Machine Learning Example 2：降維度方法

作者：廖柄煘

摘要

隨著科技發展，資料大量的資訊化形成巨量資料，多維度與大量資料容易造成可閱讀性低，所以降維度技術日趨重要。主成份分析(PCA)、線性判別分析(LDA)、局部線性嵌入（LLE）是常見降維度技術。Iris Data Set （鳶尾屬植物數據集）、Optical Recognition of Handwritten Digits Dataset、Wine 資料集，視資料特性採取合適方法，透過降維度方法能讓龐大的資料集顯示出過去未知的有用數據。

關鍵詞:降維度技術、PCA、LDA、LLE

一、緒論

1.1 動機

Iris Data Set （鳶尾屬植物數據集）包含三種鳶尾花的品種（Iris Setosa, Iris Versicolour, Iris Virginica)，另外還有萼片長度(cm)、萼片寬度(cm)、花瓣長度(cm) 、花瓣寬度(cm) 4 個特徵差別，如能進行有效的分類能達資料視覺化。Optical Recognition of Handwritten Digits Dataset 由於相似字間筆劃結構非常類似，因此必須使用能夠精確顯示些微差異的特徵，隨著電腦化的發展，如能將手寫字進行光學辨識，開發出文字辨識系統，當某個使用者一再的輸入手寫字時，系統即能逐漸地學習到這個使用者的書寫風格，能加速數位化的工作、降低人工成本。本研究葡萄酒資料集由 13 種化學原料所組成，可分成 3 種酒的類別。

1.2 目的

目前已許多降維度技術被廣泛運用，主要包含線性及非線性，兩個常被使用的線性方法為主成份分析(PCA)、線性判別分析(LDA)。然而線性降維度方法不適合處理非線性、彎曲的資料，因此許多非線性的方法提出來，如局部線性嵌入（LLE）、Isomap 等方法，採取合適的方法不僅能達到資料視覺化，可以進一步分析找出隱藏在資料中有用的訊息。

機器學習領域中的降維度技術就是指採用某種映射方法，將原高維度空間中的數據點映射到低維度的空間中。降維度的本質是學習一個映射函數 f : x->y，其中 x 是原始數據點的表達，目前最多使用向量表達形式。 y 是數據點映射後的低維度向量表達，通常 y 的維度小於 x 的維度。f 可能是顯式或隱式、線性的或非線性的。

Principal Component Analysis(PCA)是最常用的線性降維度方法，它的目標是通過某種線性投影，將高維度的數據映射到低維度的空間中表示，並期望在所投影的維度上數據的方差最大，以此使用較少的數據維度，同時保留住較多的原數據的特性。

如果把所有的點都映射到一起，那麼幾乎所有的資訊（如點和點之間的距離關係）都遺失了，而如果映射後方差盡可能的大，那麼數據點則會分散開來，以此來保留更多的資訊。可以證明，PCA 是丟失原始數據資訊最少的一種線性降維度方式。

Linear Discriminant Analysis（又稱Fisher Linear Discriminant)是一種有監督的（supervised）線性降維度方法。與PCA保持數據資訊不同，LDA是為了使得降維度後的數據點盡可能地容易被區分。

Locally linear embedding（LLE）是一種非線性降維算法，它能夠使降維後的數據較好地保持原有流形結構。LLE可以說是流形學習方法最經典的工作之一。很多後續的流形學習、降維度方法都與LLE有密切聯繫。

二、方法

本實驗程式的流程如下圖，首先使用 scikit-learn 的內建資料集 load_datasetName API 來讀取 Iris、Optical Digits、Wine 資料集。接著使用 PCA、LDA、LLE 三種降維度方法將資料縮減，最後透過 matplotlib 繪出圖形。

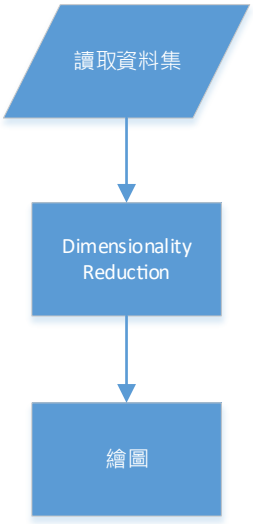

圖 1：程式流程

程式專案資料夾有 Ex2_Iris.py、Ex2_Optdigits.py、Ex2_Wine.py 三個檔案分別用來處理 Iris、Digit、Wine 資料集的程式，每個程式都是以 PCA、LDA、LLE 之順序來做 dimensionality reduction。

三、實驗

3.1 資料集

本實驗採用三個資料集：Iris Plants Database、Optical Recognition of Handwritten Digits、Wine recognition data。

3.1.1 Iris Plants Dataset

此資料集共有 150 筆資料，分為三類，每一類各 50 筆。每筆資料有五種屬性分別為 input attributes：萼片長度(cm)、萼片寬度(cm)、花瓣長度(cm)、花瓣寬度(cm)、以及用來分類的 output attribute（有 Setosa、Versicolour、Virginca)。

表 1：Iris 資料集欄位

萼片長度	萼片寬度	花瓣長度	花瓣寬度	類別
5.1	3.5	1.4	0.2	setosa
5.3	3.7	1.5	0.2	setosa
5	3.3	1.4	0.2	setosa
6.2	2.9	4.3	1.3	versicolor
5.1	2.5	3	1.1	versicolor
5.7	2.8	4.1	1.3	versicolor
6.3	3.3	6	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3	5.9	2.1	virginica

3.1.2 Optical Recognition of Handwritten Digits Dataset

此資料集共有 1797 筆資料，每一筆資料有一個 8x8 的圖形矩陣，也就是有共 64 個 input attributes，最後一個屬性為 output attribute，用來表示此圖形是哪一個數字，例如資料集的第一筆資料的圖形矩陣為下圖：

```
[ 0.  0.  5. 13.  9.  1.  0.]
[ 0.  0. 13. 15. 10. 15.  5.]
[ 0.  3. 15.  2.  0. 11.  8.]
[ 0.  4. 12.  0.  0.  8.  8.]
[ 0.  5.  8.  0.  0.  9.  8.]
[ 0.  4. 11.  0.  1. 12.  7.]
[ 0.  2. 14.  5. 10. 12.  0.]
[ 0.  0.  6. 13. 10.  0.  0.]
```

圖 2：數字 0 的矩陣

將上圖以灰階的方式畫出，會得到下圖：

圖 3：數字 0 矩陣的灰階圖

顯示的數字正好是第一筆資料 output 屬性的分類結果：數字 0。所有 input attributes 的數值範圍為 0~16；output attribute 數值範圍為 0~9。

表 2：Digits 資料筆數分布

數字	筆數
0	178
1	182
2	177
3	183
4	181
5	182
6	181
7	179
8	174
9	180

3.1.3 Wine Recognition Dataset

共有 178 筆資料，每一筆有 13 種化學分析名稱，分別為(1)Alcohol、(2)Malic acid、(3)Ash、(4)Alcalinity of ash、(5)Magnesium、(6)Total phenols、(7)Flavanoids、(8)Nonflavanoid phenols、(9)Proanthocyanins、(10)Color intensity、(11)Hue、(12)OD280/OD315 of diluted wines、(13)Proline，以及類別。

表 3：Wine 資料集欄位

類別	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr7	Attr8	Attr9	Attr10	Attr11	Attr12	Attr13
1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
1	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
2	11.87	4.31	2.39	21	82	2.86	3.03	0.21	2.91	2.8	0.75	3.64	380
2	12.07	2.16	2.17	21	85	2.6	2.65	0.37	1.35	2.76	0.86	3.28	378

2	12.37	1.63	2.3	24.5	88	2.22	2.45	0.4	1.9	2.12	0.89	2.78	342
2	12.04	4.3	2.38	22	80	2.1	1.75	0.42	1.35	2.6	0.79	2.57	580
3	12.86	1.35	2.32	18	122	1.51	1.25	0.21	0.94	4.1	0.76	1.29	630
3	12.88	2.99	2.4	20	104	1.3	1.22	0.24	0.83	5.4	0.74	1.42	530
3	12.81	2.31	2.4	24	98	1.15	1.09	0.27	0.83	5.7	0.66	1.36	560

表 4：Wine 資料筆數分布

類別	筆數
1	59
2	70
3	48

3.2 實驗設計（實驗如何進行、參數如何設定等）

將 PCA、LDA、LLE 三個方法的 n_components 參數設定為 2，就可以取得前兩大特徵向量，除了 LLE 有額外設定 n_neighbors 為 30 與 method 為 standard 之外，其餘的參數皆使用預設值。

3.3 實驗結果

3.3.1 Iris 資料集

PCA 前兩大特徵向量可解釋的變異量比例為：[0.92461621 0.05301557]

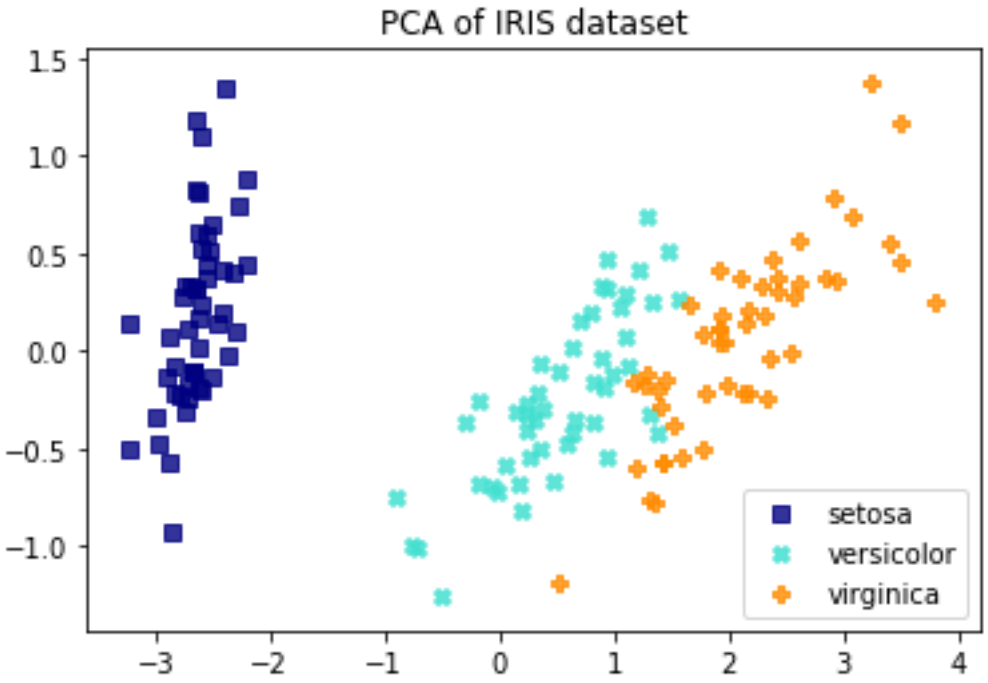

圖 4：Iris PCA 結果

LDA 前兩大特徵向量可解釋的變異量比例為：[0.99147248 0.00852752]

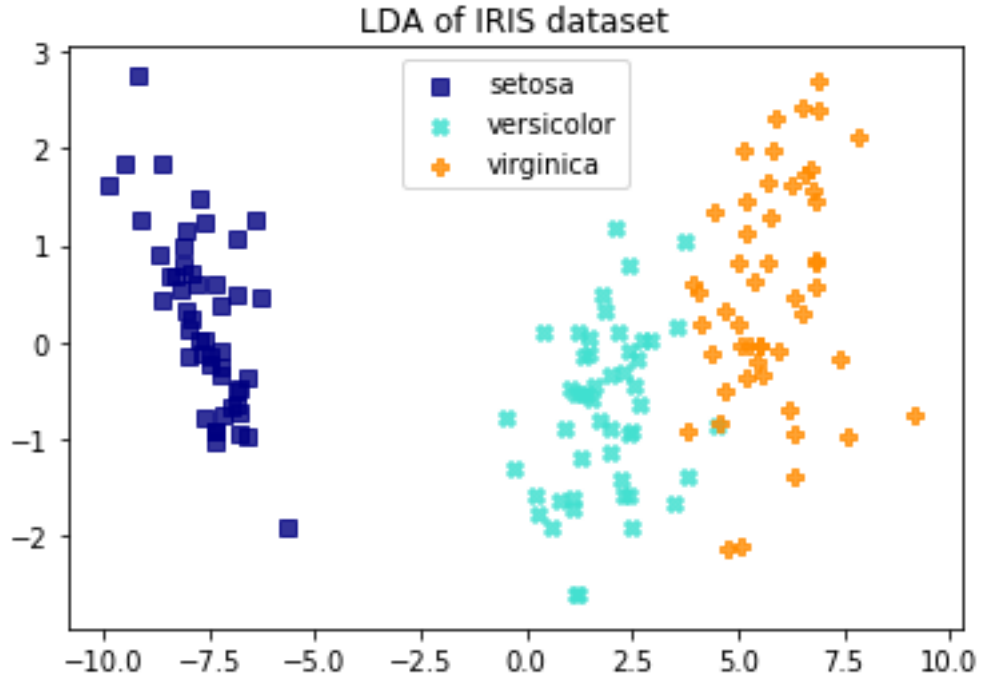

圖 5：Iris LDA 結果

LLE 重建錯誤率：1.7181595957426014e-05

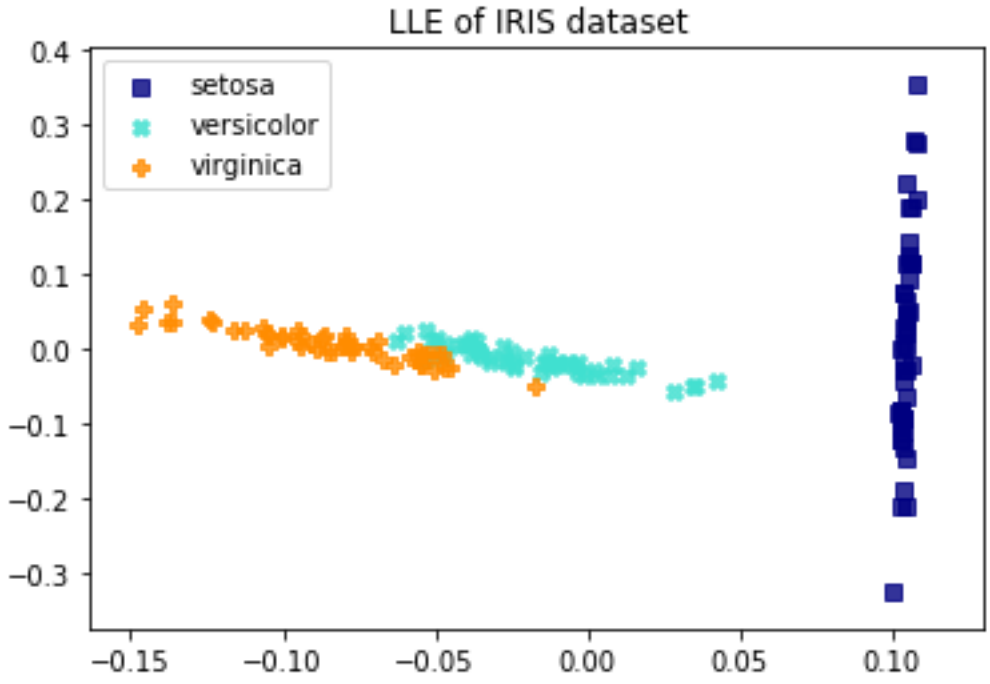

圖 6：Iris LLE 結果

3.3.2 Optical Recognition of Handwritten Digits 資料集

下圖呈現 Digits 資料集的前 400 筆資料的灰階圖。

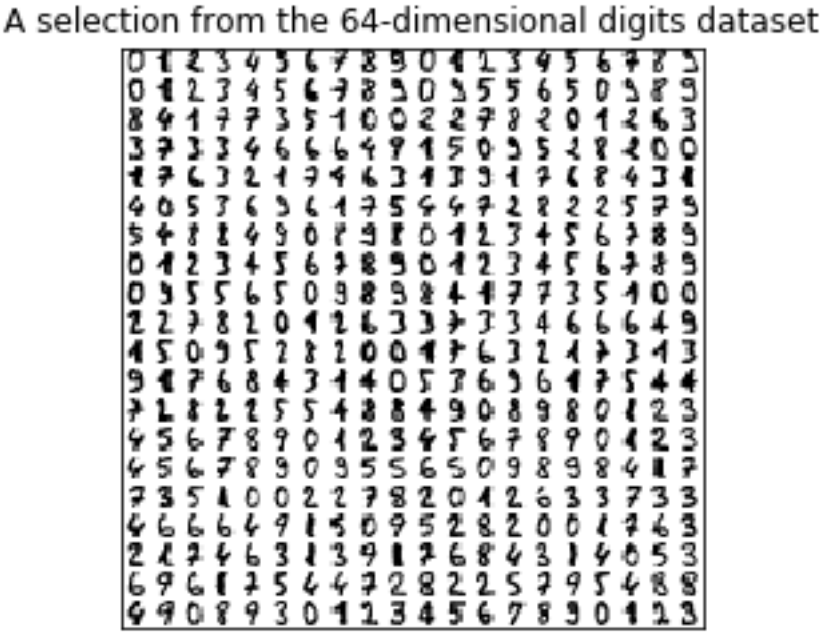

圖 7：資料集部分灰階圖

PCA 前兩大特徵向量可解釋的變異量比例為：[0.14890594 0.13618771]

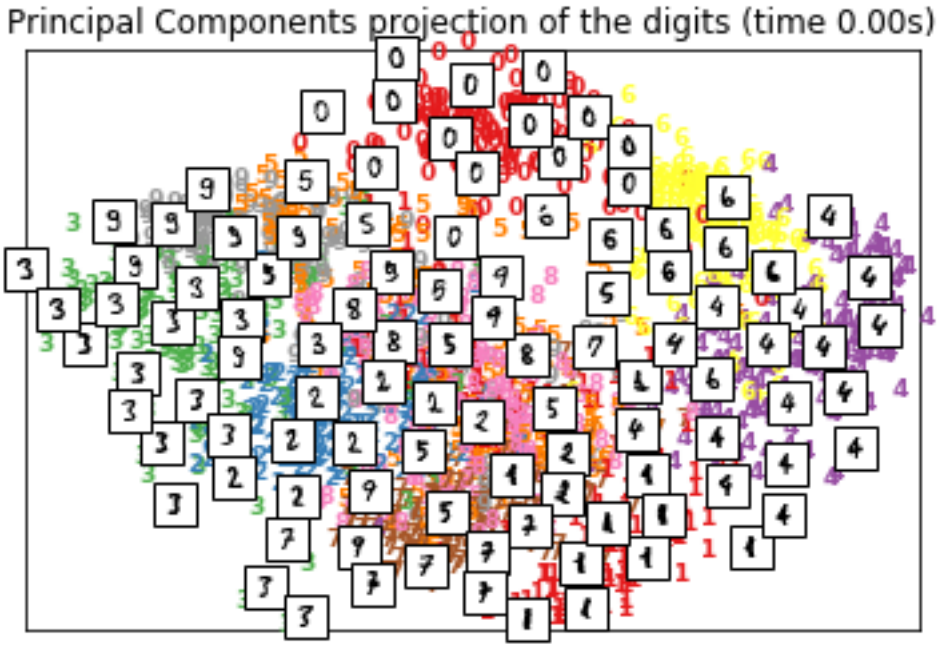

圖 8：Digit PCA 結果

LDA 前兩大特徵向量可解釋的變異量比例為：[0.28901578 0.18252926]

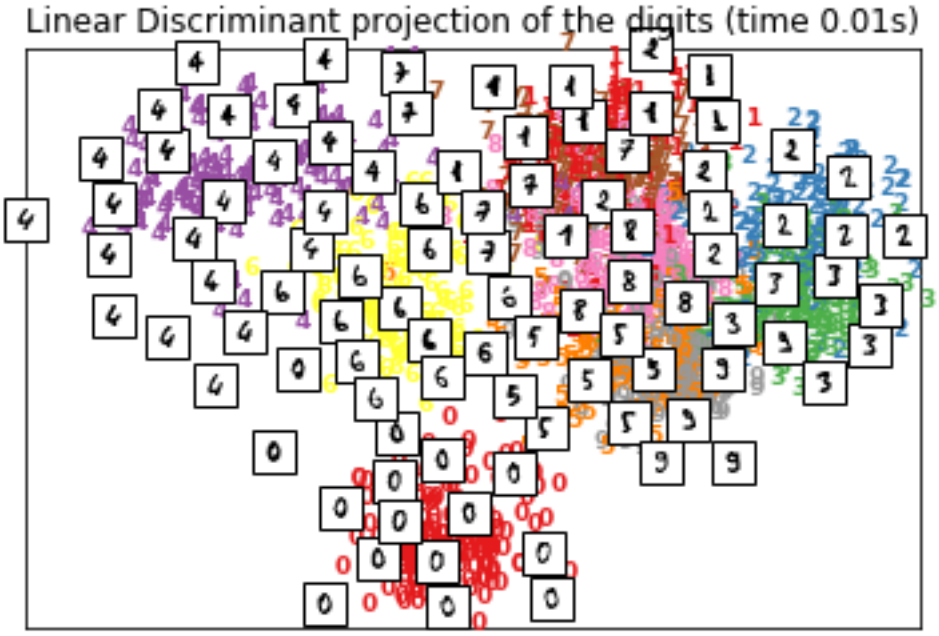

圖 9：Digit LDA 結果

LLE 重建錯誤率：5.16179e-07

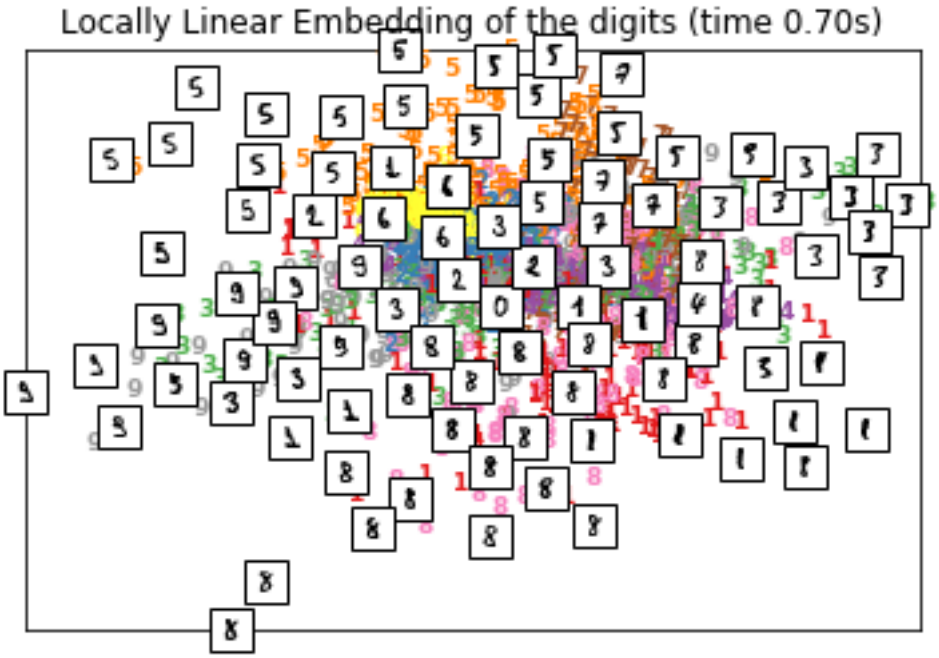

圖 10：Digit LLE 結果

3.3.3 Wine 資料集

PCA 前兩大特徵向量可解釋的變異量比例為：[0.99809123 0.00173592]

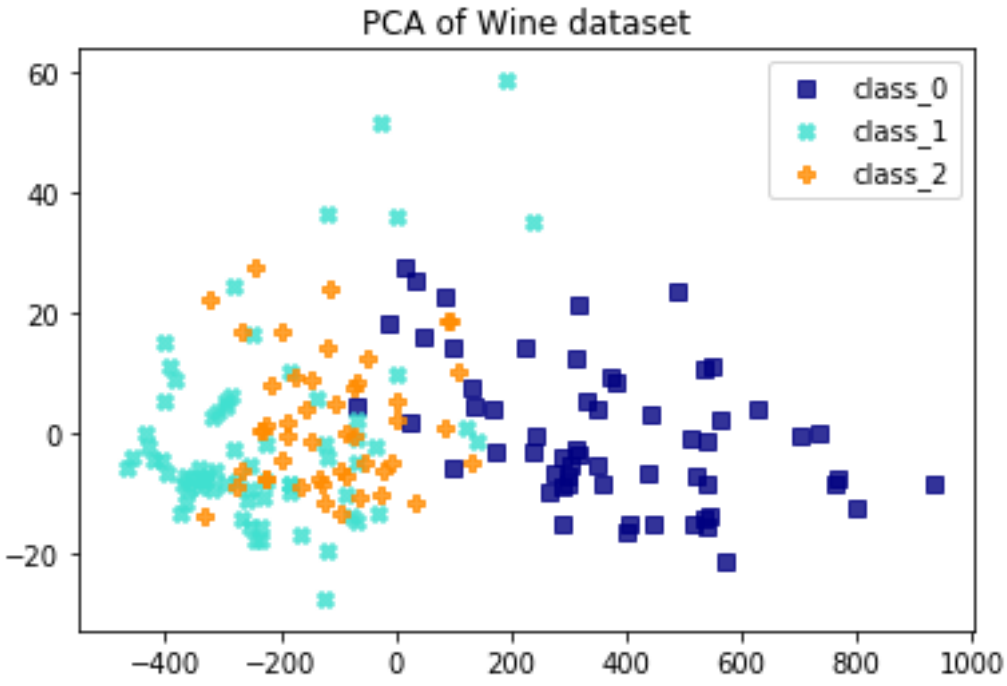

圖 11：Wine PCA 結果

LDA 前兩大特徵向量可解釋的變異量比例為：[0.68747889 0.31252111]

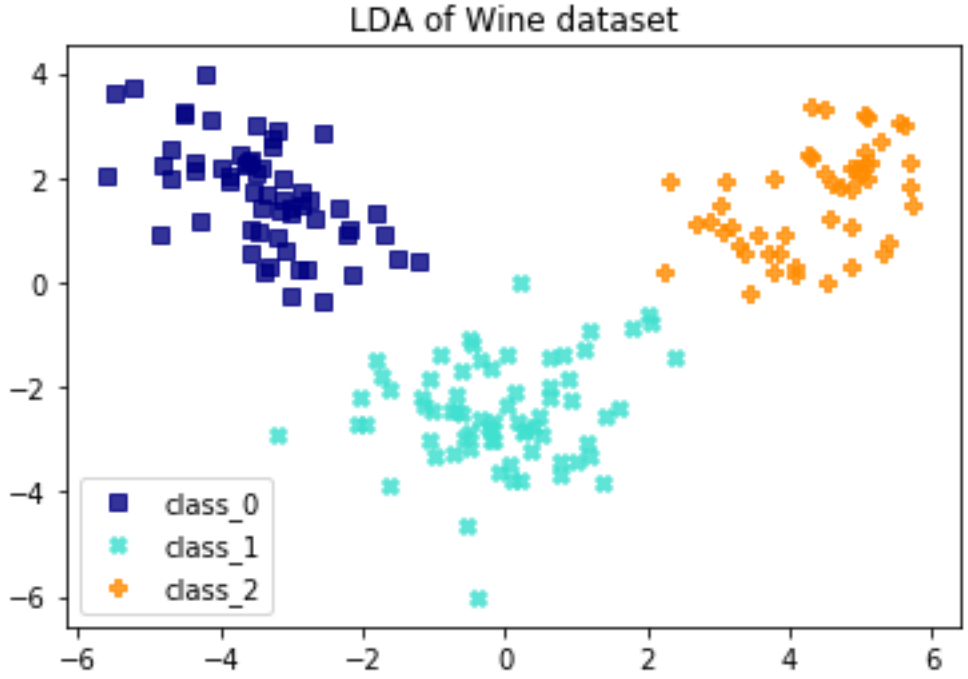

圖 12：Wine LDA 結果

LLE 重建錯誤率： 0.0001739967328243908

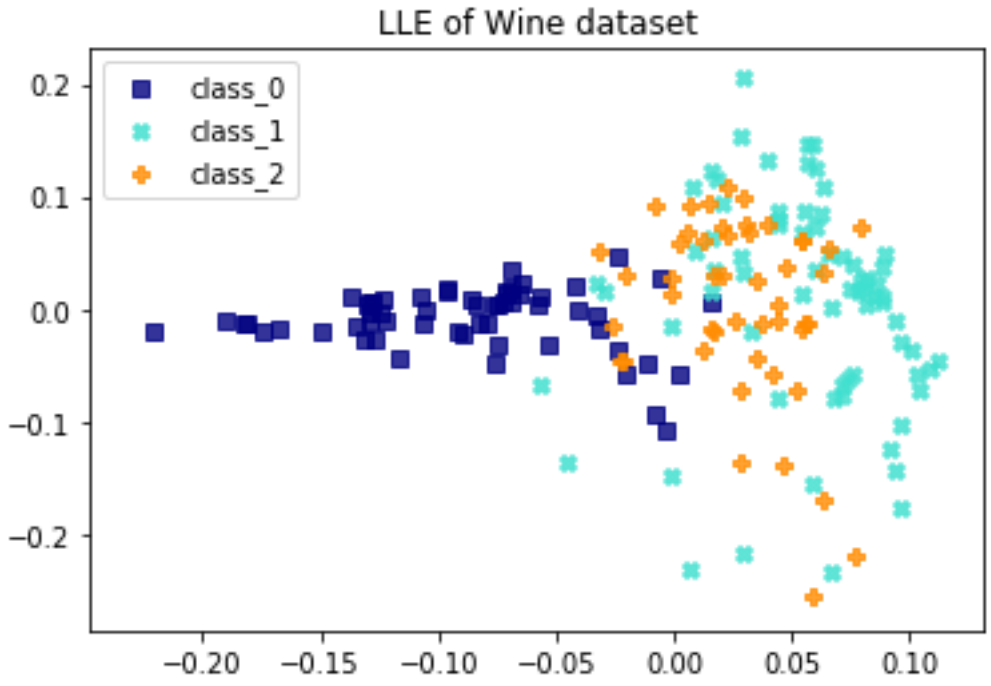

圖 13：Wine LLE 結果

四、結論

以 PCA、LDA、LLE 三種方法來處理 Iris、Wine 兩種資料集時，使用前兩大的特徵向量約可代表 98%以上的原始資料。而 Hand written digits 使用 PCA 與 LDA 約可解釋 27%與 46%的原始資料，LLE 反而能解釋更多，這是因為 PCA 與 LDA 都是 Linear dimensionality reduction 而 LLE 為 Non-Linear dimensionality reduction。

Iris 資料集使用 LDA 及 PCA 降維度技術似乎差異不大，優過於 LLE。Wine 資料集使用 LDA 降維度技術似乎優於 PCA 及 LLE，Class 之間距離較遠彼此間分隔比較清楚。Optical Recognition of Handwritten Digits 資料集使用 LDA 降低資料的維度不只能提高辨識率，似乎優於 PCA 及 LLE。最後整體顯示 LDA 有最好的效果。