資料科學的特徵工程

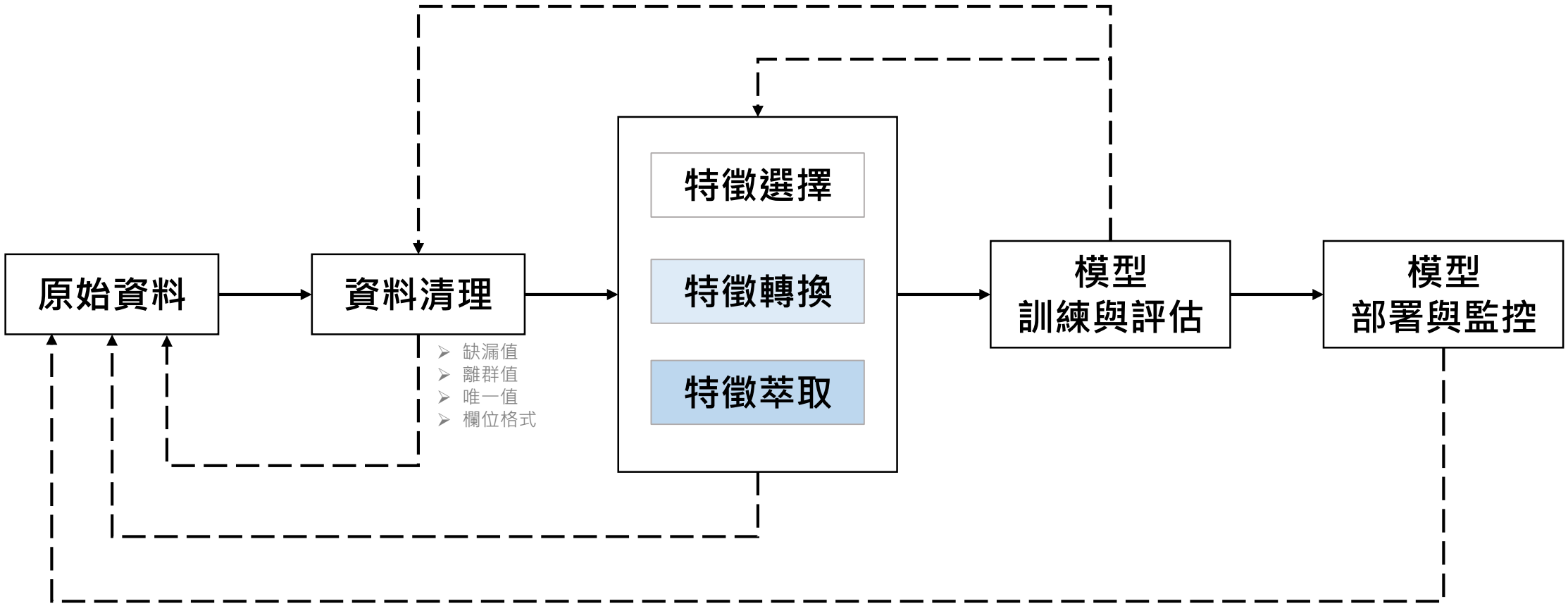
特徵轉換與特徵萃取的概念與手法

Feature Transformation

Feature Extraction

吳彥霖

ML建模的步驟



特徵轉換與特徵萃取

特徵轉換
Feature Transformation

- 改變或調整原始特徵變數的表現形式，提升特徵變數的可解釋性，與 ML 模型的適配性。
- 常見的方法：
  - 特徵分箱 Feature Binning
  - 特徵縮放 Feature Scaling

特徵萃取
Feature Extraction

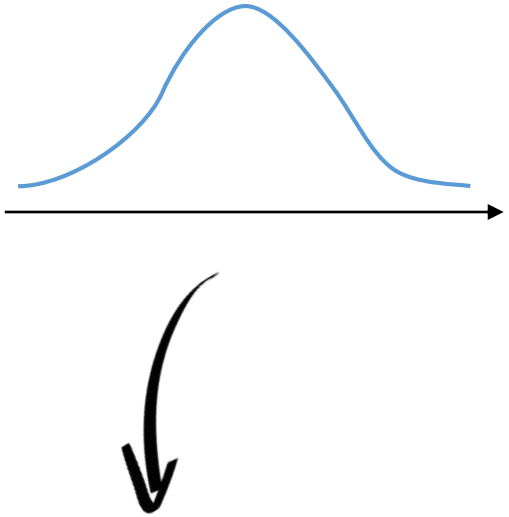
- 係一種特徵降維 Dimensionality Reduction 的處理步驟。
- 將原始特徵資料集，轉換成更具代表性且除去冗餘資訊的新特徵資料集，強化數據重要資訊，以提升 ML 模型的表現。
- 常見的方法：
  - 主成份分析 Principal Component Analysis, PCA

特徵分箱

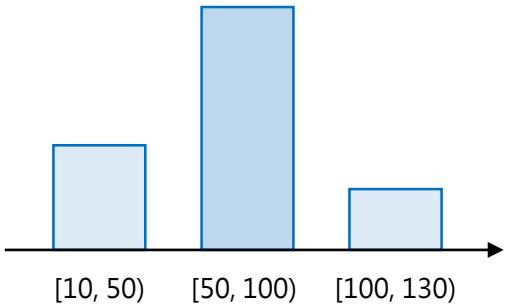
Feature Binning

特徵分箱 Feature Binning

原始特徵變數為連續型



分箱後為類別型



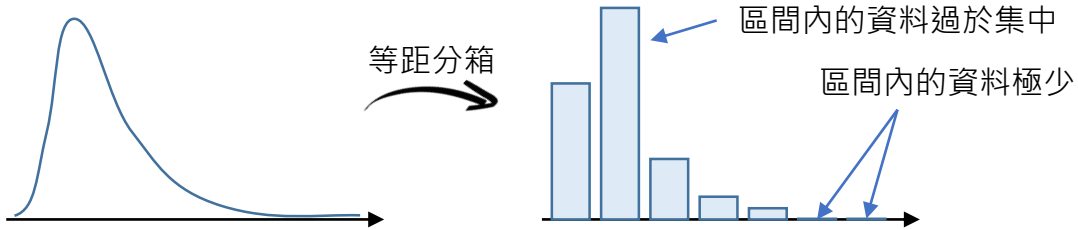
- 什麼是**特徵分箱**？
  - 目的係將連續型數據，轉換為離散型數據，具體而言，將特徵變數的數據範圍，劃分成多個區間，將同一個區間內的數據視為相同的一類。
- 為什麼需要**特徵分箱**？
  - **特定 ML 模型的需求**，有些 ML 模型的輸入須為離散型數據，例如：樸素貝葉斯分類器 Naive Bayes Classifier、貝葉斯網絡 Bayesian Networks。
  - 雖然，**樹模型沒有一定需先特徵分箱**，然而，分箱能減少樹的深度，提升運算效率、模型解釋性、降低過擬合的風險。
  - **降低離群值(或異常值)對 ML 模型的影響**，因為，分箱時會將其歸為某一區間的類別，避免過度影響 ML 模型。
- 常見的**特徵分箱**方法
  - 等距分箱 Equal Width Binning
  - 等頻分箱 Equal Frequency Binning
  - 聚類分析 Clustering Analysis

等距分箱與等頻分箱

分箱
Binning

等距分箱
Equal Width Binning

- 每一個**區間寬度** Bin Width **相等**，
  每一個區間的資料量不盡相同。
- **不適合**原始特徵變數**分布高度偏斜**的情況。



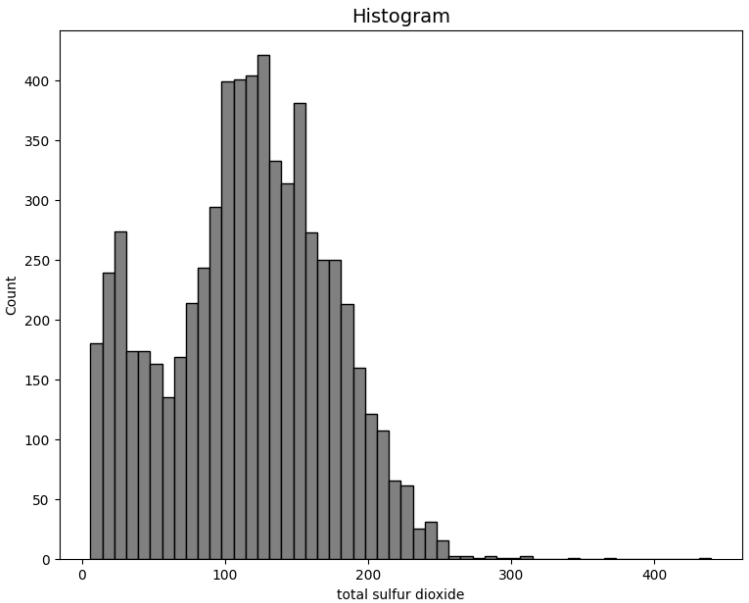
類別不平衡

等頻分箱
Equal Frequency Binning

- 每一個區間寬度 Bin Width 不盡相同，
  每一個**區間的資料量相等**。
- 利用**分位數**(Quantiles) 決定每一個區間的上下界。

等距分箱與等頻分箱

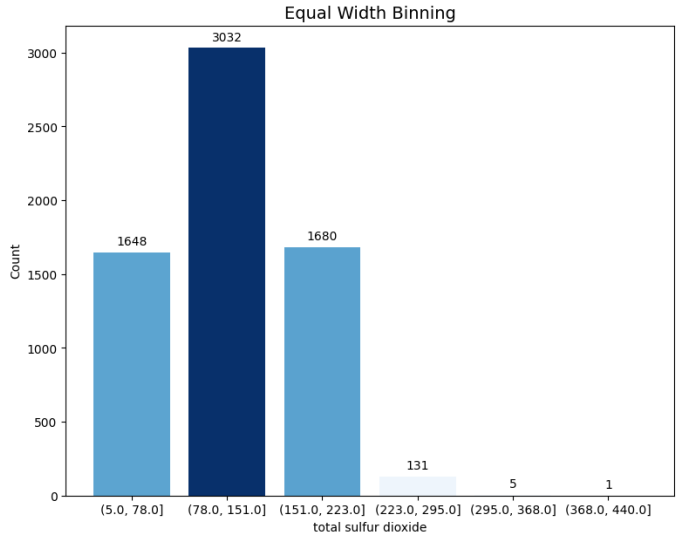
原始特徵變數的分布

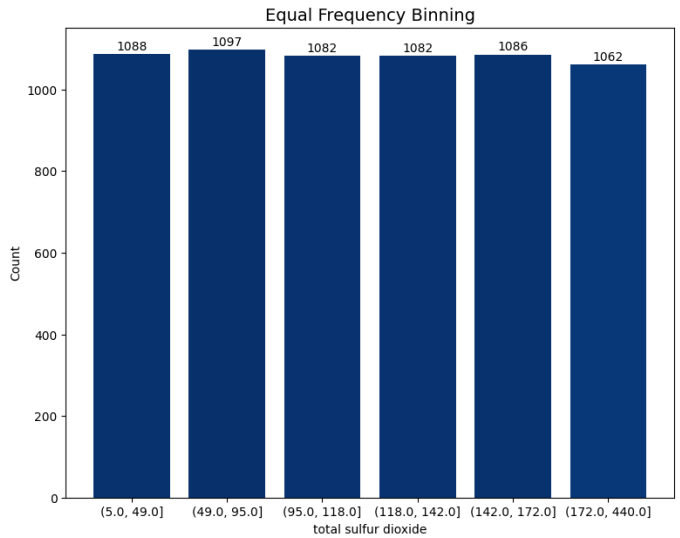


資料來源 : Kaggle Wine Quality

等距分箱
Equal Width Binning

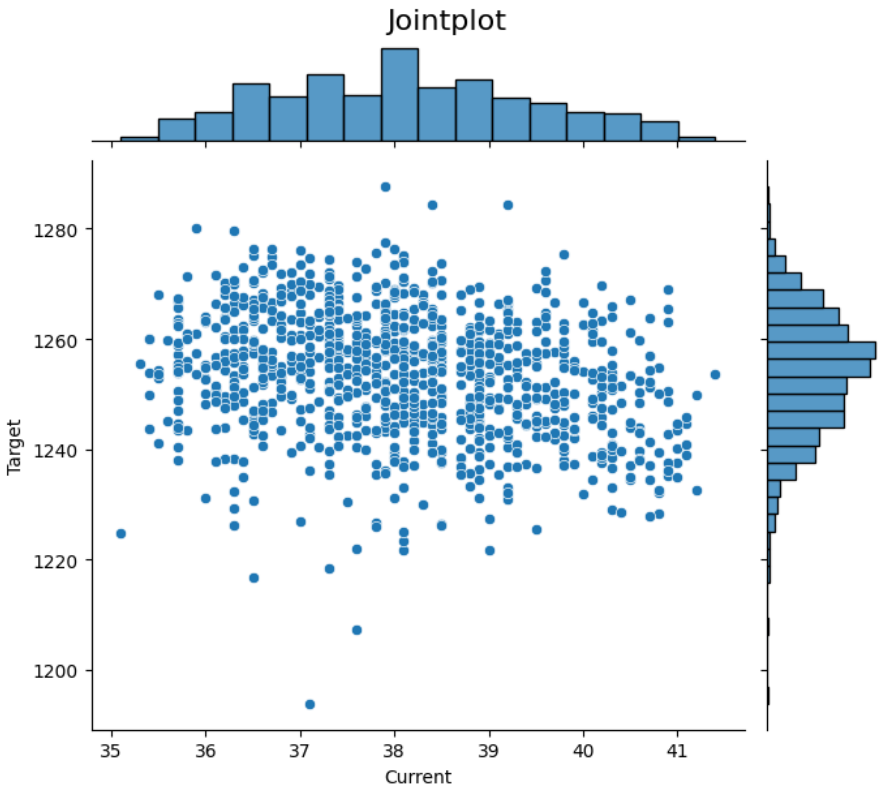
等頻分箱
Equal Frequency Binning





特徵分箱 Feature Binning

- 協助尋找資料中**隱性的差異**



$Pearson\ Correlation \approx -0.242$

【案例】

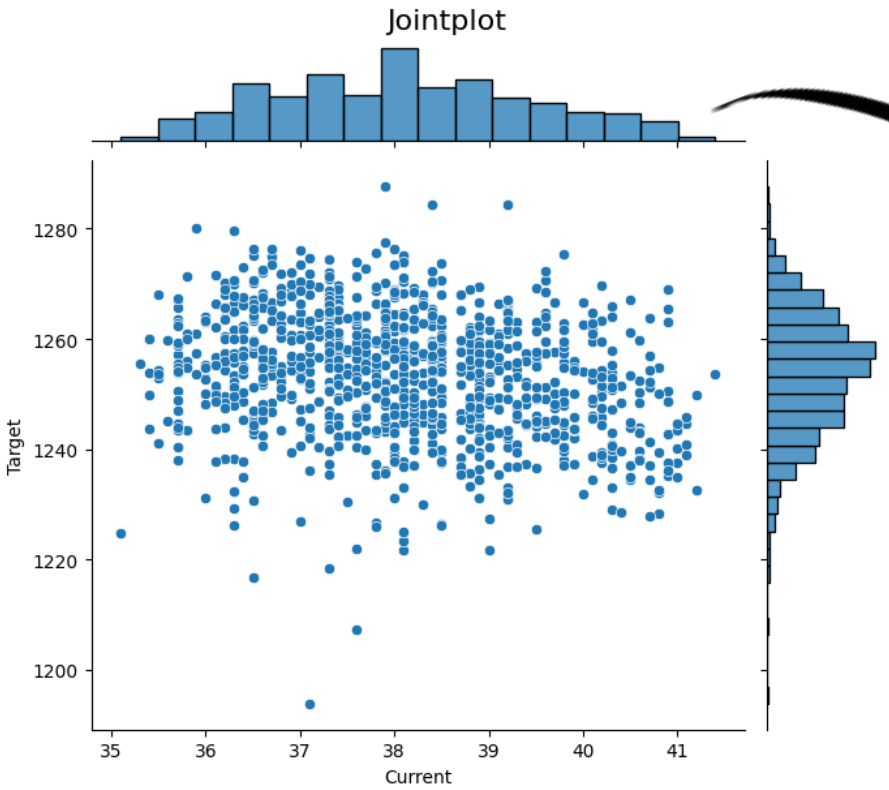
靶材電源供應器與量測值 Target 的關係？

- 電流 Currnet 與量測值 Target
  - 相關係數不高；
  - 散佈圖中，無法觀察出特殊的趨勢或形態。

特徵分箱 Feature Binning

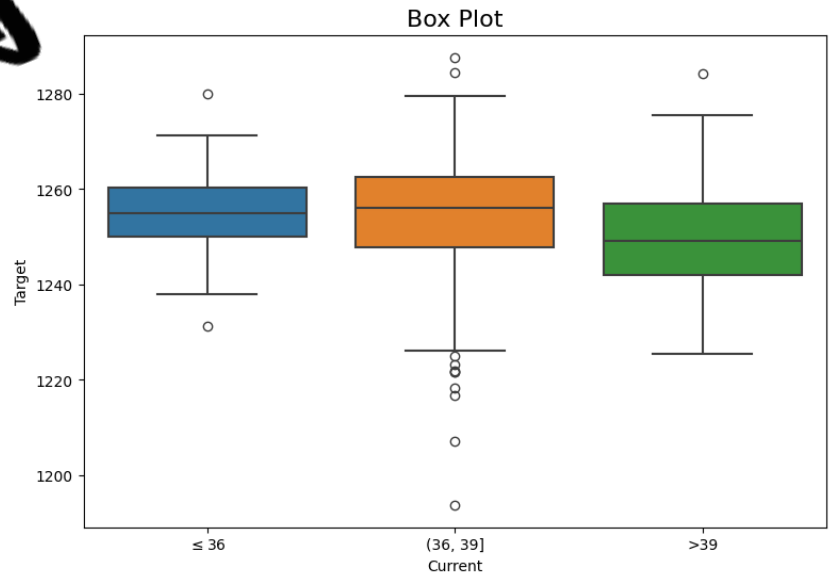
- 協助尋找資料中**隱性的差異**

【案例】



靶材電源供應器與量測值 Target 的關係？

Current
自定義分箱



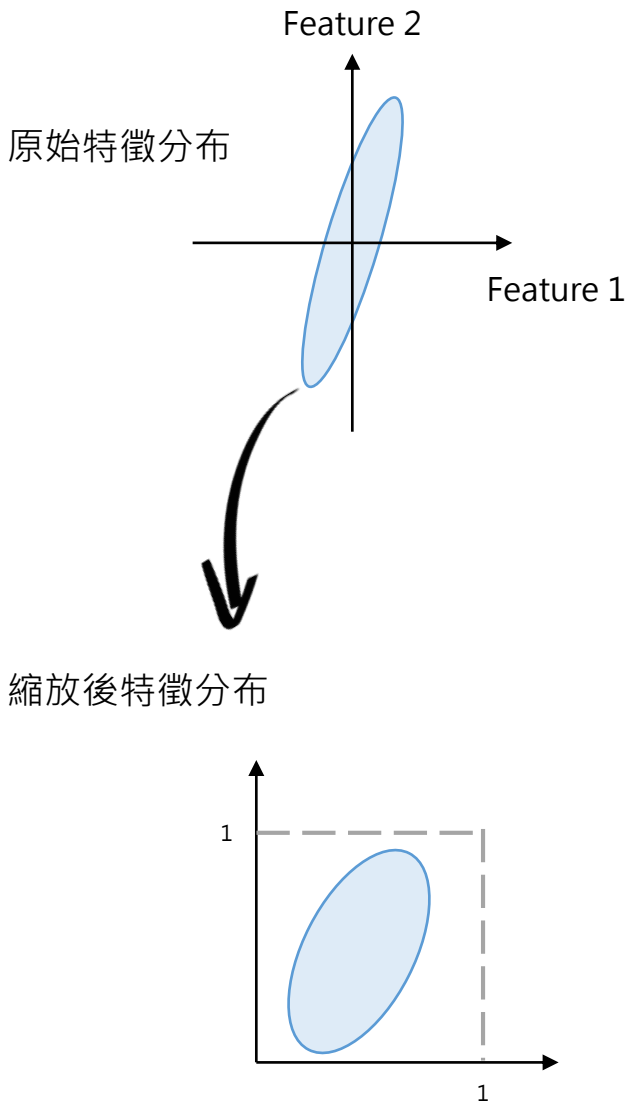
$Pearson\ Correlation \approx -0.242$

- 不同電流範圍的量測值 Target 變異數有顯著差異 (in Brown-Forsythe Test)
- 不同電流範圍的量測值 Target 中位數有顯著差異 (in Kruskal-Wallis Test)

特徵縮放

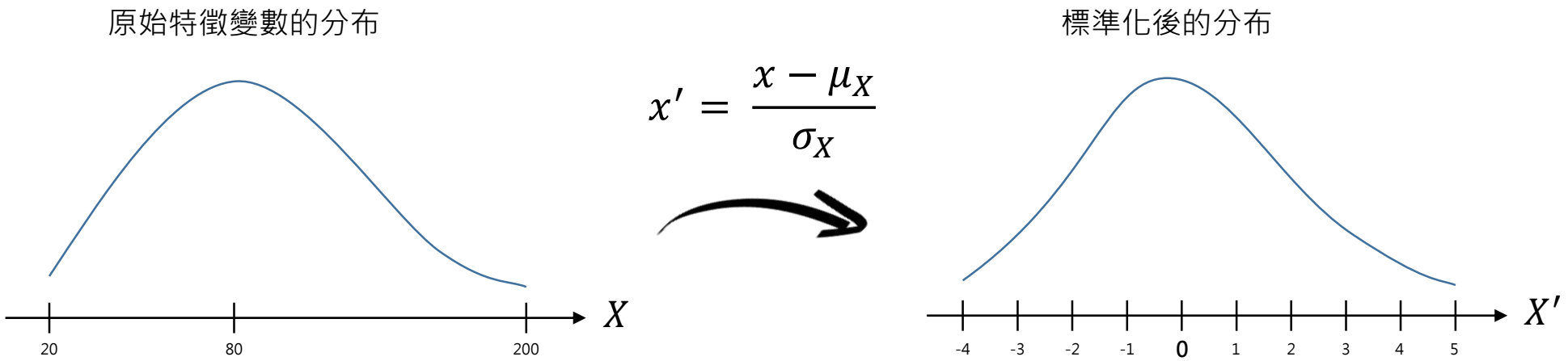
Feature Scaling

特徵縮放 Feature Scaling



- 什麼是**特徵縮放**？
  - 目的係將不同數值尺度的特徵變數，轉換到相同(或相近)的數值區間，避免因為特徵變數之間數值尺度的差異，造成解讀與判斷的偏差，幫助我們能更客觀理解與比較數據。
- 為什麼需要**特徵縮放**？
  - 對於一些基於距離或梯度收斂的 ML 演算法，特徵縮放能提升 ML 模型的訓練效率，與確保模型的預測表現。
  - 概略而言，特徵縮放是將不同的特徵變數，數值調整到同一個尺度上，使得每個特徵對 ML 模型的影響力更均衡。
- 常見的**特徵縮放**方法
  - 標準化 Standardization
  - 最小最大縮放 Min-Max Scaling
  - 絕對最大縮放 Max-Abs Scaling

標準化 Standardization



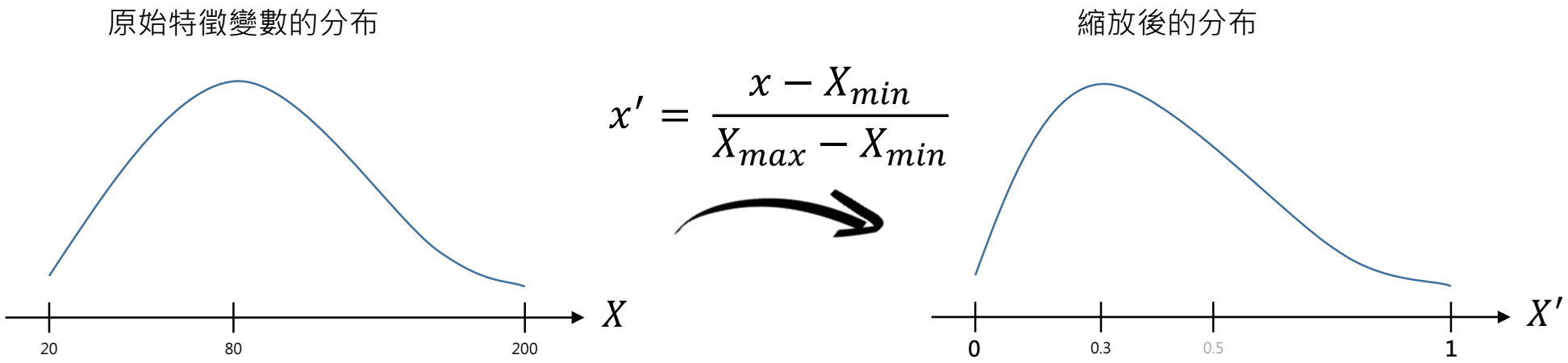
目的

- 將原始特徵變數轉換成平均值 $\mu = 0$ 、標準差 $\sigma = 1$。
- 若原始特徵變數近似於常態分布，經標準化後，會近似於標準常態分布。

注意事項

- 標準化的過程，會受離群值(或異常值)所影響，建議在標準化前，應先處理特徵變數的離群值(或異常值)。

最小最大縮放 Min-Max Scaling



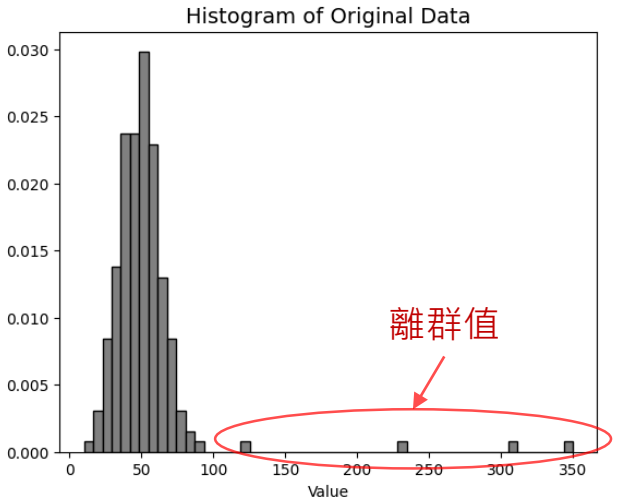
目的

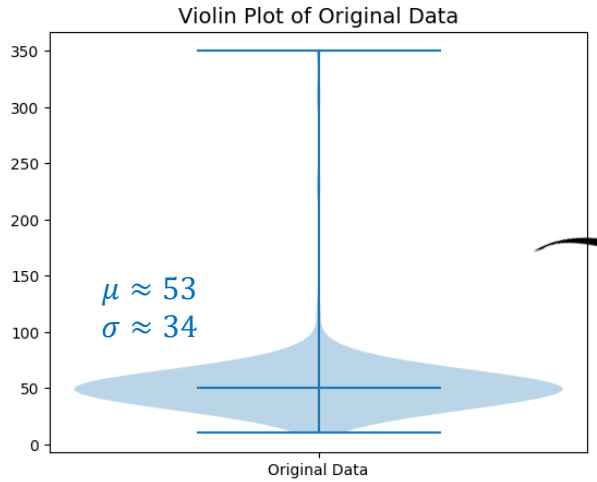
- 將原始特徵變數轉換到 $[0, 1]$ 固定區間中 。
- 最小最大縮放會保持原始特徵的數值相對比例。

注意事項

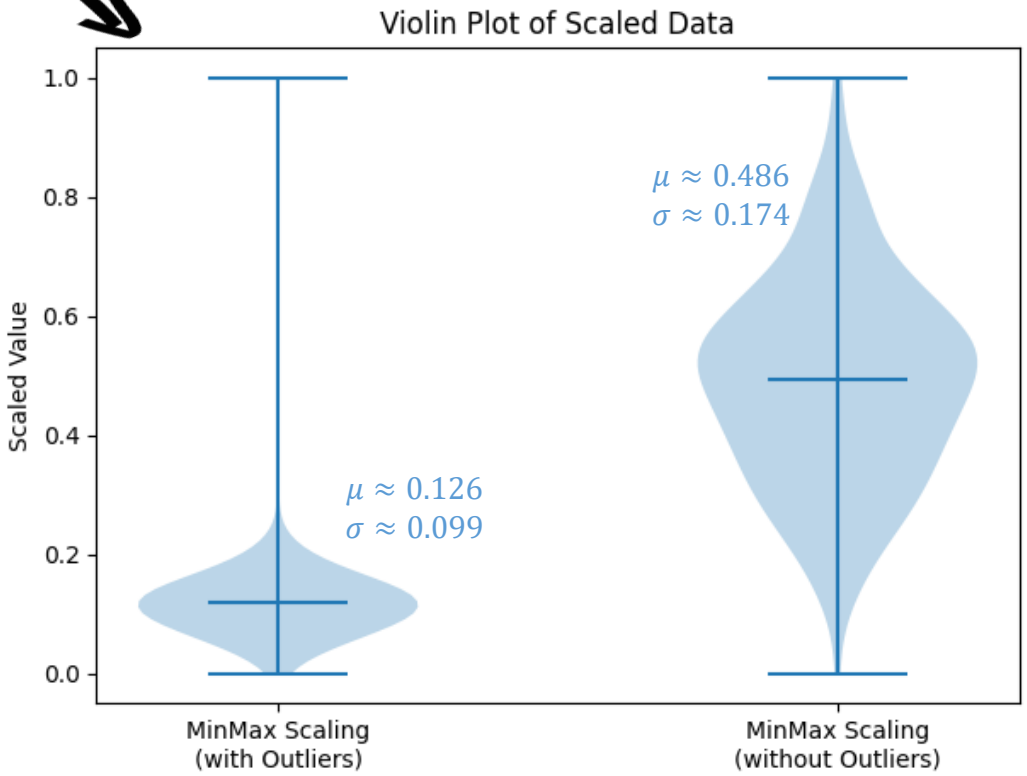
- 最小最大縮放會受離群值(或異常值)的影響，造成縮放後的分布極度不均。
  建議：可刪除離群值(或異常值)，或使用對離群值較不敏感的方法，例如：穩健縮放 Robust Scaling 。

最小最大縮放 Min-Max Scaling



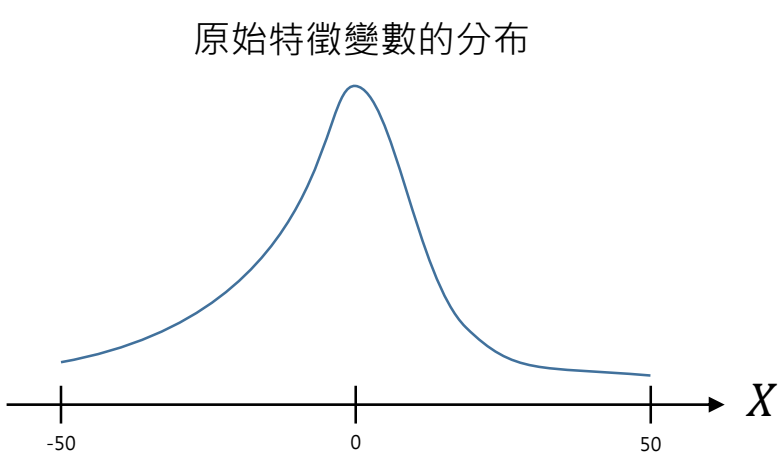


$$x' = \frac{x - X_{min}}{X_{max} - X_{min}}$$

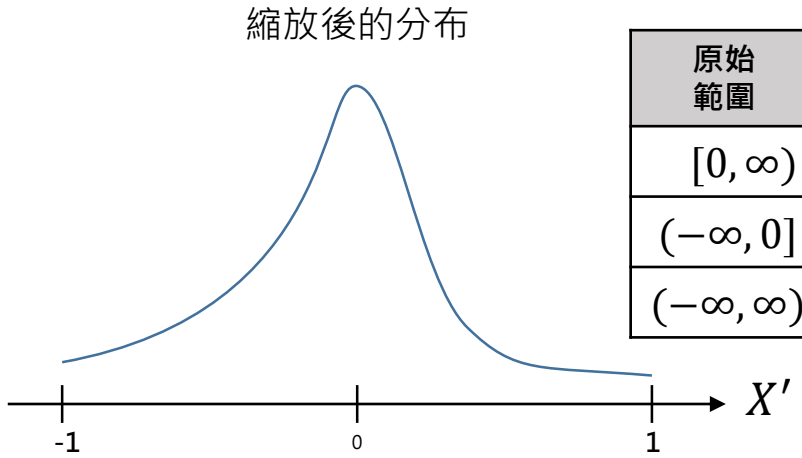


- 當特徵變數存在**離群值**(或**異常值**)時，易導致**縮放後的數值集中在極小的範圍內**；
- 若，後續需與其他特徵變數，進行縮放後的分布比較，則，不易分辨出分布的差異。

最大絕對縮放 Max-Abs Scaling



$$x' = \frac{x}{\max|X|}$$



原始範圍	轉換後範圍
$[0, \infty)$	$[0, 1]$
$(-\infty, 0]$	$[-1, 0]$
$(-\infty, \infty)$	$[-1, 1]$

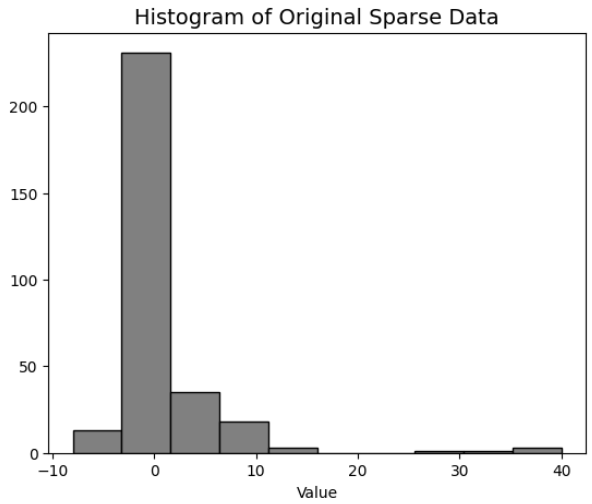
目的

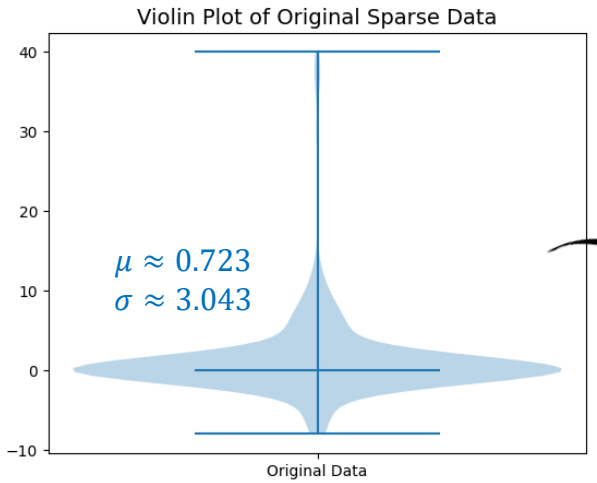
- 將原始特徵變數轉換到固定的區間中 。
- 特別適合處理稀疏資料(Sparse Data)，也就是，絕大部分的數值為 0。

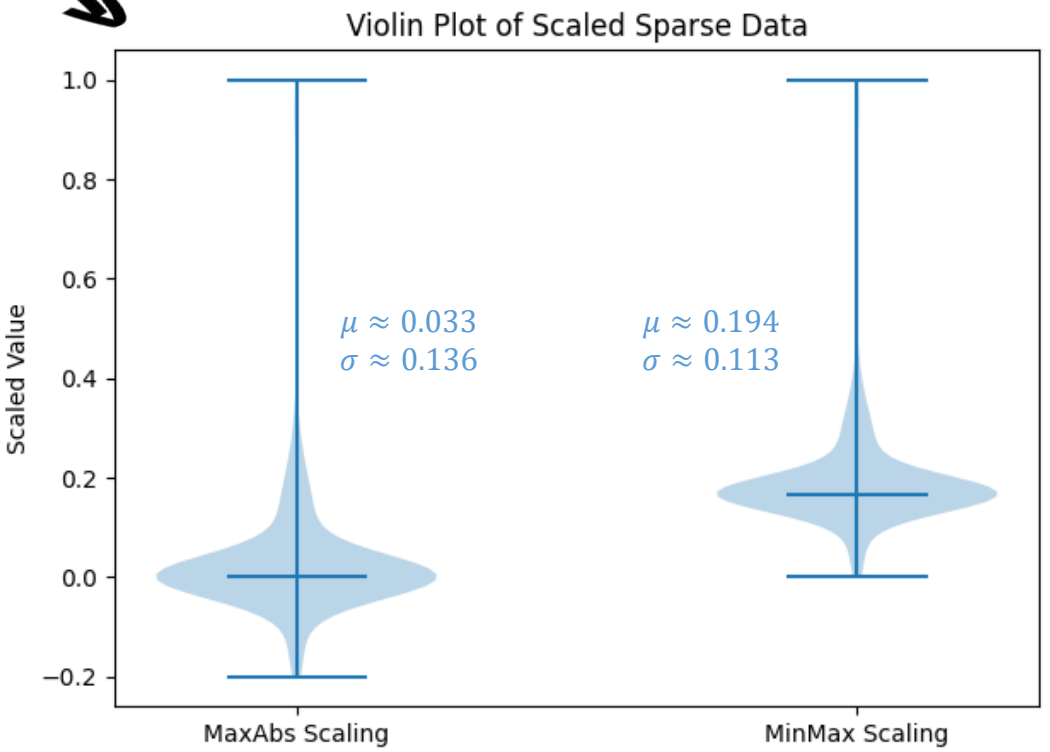
注意事項

- 最大絕對縮放會受離群值(或異常值)的影響，造成縮放後的分布極度不均。

最大絕對縮放 Max-Abs Scaling

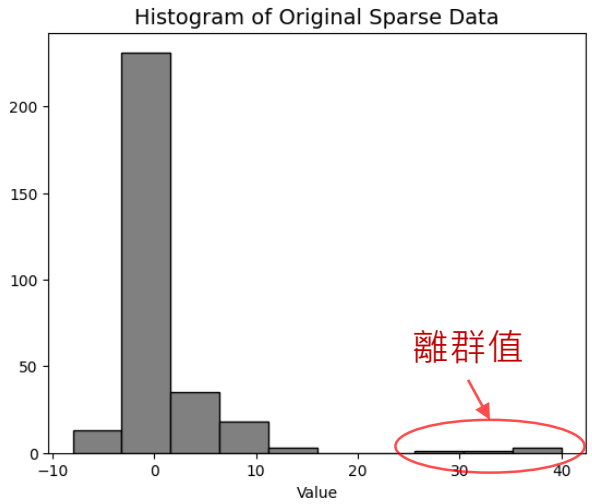


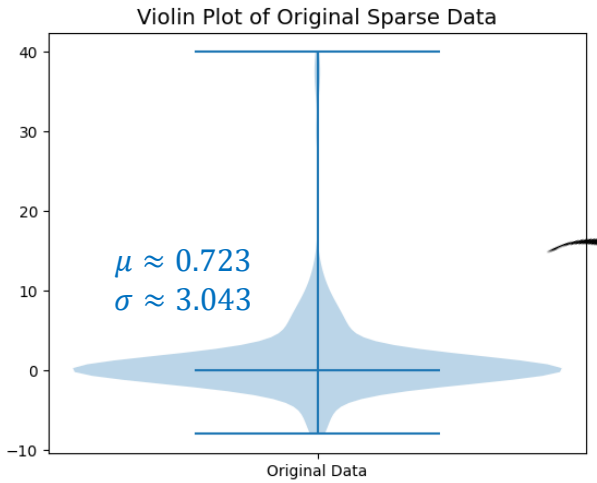




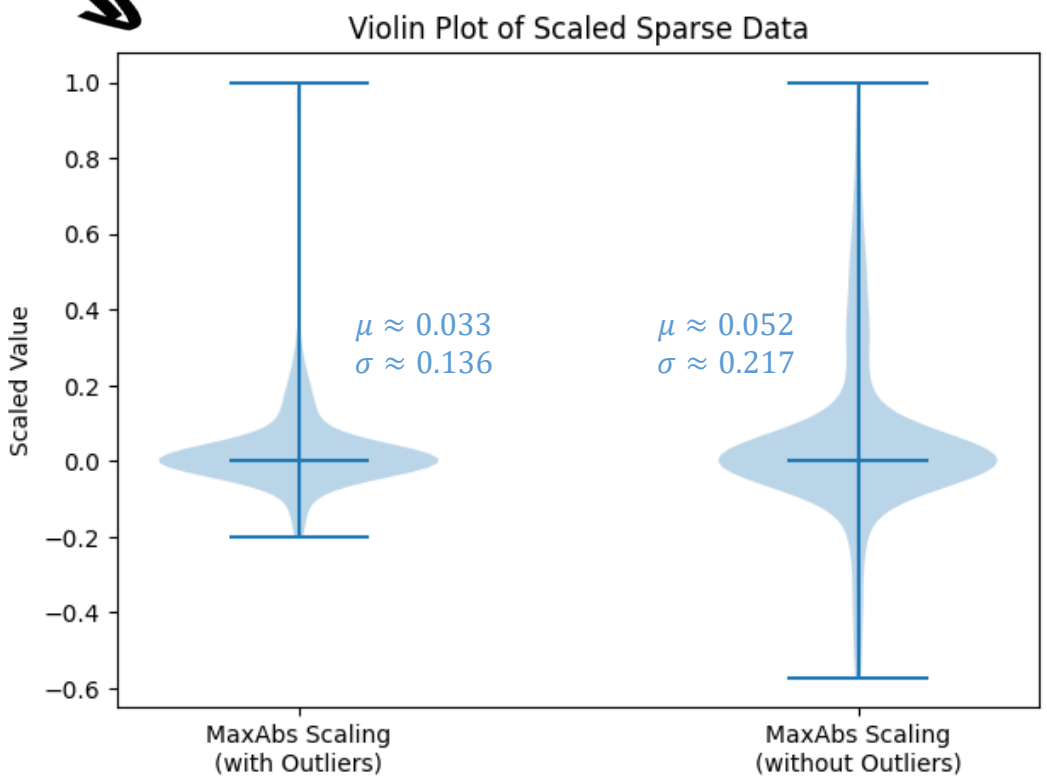
- 當特徵變數絕大多數為 0 時，稱為稀疏資料 Sparse Data；
- **最大絕對縮放**不改變原始稀疏資料的特性，
  - ✓ **只對原始的非 0 元素進行縮放**，
  - ✓ 對於原始數值為 0 不造成任何改變。

最大絕對縮放 Max-Abs Scaling





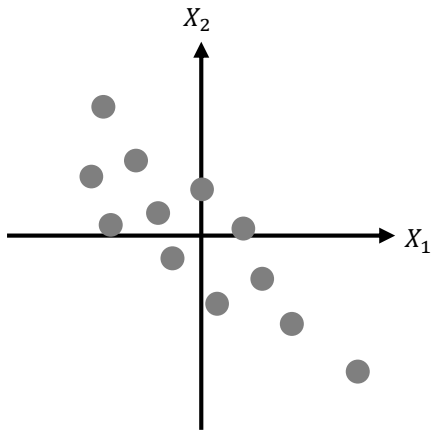
$$x' = \frac{x}{\max|X|}$$



- 當特徵變數存在**離群值**(或**異常值**)時，易導致**縮放後的數值集中在極小的範圍內**；
- 若，後續需與其他特徵變數，進行縮放後的分布比較，則，不易分辨出分布的差異。

特徵縮放 Feature Scaling

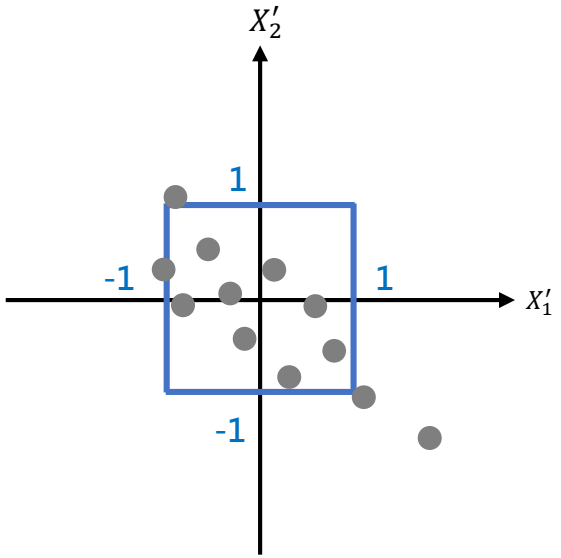
原始特徵分布



$$x' = \frac{x - \mu_X}{\sigma_X}$$

標準化

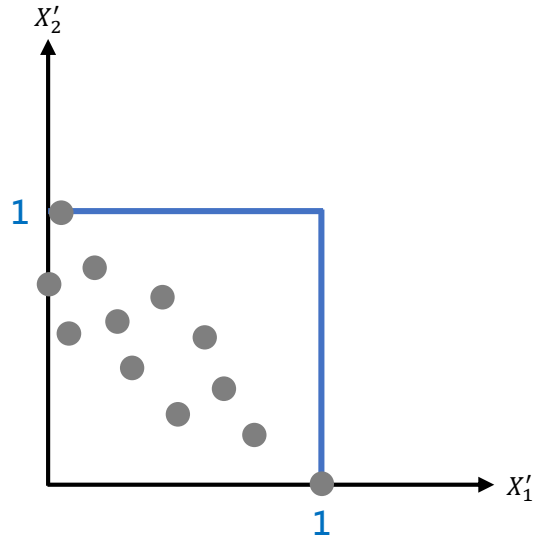
Standardization



$$x' = \frac{x - X_{min}}{X_{max} - X_{min}}$$

最小最大縮放

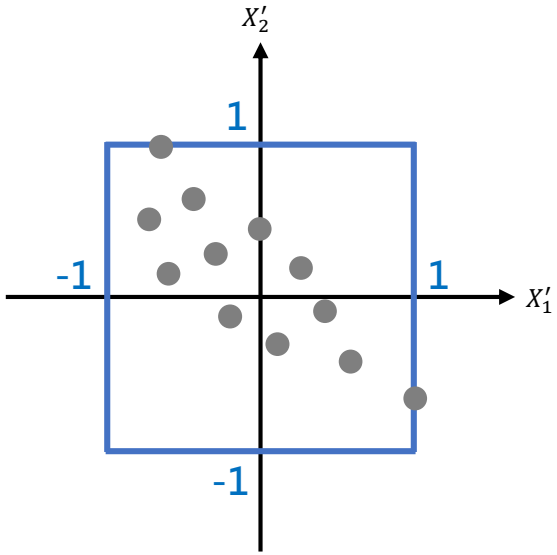
Min-Max Scaling



$$x' = \frac{x}{\max|X|}$$

最大絕對縮放

Max-Abs Scaling



主成份分析

Principal Component Analysis

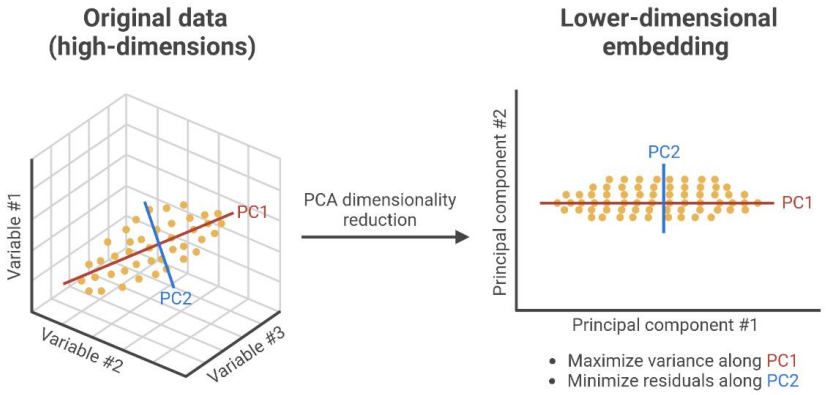
主成份分析 Principal Component Analysis

3 維

V1	V2	V3
10	0.9	150
20	0.2	165
⋮	⋮	⋮
15	0.5	140

2 維

PC1	PC2
0.95	0.4
0.55	-0.2
⋮	⋮
0.05	0.3



概念

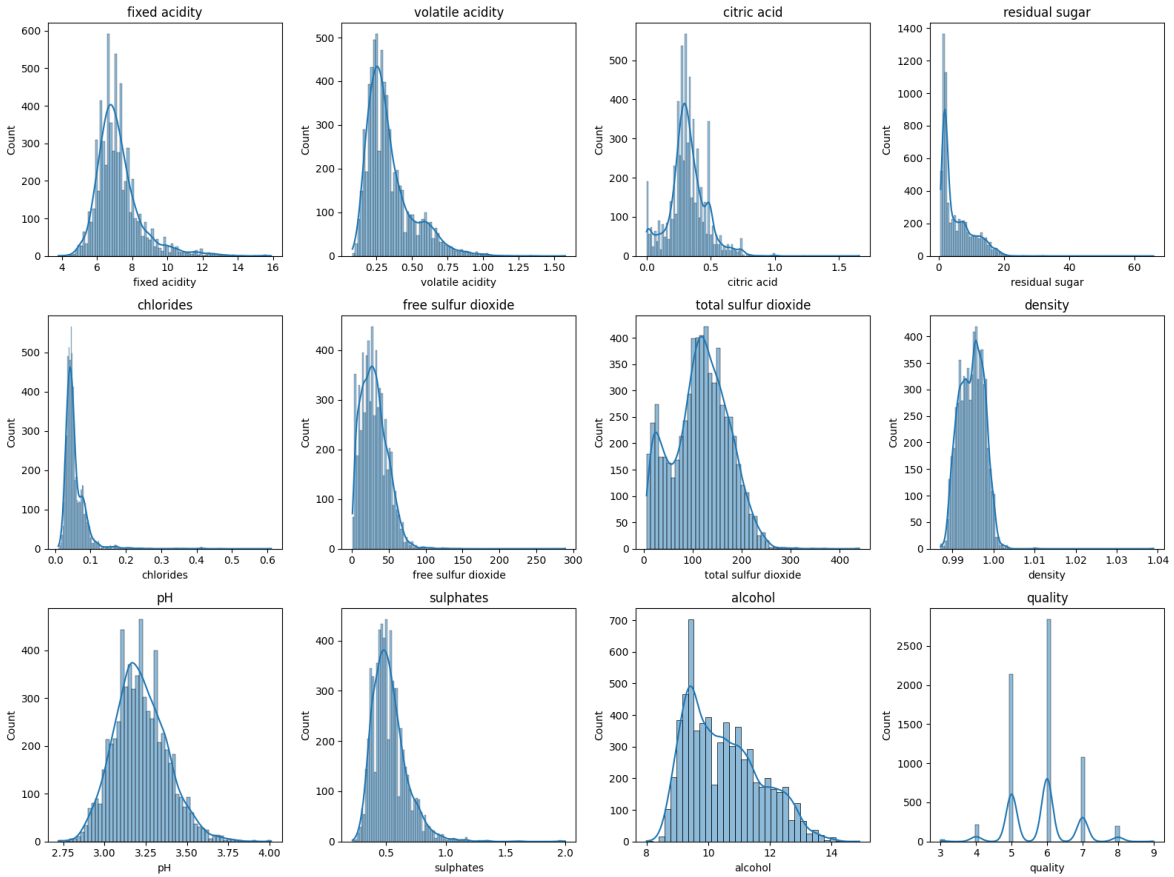
- 一種特徵降維的技術，將資料從 p 個特徵，壓縮成 k 個$(k < p)$主成份(Principal Component)特徵。
- 在維持資料最大變異性的前提下，將高維資料轉換成低維資料，能避免特徵變數間的共線性問題。
- 主成份的產生，是由各特徵變數的線性組合而成，主要依資料的變異程度，由大到小重新選定正交(Orthogonal)座標系，第一主成份為變異最大的方向、第二主成份為第二大變異的方向，依此類推。

注意事項

- **解釋性降低**：
  主成份為數學上最大變異性方向的結果，通常缺乏明確的物理與業務解釋，不一定有明確的意義。
- 特徵間具高度**線性相關**：
  若資料為高度非線性的關係，PCA 無法很好地捕捉資料的結構。
- 資料**標準化** Standardization：
  在進行 PCA 之前，會將每個特徵變數做標準化，能避免特徵變數之間不同單位尺度的差異，且在相同的基準下進行比較分析。

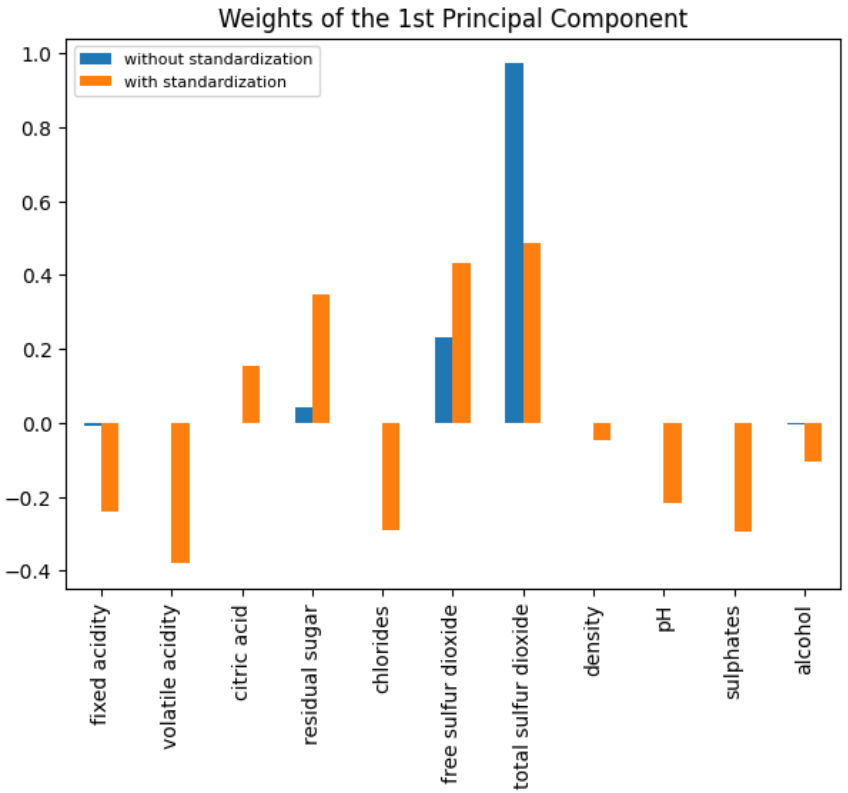
PCA 的資料前處理

【案例】



- 數據集名稱 : Wine Quality 資料來源 : [Kaggle Wine Quality]
  - 有關葡萄酒的品質、物理或化學特性值的紀錄資料
  - 12 個特徵變數、6,498 筆資料

PCA 的資料前處理



【案例】

- 數據集名稱 : Wine Quality 資料來源 : Kaggle Wine Quality
  - 有關葡萄酒的品質、物理或化學特性值的紀錄資料
  - 12 個特徵變數、6,498 筆資料

- 資料**無標準化**
  第一主成份的特徵組合(左圖藍色)，以 total sulfur dioxide 占絕大部分的權重，而， total sulfur dioxide 相較其他特徵變數的尺度來得大。

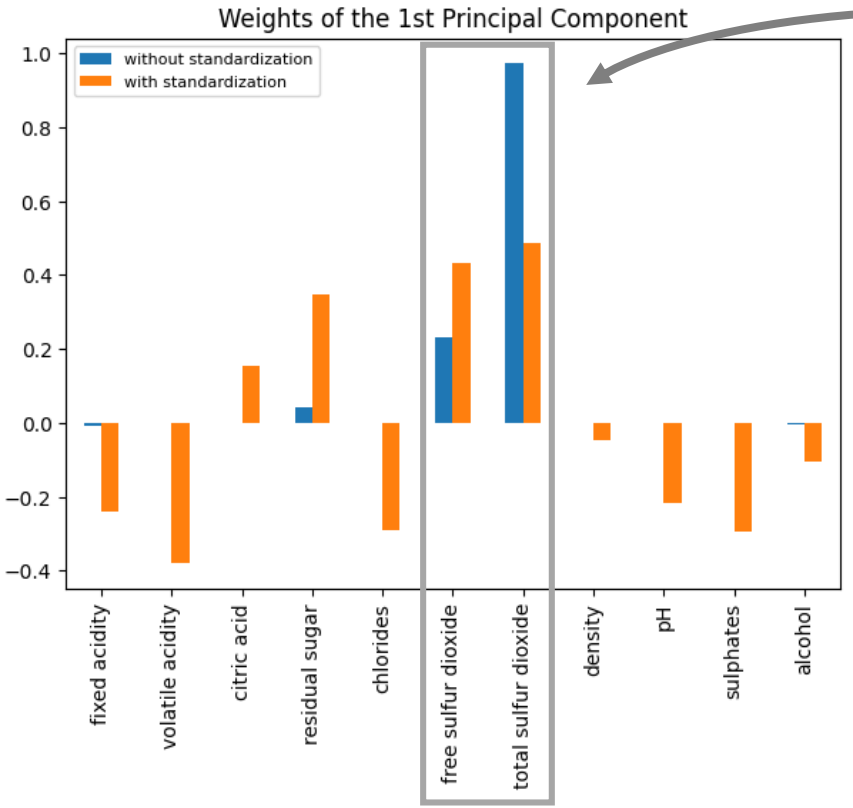
- 資料**有標準化**
  第一主成份的特徵組合(左圖橘色)，各特徵的權重分布較均衡，不偏向某些特定的特徵變數。

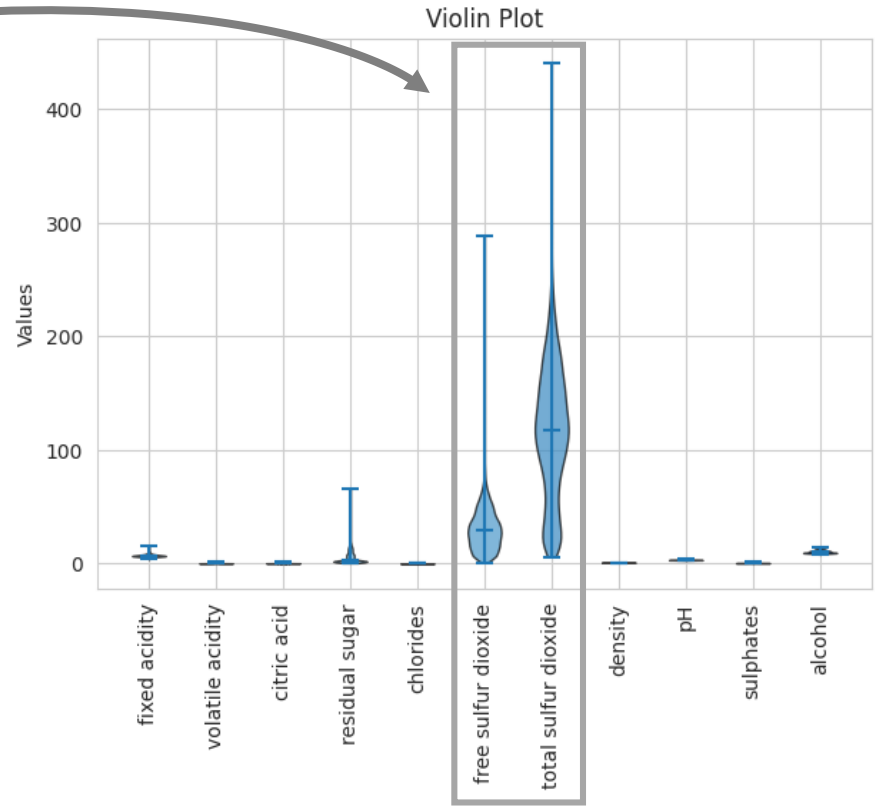
PCA 的資料前處理

【案例】

- 若特徵資料無標準化，直接進行 PCA，則，主成份的特徵權重，會受數據尺度的大小影響，會偏向關注尺度較大的特徵變數。

(因為，尺度越大變異相對容易越大。)

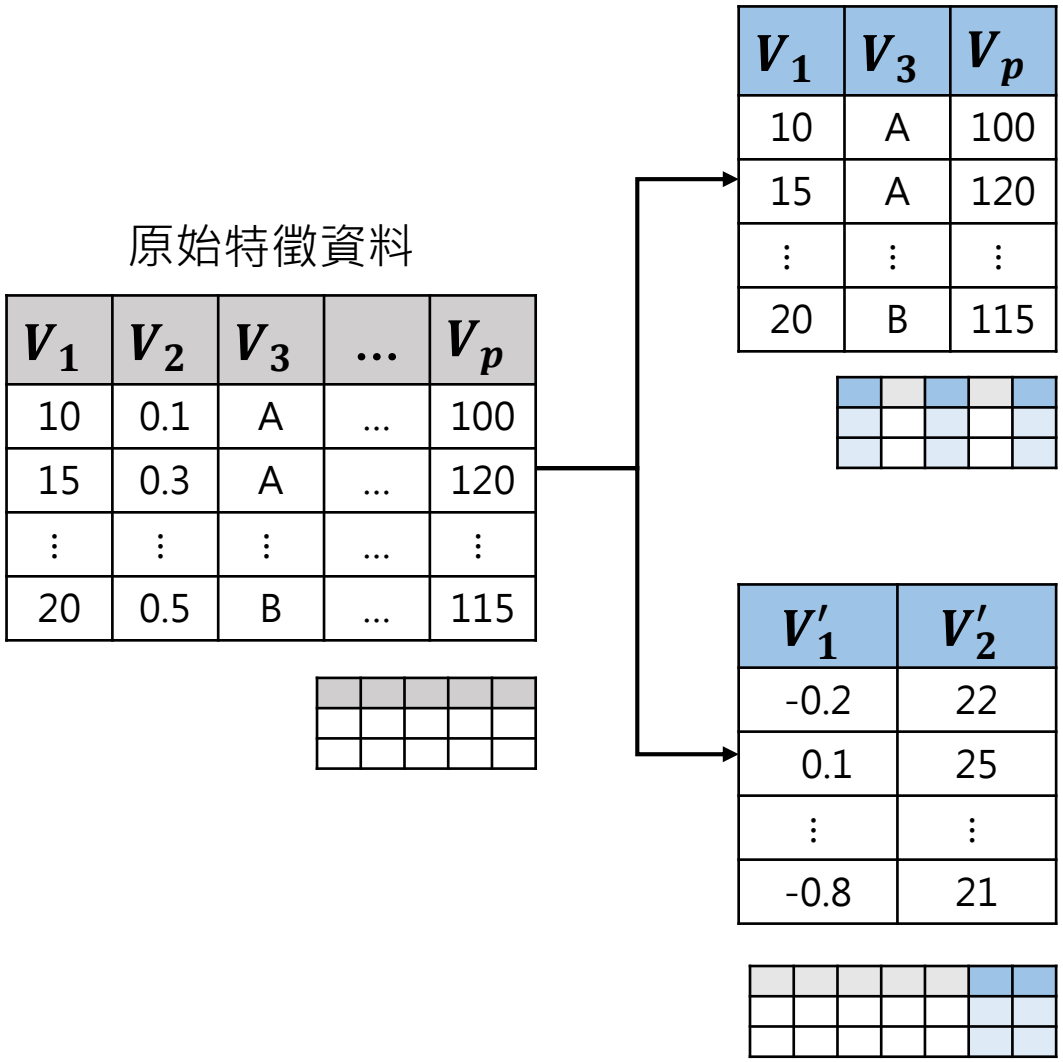




優化 ML 模型的策略

- ML 模型欠擬合與過擬合的資料面處理策略

特徵篩選與萃取



特徵篩選 Feature Selection

- 從原始特徵資料中，選取部分原始特徵變數。
- **目的**
  排除冗餘或不必要的資訊，提升 ML 模型訓練的效率與泛化能力。

特徵萃取 Feature Extraction

- 針對部分或全部原始的特徵變數，進行轉換創造新的特徵變數。
- **目的**
  降低特徵資料的維度，強化數據的重要資訊，去除冗餘的資訊，提升 ML 模型的表現。

優化ML模型的策略

	欠擬合 Underfitting 確保 ML 模型能學習到足夠的代表性資訊	**過擬合** Overfitting 避免 ML 模型過度學習噪音或冗餘的資訊
特徵篩選 Feature Selection	避免過度地刪除特徵變數。 例如： ● 放寬篩選閾值(Threshold) ● 降低正則化的強度 ⋮	濾除冗餘或無關的特徵變數。 例如： ● 移除高度相關性的特徵 ● 移除高度共線性的特徵 ⋮
特徵萃取 Feature Extraction	改變或組合現有的特徵變數，添加新的代表性資訊。 例如： ● 增加主成份分析的新特徵 ● 增加特徵分箱的類別型特徵 ● 增加非線性轉換出更複雜的關係 ⋮	降低特徵資料的維度，去除噪音且強化重要的訊息。 例如： ● 主成份分析 PCA ● 線性判別分析 LDA ⋮

數據是燃料，**特徵工程**是精煉技術！

CONTACT ME

吳彥霖 yenlinwu1112@gmail.com