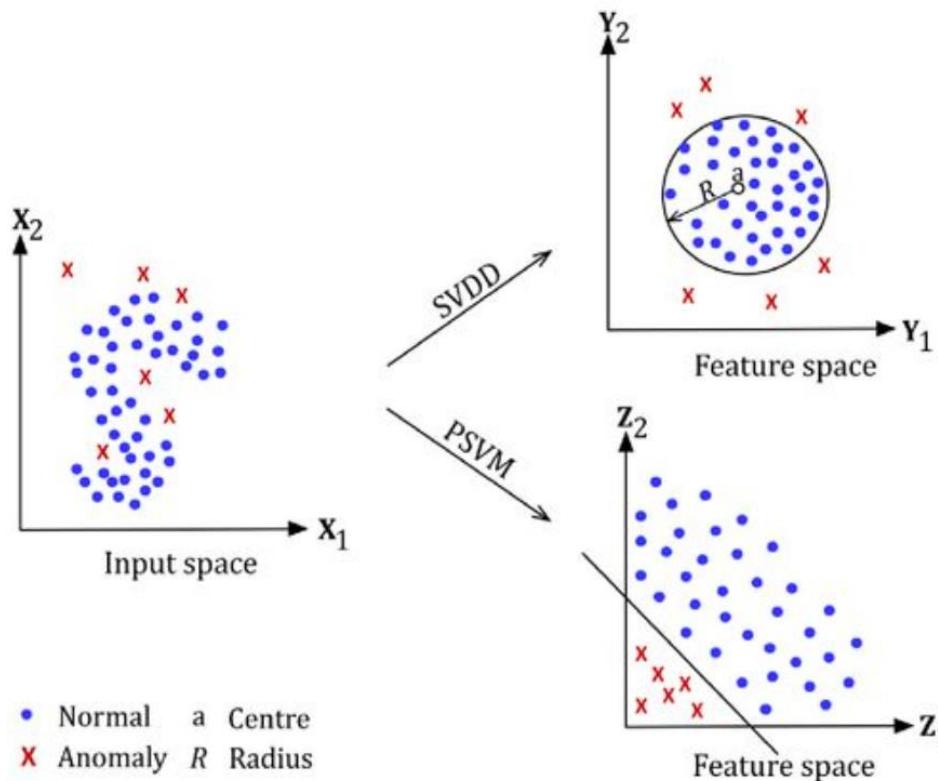


- Supervised, binary and linear classifier
- Decision surface is a hyperplane
- Maximizes the margin between classes (hard-margin SVM)
- If data non-separable, maximizes the margin and minimizes misclassifications (soft-margin SVM)
- If data non-separable, maps the input space, to a higher dimensional feature space (kernels)

Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*, 58, 121-134. <https://slideplayer.com/slide/15169258/>

□ 一分類學習方法 (one-class SVM)

- Plane-based one-class SVM (Schölkopf et al., 2001)
- Support vector data description (SVDD) (Sun and Tsung, 2003)



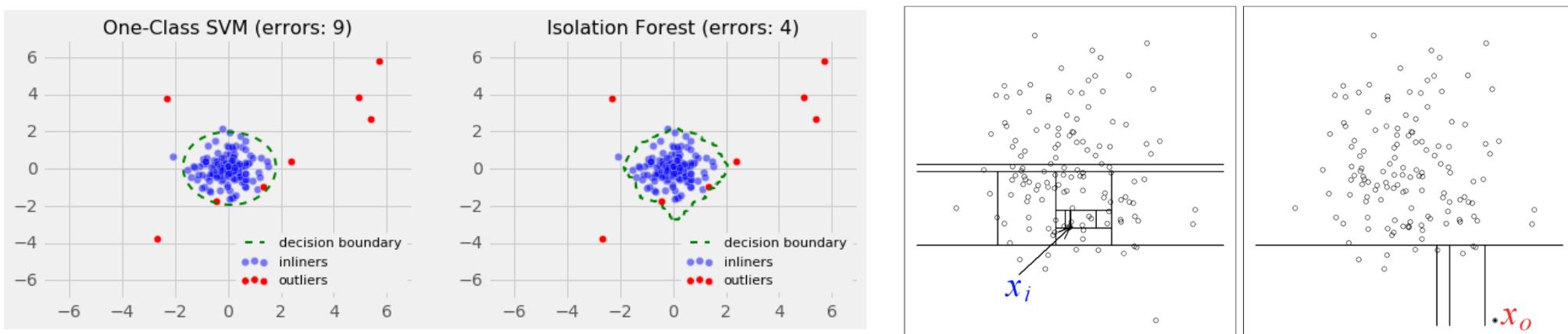
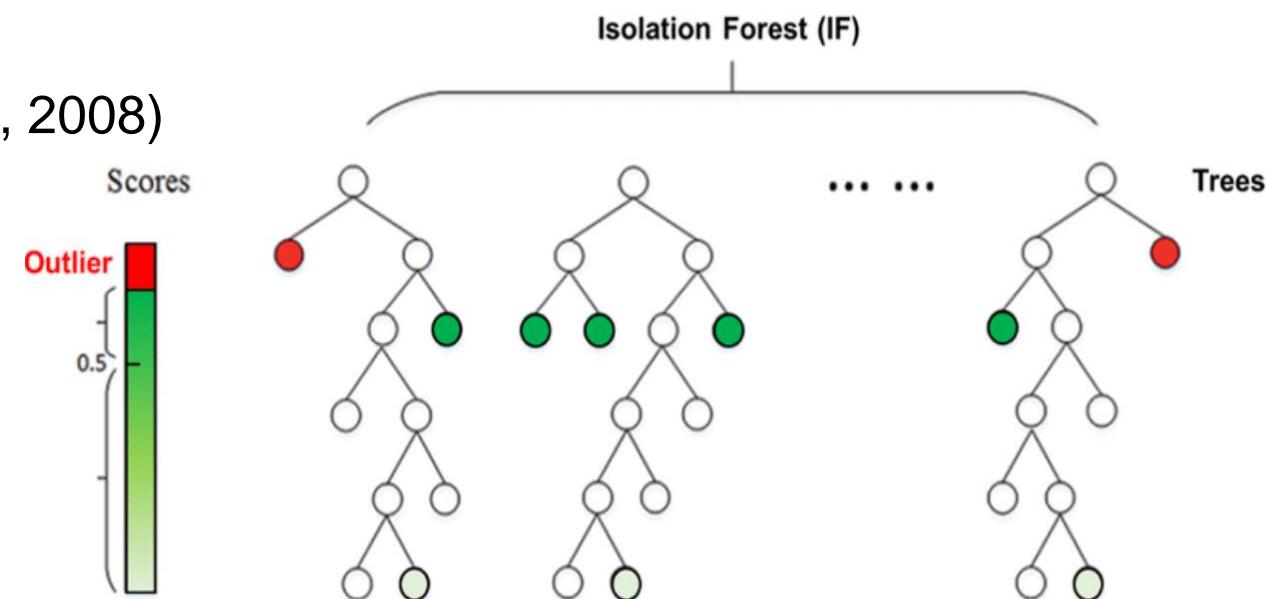
- Determines a smooth surface or boundary, that separates the transformed vectors into normal and anomalous observations
- Unsupervised separation, that lacks the need of expensive and laborly created labeled datasets
- Constructs a model of normal behavior, where data points that deviate from that model are classified as anomalies

Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*, 58, 121-134. <https://slideplayer.com/slide/15169258/>

□ 一分類學習方法

● Isolation Forest (Liu et al., 2008)

- Using all training data
- Randomly select branching feature/value
- May cause “deep tree”



<https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e?gi=39e026195409>

<https://albahnsen.com/2016/10/31/benefits-of-anomaly-detection-using-isolation-forests/>

https://www.researchgate.net/figure/Overview-of-the-isolation-forest-method-Light-green-circles-represent-common-normal_fig3_341629782

□ 一分類學習方法

- 區域離群因子 (Local Outlier Factor, LOF)

- 和周遭點的密度進行比較 判斷是否離群

- 透過定義每個資料點與其周遭點的距離來定義密度，藉此判斷新資料點(黑點 Point B)是否為離群值

- B為正常值

B點密度和臨域點皆很高

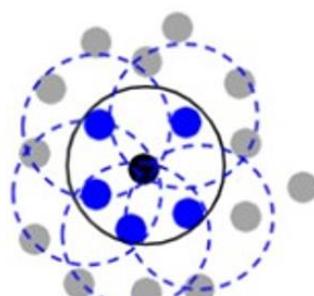
- B為異常值

B點密度較臨域點密度低

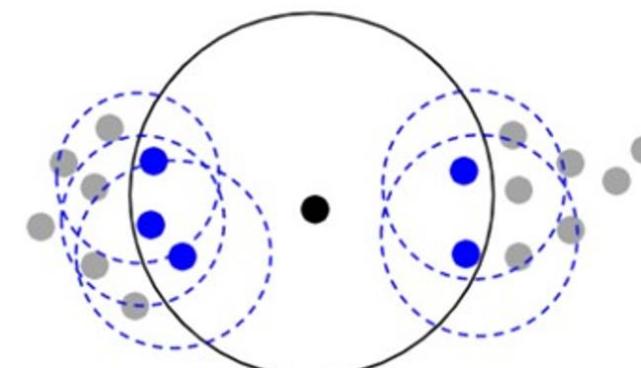
- B為正常值

B點密度和臨域點皆很低

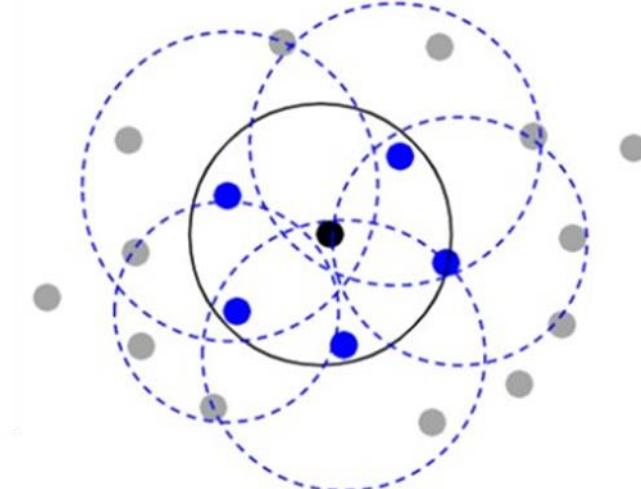
Case 1



Case 2



Case 3



● : 測試點(test point B)

● : 黑點的周遭點

● : 藍點的周遭點

□ 方法優劣與比較

- 當我們遇到數據不平衡時，要如何在這眾多的方法論中做出抉擇呢？
- 這裡提出以下三個原則
 1. 確立問題與動機
 2. 剖析手上的數據
 3. 釐清方法的假設、使用時機、限制、優與劣
- 掌握這三個原則便可有依據且合理的選擇合適的方法。

- 舉例而言，假設今天我們要建立一個預測產品不良種類的模型，而背後的動機在於特定不良種類的發生，可能是
 1. 機台參數調控不佳與環境因子的影響
 - 根因分析，這時關注的是模型的解釋性
 2. 此產品若有後續的製程將導致機台異常
 3. 產品全部檢驗的成本過高
 - 第2與第3個動機牽涉不良種類所帶來的影響，我們希望盡可能的找出特定不良種類的產品，這時關注的則是不同誤判產生的代價或成本。

□ 方法優劣與比較

- 接著，我們可能發現了數據不平衡的問題，這時**數據量與不平衡比例 (IR)**這兩個特性最為關鍵，當數據量大但不平衡比例高時，或許還能用數據去彌補不平衡的問題，一旦數據量小且不平衡比例極高時，問題便變得相當棘手。
- 尤其是，多個類別分類的問題，在機台異常類別 (down code) 或產品不良類別 (defect code) 問題時常屬於多個類別分類的問題，若將這些異常與不良的類別去剖析，可能有階層、重疊的關係，或是特定類別的量極少且影響較小，這時可以考慮**合併或刪除**某些類別，在合併後來減輕數據不平衡的狀況，而這種數據預處理也需具備一定的領域知識。

□ 方法優劣與比較

表 12.4 數據層面的數據平衡方法比較

方法	假設	使用時機				優	劣	
		數量	不平衡率	運算資源	代價矩陣			
上抽樣 over-sampling	-	小	高	少	無	使用容易	資訊過於特定導致過度配適	
下抽樣 under-sampling	-	大	高	少	無	使用容易	損失原有的資訊量	
混合抽樣 over- and under-sampling	-	皆可	非常高	少	無	使用容易	包含上述兩個缺點，但個別相對不嚴重	
數據生成	SMOTE	少數群間在特徵空間上具高相似度	小	高	足夠	無	生成樣本具有差異	1. 不同類別的樣本可能重疊 2. 新的資訊量不具解釋力
	ADASYN	少數群間在特徵空間上具高相似度	小	高	足夠	無	少數群樣本比例低的空間有更多資訊量	1. 同 SMOTE 2. 對離群值較為敏感
	B-SMOTE	少數群間在特徵空間上具高相似度	小	高	足夠	無	生成樣本在分類邊界上	1. 同 SMOTE 2. 分區界線隱含多個超參數需設定
數據清理	ENN	多數群間在特徵空間上具高相似度	大	高	足夠	無	移除樣本後的邊界明確	1. 多個超參數需設定 2. 可能移除了邊界上的重要資訊
	Tomek Links	多數群間在特徵空間上具高相似度	大	高	足夠	無	移除樣本後的邊界明確	移除大量樣本可能造成重要資訊的損失
混合策略	B-SMOTE+ENN	樣本間在特徵空間上具高相似度	皆可	非常高	足夠	無	同時具備數據生成與清理的優點	同時具備數據生成與清理的缺點
具資訊抽樣	EasyEnsemble	-	大	高	多	無	引入集成學習的 Bagging，且無增加與損失數據	1. 多個超參數需設定 2. 不平衡率過高時運算效率低
	BalanceCascade	-	大	高	多	無	引入集成學習的 Boosting，且無增加與損失數據	同 EasyEnsemble

□ 方法優劣與比較

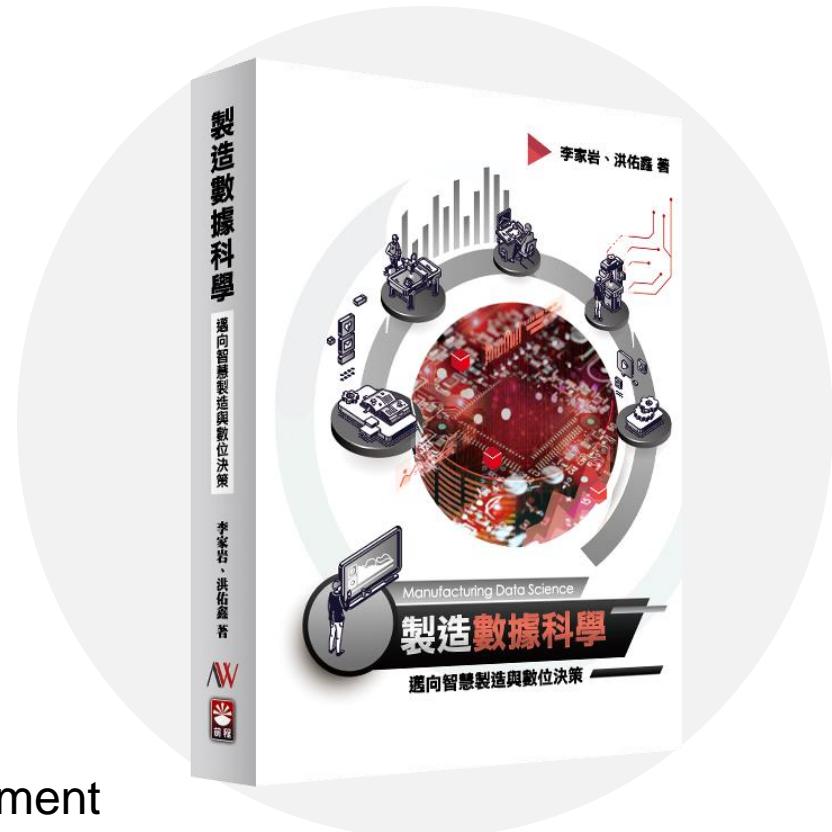
表 12.5 演算法層面的數據平衡方法比較

方法		假設	使用時機				優	劣
			數據量	不平衡率	運算資源	代價矩陣		
代價敏感學習	Thresholding	模型結果需為機率	小	高	少	給定	使用容易，可與成本連結	受限於模型結果為機率
	DT,SVM,NN (embedded)	-	小	高	足夠	給定	可用於解釋性模型，並與成本連結	嵌入技巧較為複雜
	AdaCost	-	大	高	多	給定	引入集成學習的 Boosting，可與成本連結	多個超參數需設定
	Cost-Sensitive RF	-	大	高	多	給定	引入集成學習的 Bagging，可與成本連結	多個超參數需設定
一分類學習	SVDD	樣本間在特徵空間上具高相似度	皆可	極高	足夠	無	不需考慮任何少數群樣本	超參數需設定
	Isolation Forest	異常點的特徵值與正常點特徵值有所差異	皆可	極高	少	無	不需考慮任何少數群樣本；基於集成學習的線性時間複雜度	超參數需設定

□ 數據分析中重要的三大關鍵—特徵工程、數據增強與數據平衡

- 除了**模型演算法**本身的選擇會影響預測準確度之外，這三大步驟對於數據上所下的功夫，**也是能有效提升預測準確度的重要關鍵步驟**。對於數據科學實務上的應用，有著舉足輕重的角色。
- 事實上，模型演算法已有大量的**模組(module)**與**套件 (package)**去實踐，程式撰寫過程易於使用；然而，這三大步驟是對於**數據品質 (data quality)**的改善，經常需要**人與領域知識**的參與。這也說明了自動化與智能化，不單只是專注於設備機台或數據算法，同時還需考慮「**人+機**」的重要性，以達到「**人機協作**」(**human-robot collaboration**)之綜效。

Thanks for your attention



NTU Dept. of Information Management
name: 李家岩 (FB: Chia-Yen Lee)
phone: 886-2-33661206
email: chiayenlee@ntu.edu.tw
web: <https://polab.im.ntu.edu.tw/>