Day 60

非監督式機器學習

PCA 觀察：使用手寫辨識資料集

出題教練

周俊川 / 陳明佑

知識地圖 非監督學習

CUPOY

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

前處理 Processing → 探索式數據分析 Exploratory Data Analysis → 特徵工程 Feature Engineering → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

非監督式學習
Unsupervised Learning

分群 Clustering

降維 Dimension Reduction

非監督學習
Unsupervised learning

非監督簡介

分群
Clustering

- K-平均算法 K-Mean
- 階層分群法 Hierarchical Clustering

降維
Dimension Deduction

- 主成分分析PCA(Principal components analysis)
- T 分佈隨機近鄰嵌入 t-SNE

本日知識點目標

- 知道手寫資料集的來源與用途
- 知道為什麼使用手寫資料集來觀察主成分分析的降維效果

註 因為非監督模型的效果，較難以簡單的範例看出來，所以非監督偶數日提供的檢視工具，僅供觀察非監督模型的效果，與後續其他部分及程式寫作無關，同學只要能感受到這些非監度模型的效果即可，不用執著於完全搞懂該章節所使用的工具

手寫辨識資料集 (MNIST) (1 / 2)

CUPOY

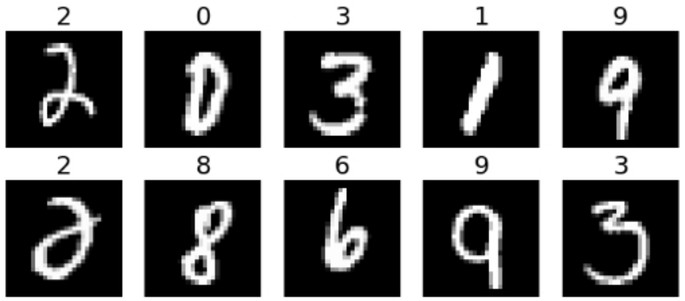
- 手寫辨識資料集的來源
  - 手寫辨識資料集 (MNIST, Modified National Institute of Standards and Technology databas) 原始來源的NIST，應該是來自於美國人口普查局的員工以及學生手寫所得，其中的 Modified 指的是資料集為了適合機器學習做了一些調整：將原始圖案一律轉成黑底白字，做了對應的抗鋸齒的調整，最後存成 28x28 的灰階圖案，成為了目前最常聽到的基礎影像資料集

(圖：原始的NIST圖案-來源: wiki)

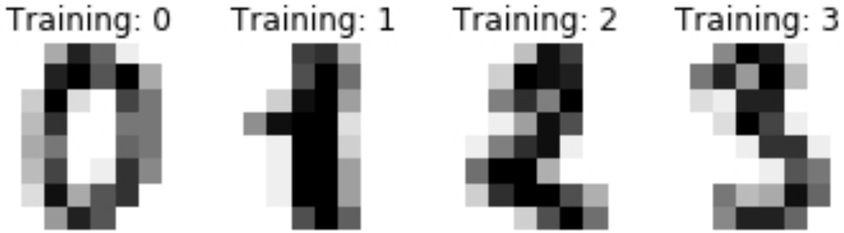
手寫辨識資料集 (MNIST) (2 / 2)

CUPOY

- sklearn 中的手寫辨識資料集
  - 與完整的MNIST不同，sklearn為了方便非深度學習的計算，再一次將圖片的大小壓縮到 8*8 的大小，雖然仍是灰階，但就形狀上已經有點難以用肉眼辨識，但壓縮到如此大小時，每張手寫圖就可以當作64 (8*8=64) 個特徵的一筆資料，搭配一般的機器學習模型做出學習與預測



來源：原 MNIST 資料集
(取自Kaggle練習題)

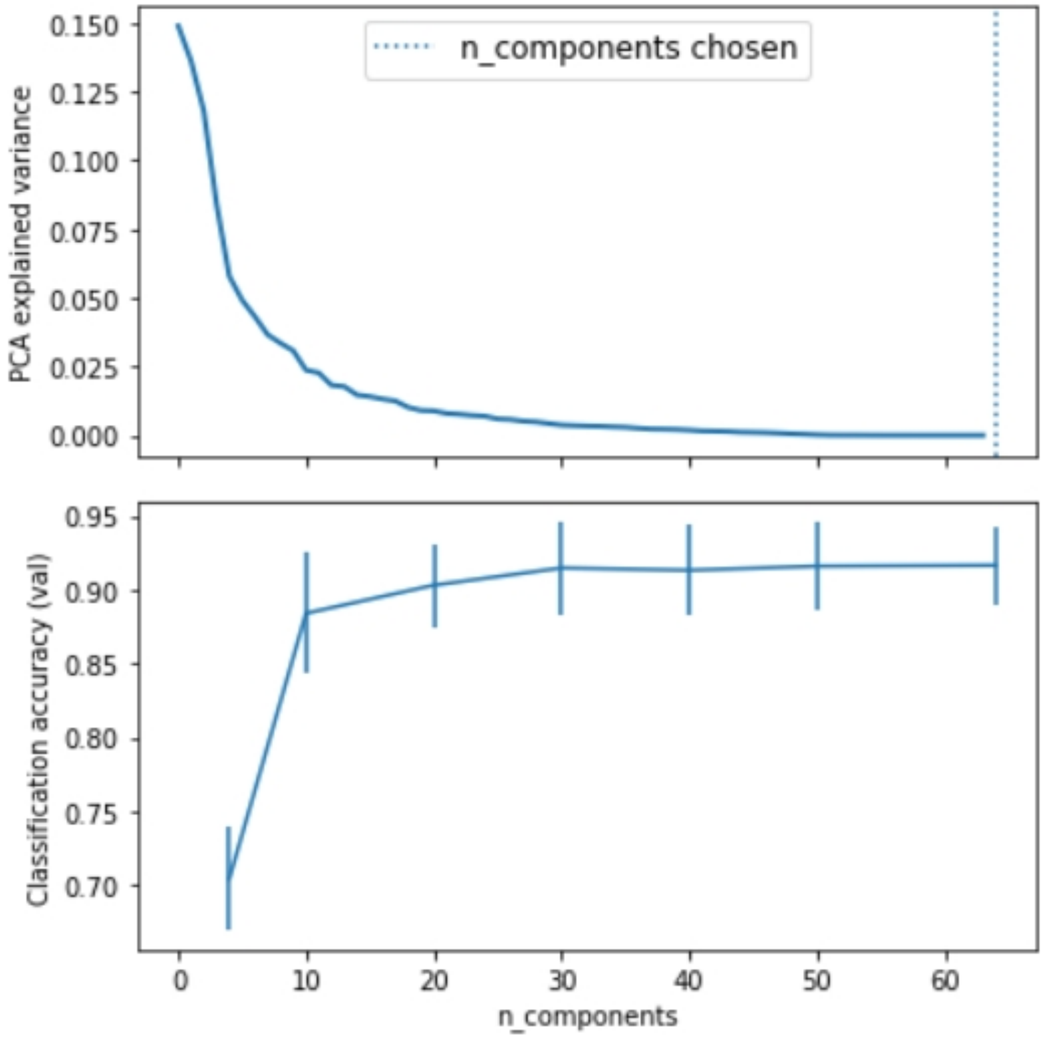


來源：sklearn 上的 MNIST 資料集

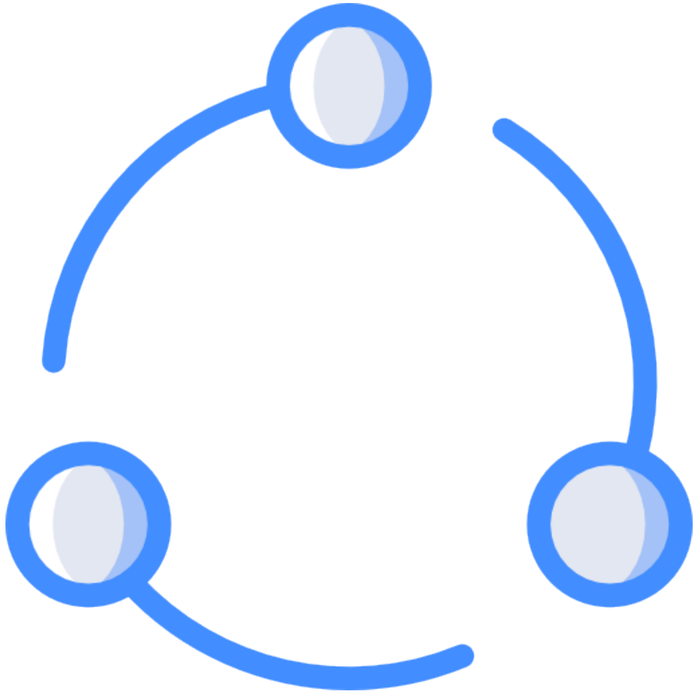
為什麼挑 MNIST 檢驗 PCA 的降維效果

CUPOY

- 高維度.高複雜性 / 人可理解的資料集
  - 由於 PCA 的強大，如果資料有意義的維度太低，則前幾個主成分就可以將資料解釋完畢
  - 使用一般圖形資料，維度又會太高，因此我們使用 sklearn 版本的 MNIST 檢驗 PCA，以兼顧內容的複雜性與可理解性
  - 由範例的折線圖可以看出來：前幾個維度就能解釋75%以上的變數



重要知識點複習



- 手寫資料集是改寫自手寫辨識集NIST的，為了使其適合機器學習，除了將背景統一改為黑底白字的灰階圖案，也將大小統一變更為 28*28

- 為了兼顧內容的複雜性與可理解性，我們使用圖形當中最單純的 sklearn 版 MNIST 作為觀察 PCA 效果的範例

解題時間
Coding Time

請跳出PDF至官網Sample Code & 作業
開始解題