Day 59

非監督式機器學習

降維方法 - 主成份分析

周俊川
出題教練

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

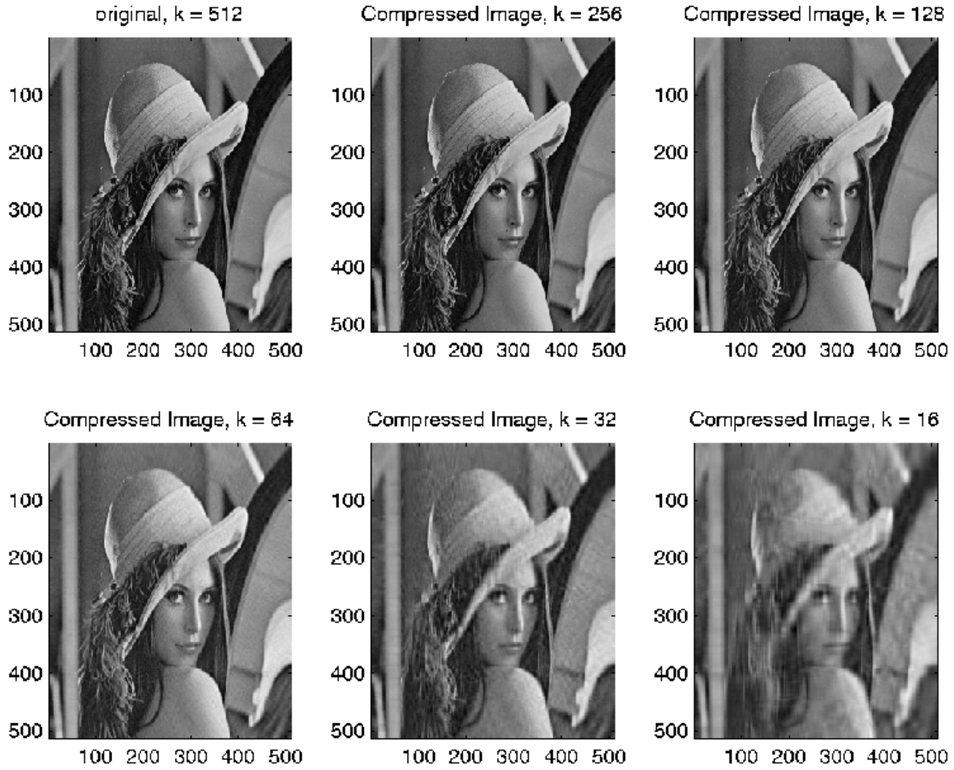
- 主成分分析PCA(Principal components analysis)
- T 分佈隨機近鄰嵌入 t-SNE

本日知識點目標

- 降低維度的好處，及其應用領域
- 主成分分析 (PCA) 概念簡介

爲什麼需要降低維度？壓縮資料

- 有助於使用較少的 RAM 或 disk space，也有助於加速 learning algorithms
- 影像壓縮
  - 原始影像維度爲 512, 在降低維度到 16 的情況下，圖片雖然有些許模糊，但依然保有明顯的輪廓和特徵

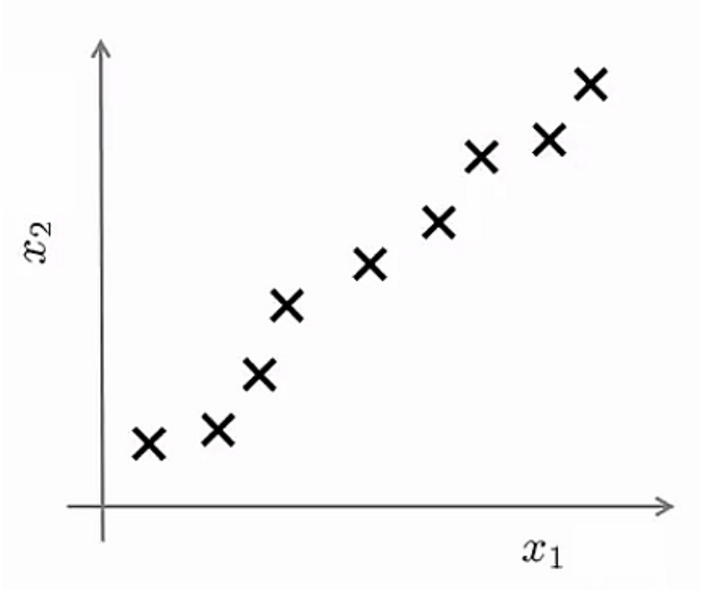


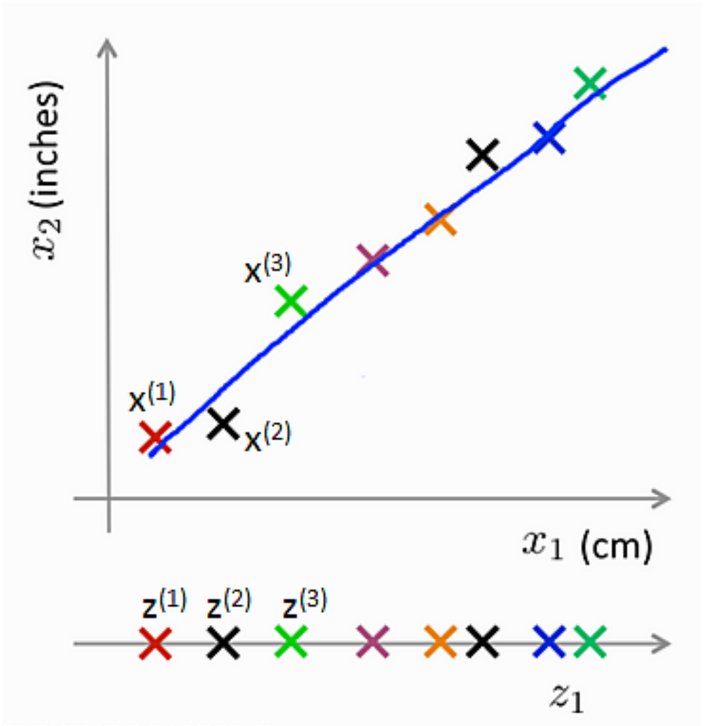
圖片來源：fourier.eng.hmc.edu

爲什麼需要降低維度？特徵組合及抽象化

CUPOY

- 壓縮資料可進而組合出新的、抽象化的特徵，減少冗餘的資訊。
- 左下圖的 x1 和 x2 高度相關，因此可以合併成 1 個特徵 (右下圖)。
  - 把 x(i) 投影到藍色線，從 2 維降低爲 1 維。



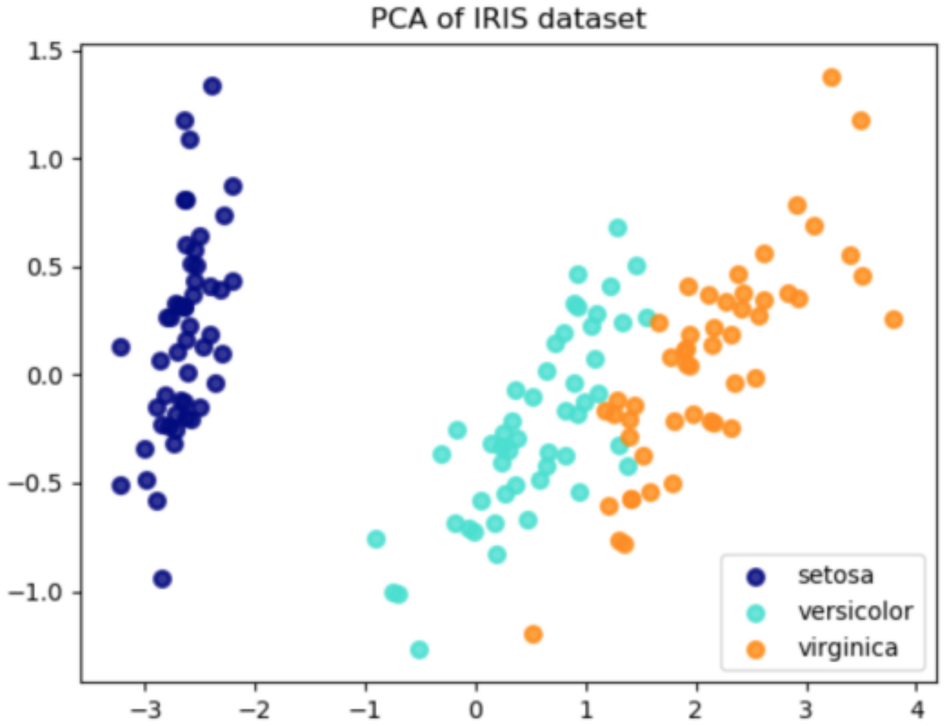


圖片來源：murphymind.blogspot

爲什麼需要降低維度？資料視覺化

CUPOY

- 特徵太多時，很難 visualize data, 不容易觀察資料。
- 把資料維度 (特徵) 降到 2 到 3 個 , 則能夠用一般的 2D 或 3D 圖表呈現資料



圖片來源：scikit-learn

主成份分析 (PCA)

CUPOY

- 實務上我們經常遇到資料有非常多的 features, 有些 features 可能高度相關，有什麼方法能夠把高度相關的 features 去除？
- PCA 透過計算 eigen value, eigen vector, 可以將原本的 features 降維至特定的維度
  - 原本資料有 100 個 features，透過 PCA，可以將這 100 個 features 降成 2 個 features
  - 新 features 為舊 features 的線性組合

新 features 彼此不相關

CUPOY

$$Z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n$$
$$Z_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n$$

Uncorrelated

$$\vdots$$

$$Z_n = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n$$

應用：加速監督式學習

CUPOY

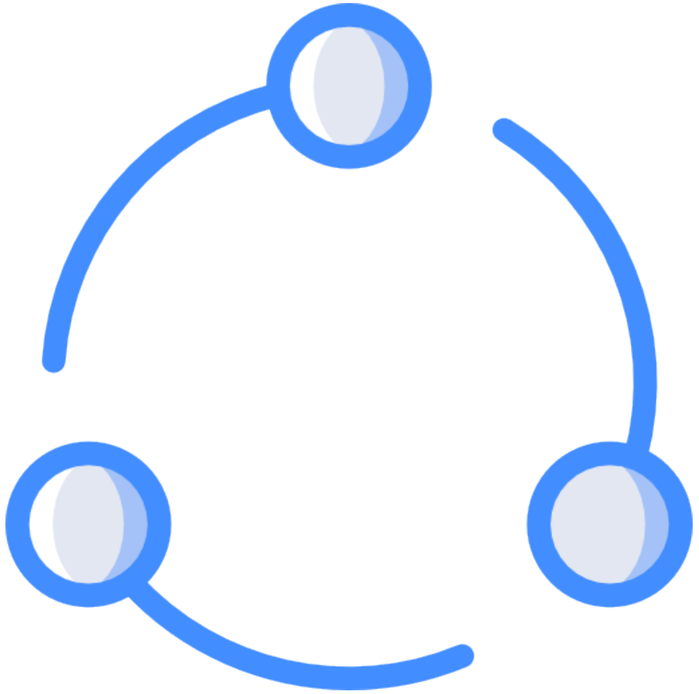
- 組合出來的這些新的 features 可以進而用來做 supervised learning 預測模型
- 以判斷人臉爲例，最重要的特徵是眼睛、鼻子、嘴巴，膚色和頭髮等都可以捨棄，將這些不必要的資訊捨棄除了可以加速 learning，也可以避免一點 overfitting。

PCA 應用在監督式學習的注意事項

CUPOY

- 不建議在早期時做，否則可能會丟失重要的 features 而 underfitting .
- 可以在 optimization 階段時，考慮 PCA，並觀察運用了 PCA 後對準確度的影響

重要知識點複習



- 降低維度可以幫助我們壓縮及丟棄無用資訊、抽象化及組合新特徵、呈現高維數據。常用的算法爲主成分分析。
- 在維度太大發生 overfitting 的情況下，可以嘗試用 PCA 組成的特徵來做監督式學習，但不建議一開始就做。

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題