Day 7

非監督式學習 降維 (1)

[全民瘋AI系列]

第12屆 iT邦幫忙 鐵人賽

Day 7 學習目標

01 **降維觀念**

何謂降維? 降維有什麼優點?

02 **常見兩種降維方法**

PCA & T-sne

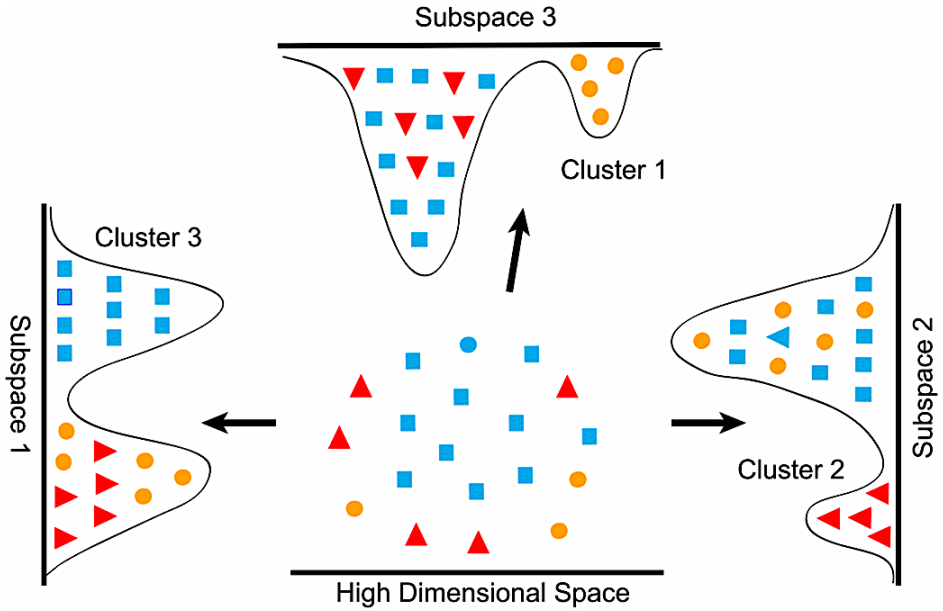
Part 1

非監督式學習 降維
觀念講解

第12屆 iT邦幫忙 鐵人賽

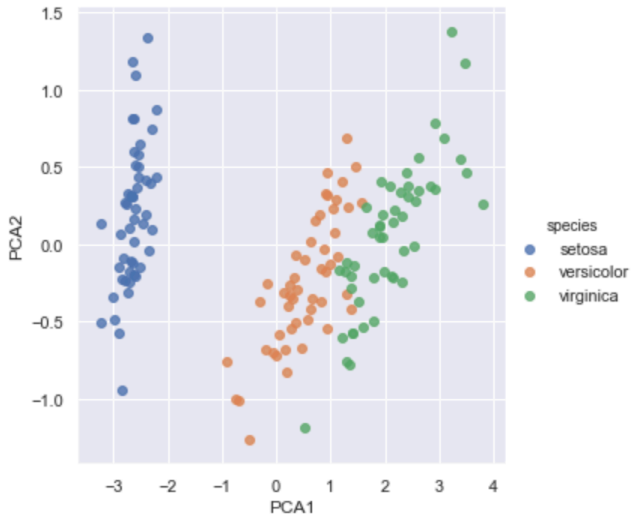
降維 (Dimension Reduction)

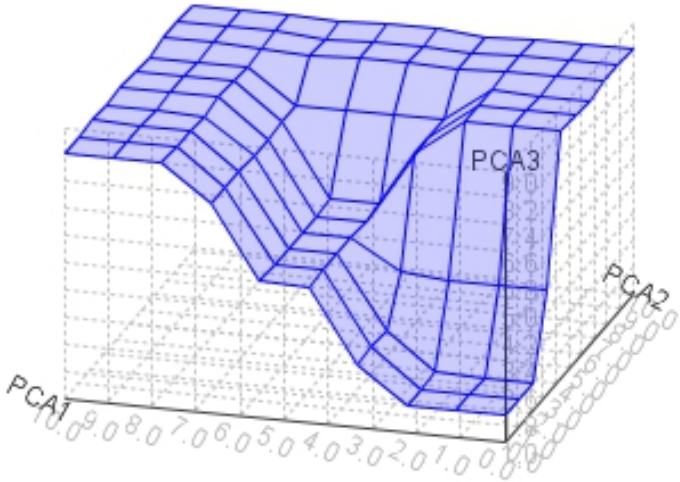
顧名思義，就是原本的Data處於在一個比較高的維度作標上，我們希望找到一個低維度的作標來描述它，但又不能失去Data本身的特質。



為什麼要降維?

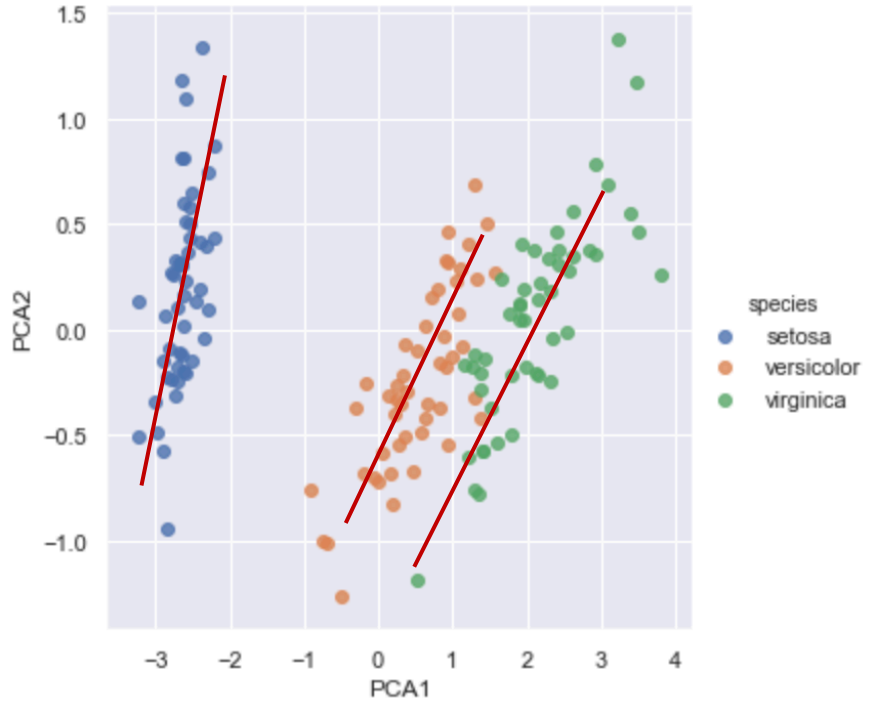
- 壓縮資料，減少計算資源。
- 降維可以幫助資料視覺化。





Principal component analysis (PCA)

- 將一個具有 n 個特徵空間的樣本，轉換為具有 k 個特徵空間的樣本，其中 k < n
- PCA的目的是把高維的點頭影到低維的空間上，並且低維度的空間保有高維空間中大部分的性質。
- PCA只允許做線性的轉換。



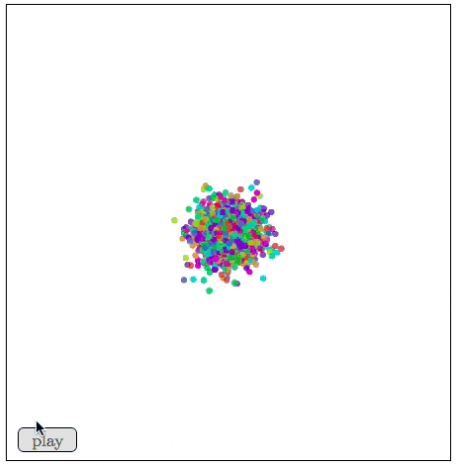
PCA的主要步驟

1. 先求出所有資料點中心μ，也就是將每一個資料點做平均
2. 將每一個資料點減去μ，這步驟是將資料點平移，平移後原點是所有點的中心
3. 計算每一個feature的variance
4. 把每一個值都除以variance

PCA 小結

- PCA 是個相當直觀且有效的降維方式，不過在三維轉換為二維時我們可以看到，有些數據的集群完全被搗成一團。這跟 PCA 的原理有關，因為 PCA 是對資料求共變異數矩陣，在進行奇異值分解。因此會被資料的差異性影響，無法很好表現**相似性以及分佈**。

- 且 PCA 是一種線性降維的方式，但如果特徵與特徵間的關聯是非線性關係的話，用 PCA 可能會導致**欠擬合（underfitting）**的情形發生。



T-Distributed Stochastic Neighbor Embedding (t-SNE)

- 目標跟PCA是一樣的，希望把高維的資料投影到低維中，並且保留高維中的點與點之間的關係與特性。兩者不同的點在於t-SNE允許非線性的轉換。

- t-SNE使用了更複雜的公式來表達高維與低維之間的關係。主要是將高維的數據用高斯分佈的機率密度函數近似，而低維數據的部分使用 t 分佈的方式來近似。

高維度

$$p_{ij} = \frac{\exp\frac{\left(-\|x_i - x_j\|^2\right)}{2\sigma^2}}{\sum\frac{\exp(-\|x_k - x_l\|^2)}{2\sigma^2}}$$

低維度

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum\left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

兩個分佈之間的相似度

- 求算兩個分佈之間的相似度，經常用 KL 距離（Kullback-Leibler Divergence）來表示，也叫做相對熵（Relative Entropy）。

$$KL\left(P||Q\right)) = D\left(P||Q\right) = \sum P\left(x\right) \log \frac{P(x)}{Q(x)}$$

9

t-SNE 小結

- 利用此方法降維後原本相近的點依然相近，反之原本距離遠的投影後依然保持遠的距離。
- t-SNE允許非線性的轉換，因此有機會在原本分開的三群在做完投影後依然是開的。

t-SNE 不適用於新資料

PCA 降維可以適用新資料，可呼叫transform() 函式即可。而 t-SNE 則不行。因為演算法的關係在 scikit-learn 套件中的 t-SNE 演算法並沒有transform() 函式可以呼叫。

PCA & t-SNE 整理

PCA和t-SNE是兩個不同降維的方法，PCA的優點在於簡單若新的點要映射時直接代入公式即可得出降維後的點。若t-SNE有新的點近來時我們沒有去計算新的點和舊的點之間的關係因此我們無法將新的點投影下去。t-SNE的優點是可以保留原本高維距離較遠的點降維後依然保持遠的距離，因此這些群降維後依然保持群的特性。

PCA允許線性的轉換

t-SNE允許非線性的轉換

Thanks

PRESENTED BY 10程式中

第12屆 iT邦幫忙 鐵人賽