Dimension reduction

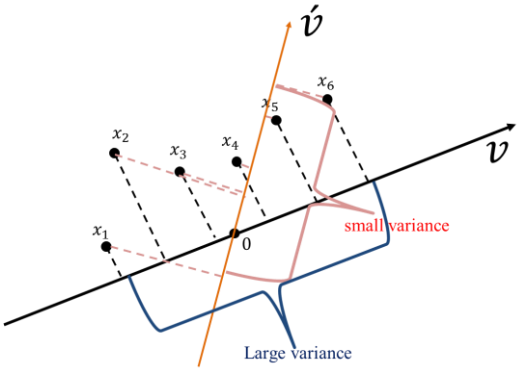
一、 原理

I. 主成分分析(Principal Component Analysis, PCA)

主成分分析在機器學習內被歸類成為降維(Dimension reduction)內特徵擷取(Feature extraction)的一種方法，降維就是希望資料的維度數減少，但整體的效能不會差異太多甚至會更好。

機器學習主要是希望用 PCA 達到 dimension reduction 的目的，主要是為了避免 Hughes 現象(Hughes Phenomenon)/ 維度詛咒(curse of dimensionality)。維度詛咒，預測/分類能力通常是隨著維度數(變數)增加而上生，但當模型樣本數沒有繼續增加的情況下，預測/分類能力增加到一定程度之後，預測/分類能力會隨著維度的繼續增加而減小。

主成份分析的基本假設是希望資料可以在特徵空間找到一個投影軸(向量)投影後可以得到這組資料的最大變異量。



- 假設有 n 的樣本點{x1, x2,…, xn}，xi∈Rd 投影軸為 v；投影後的點為{vT x1, vT x2,⋯, vT xn}。
- 變量的投影後的變異數:

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}^T\boldsymbol{x}_i)(\boldsymbol{v}^T\boldsymbol{x}_i)^T = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{v}) = \boldsymbol{v}^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v}$$

C 為共變異數矩陣(covariance matrix)

$$\boldsymbol{C}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T,\cdots\boldsymbol{x}_i = \begin{bmatrix} x_i^{(1)} \\ \vdots \\ x_i^{(d)} \end{bmatrix}$$

- 主成份分析則是在找投影向量讓投影後的資料變異量最大(最佳化問題)

$$\boldsymbol{v} = \underset{\boldsymbol{v}\in R^d,\|\boldsymbol{v}\|=1}{\arg\max}\ \boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v}$$

因為有限制式在，所以需要轉成 Lagrange 去做

$$f(\boldsymbol{v},\lambda) = \boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v} - \lambda(\|v\| - 1) = \boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v} - \lambda(\boldsymbol{v}^T\boldsymbol{v} - 1)$$

然後偏微分找解

$$\frac{\partial f(\boldsymbol{v},\lambda)}{\partial \boldsymbol{v}} = 0 \rightarrow 2\boldsymbol{C}\boldsymbol{v} - \mathbf{2}\lambda\boldsymbol{v} = \mathbf{0} \rightarrow \boldsymbol{C}\boldsymbol{v} = \lambda\boldsymbol{v}$$

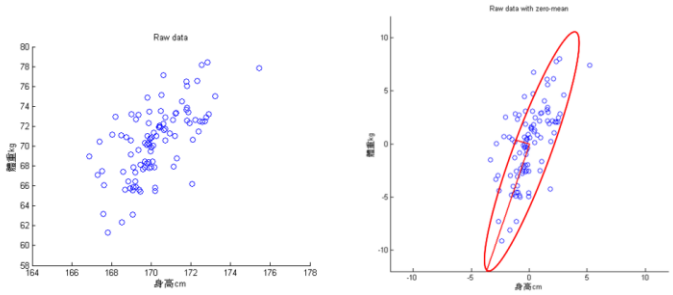
$$\frac{\partial f(\boldsymbol{v},\lambda)}{\partial \lambda} = 0 \rightarrow \boldsymbol{v}^T\boldsymbol{v} - \mathbf{1} = \mathbf{0} \rightarrow \|v\| = 1$$

找出來得解為:

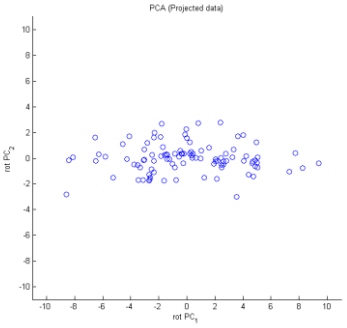
$$\boldsymbol{C}\boldsymbol{v} = \lambda\boldsymbol{v}$$
$$\|v\| = 1$$

- 就是解 C (共變異數矩陣)的特徵值(eigenvalue, λ)和特徵向量(eigenvector, v)，解出來的 eigenvalue 就是變異量(variance)，eigenvector 就是讓資料投影下去會有最大變異量的投影軸。

下左圖，經由 PCA 可以萃取出兩個特徵成分(投影軸，下圖右的兩條垂直的紅線，較長的紅線軸為變異量較大的主成份)。此範例算最大主成份的變異量為 13.26，第二大主成份的變異量為 1.23。



PCA 投影完的資料為下圖，從下圖可知，PC1 的變異足以表示此筆資料資訊。



此做法可以有效的減少維度數，但整體變異量並沒有減少太多，此例從兩個變成只有一個，但變異量卻可以保留(13.26/(13.26+1.23)= 91.51%)。

- 累積貢獻比率 (Cumulative Proportion)

從最重要的主成份開始往次重要的主成份到最不重要的主成份的變異量百分比累積。假設有 4 個變數(X1~X4)，所以萃取出的主成份會有(PC1~PC4)，累積貢

獻比率則是看前幾個主成份可以表是原始資料多少百分比的變異量，此範例只需取兩個主成份(PC1 和 PC2)則可以取得原資料的 100.000%的變異量，所以只需要 2 個主成份即可以取代原本 4 個變數然後進行後續的分類或是預測。

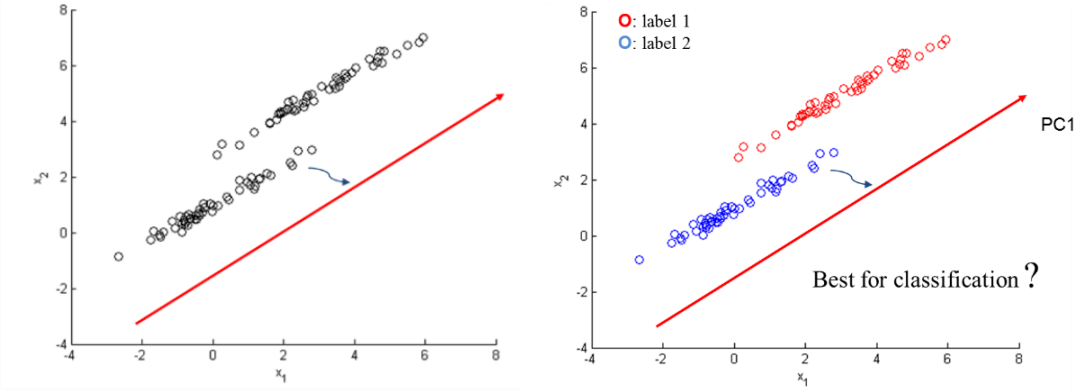
	PC1	PC2	PC3	PC4
變異量	2.987	1.013	2.220e-15	6.955e-17
變異量百分比	74.675%	25.325%	5.55e-14%	1.73875e-15
累積貢獻比率	74.675%	100.000%	100.000%	100%

II. 線性區別分析(Linear Discriminant Analysis，LDA）

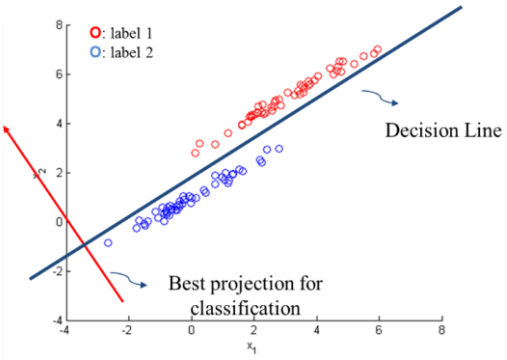
是一種 supervised learning，有些人拿來做降維(dimension reduction)，有些人拿來做分類(Classification)。

在降維度的方法上，LDA 是 PCA 延伸的一種方法。PCA 目標是希望找到投影軸讓資料投影下去後分散量最大化，但 PCA 不需要知道資料的類別。而 LDA 也是希望資料投影下去後分散量最大，但不同的是這個分散量是希望「不同類別之間的分散量」越大越好。所以 LDA 和 PCA 差異的部份，PCA 是無監督式(unsupervised learning)方法，LDA 多了這四個字(「不同類別」)是一種監督式(supervised learning)方法。

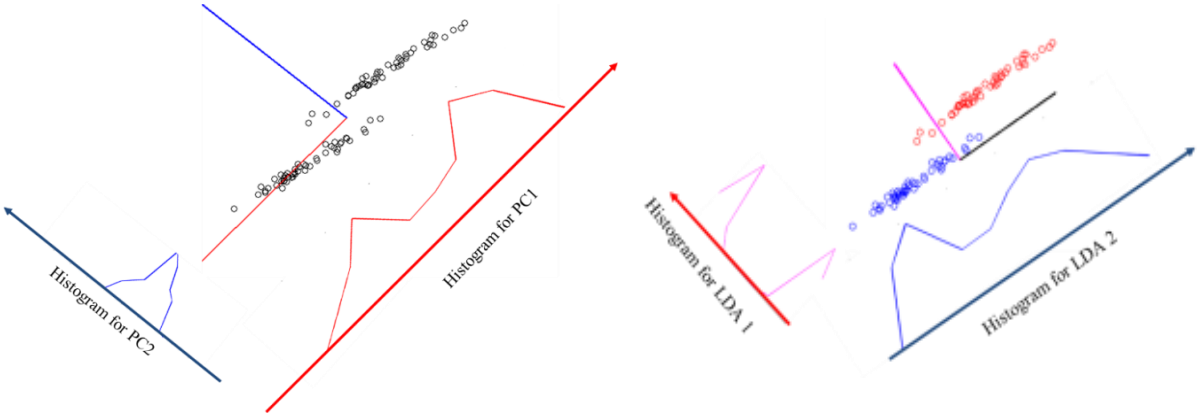
如果我們用 PCA 的方式，得到的投影軸是紅色的線(投影後變異量最大)。但如果我們的資料中有類別(監督式)的資訊，如果還是用 PCA 得到的投影軸對最後的分類問題能達到最大的幫助嗎?



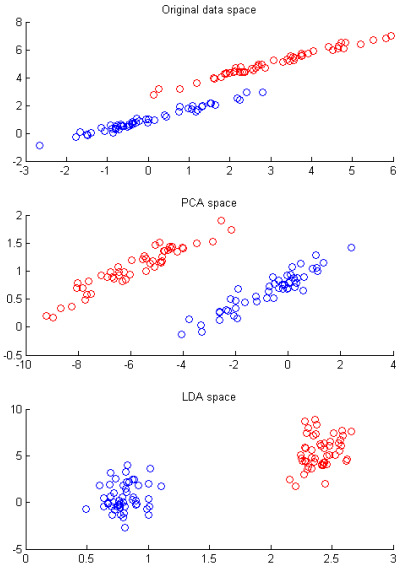
在 LDA 因為我們有了資料類別的訊息，所以可以把類別訊息考慮進算法內。LDA 部份則是希望類別跟類別之間的區別性越開越好，就是投影後紅的點跟藍色的點能越「區隔」開越好，最理想狀態是投影後用一個閾值(圖中的 decision line)就可以區隔兩類，所以這個方法叫做「線性區別分析」(下圖)。



PCA 部份，紅色線是 PCA 找到最大主成分，藍色是 PCA 找到次大的主成分，果投影後的資料可以區隔開的話，從直方圖可以很明顯看到會有兩個分佈。從 PCA 投影的資料很難看到只用一個成份可以完美區隔兩類。LDA 部份，紅色線是 LDA 找到最大成分，藍色是 LDA 找到次大的成分，從 LDA 投影的資料，可以清楚看到只需要最大的成分(紅色投影軸)就可以完美區隔兩類。



這邊我會把兩種方法投影後的資料一起呈現出來在下圖，所以可以清楚看到，LDA 的執行效果比 PCA 好。

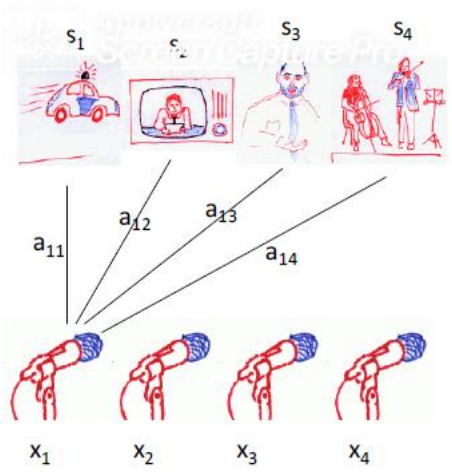


- 直觀上，對分類問題，LDA 是比 PCA 更好的一種特徵選取的技術，但有研究發現，對某些“圖像識別”的情況，使用 PCA 往往會得到較好的結果 (A. M. Martinez and A. C. Kak, “PCA Versus LDA.” IEEE Transactions on PatternAnalysis and Machine Intelligence, 23(2):228-233,2001)。

III. 獨立成分分析（ICA）

ICA 是找出構成信號的相互獨立部分(不需要正交)，對應高階統計量分析。ICA 理論認為用來觀測的混合數據陣 X 是由獨立元 S 經過 A 線性加權獲得。ICA 理論的目標就是通過 X 求得一個分離矩陣 W，使得 W 作用在 X 上所獲得的信號 Y 是獨立源 S 的最優逼近。

經典的雞尾酒宴會問題（cocktail party problem）。假設在 party 中有 n 個人，他們可以同時說話，我們也在房間中一些角落里共放置了 n 個聲音接收器（Microphone）用來記錄聲音。宴會過後，我們從 n 個麥克風中得到了一組數據 x (i) (x1(i) ,x2(i) ,....xn(i));i=1,...,m，i 表示採樣的時間順序，也就是說共得到了 m 組採樣，每一組採樣都是 n 維的。我們的目標是單單從這 m 組採樣數據中分辨出每個人說話的信號。



This is recorded by the microphones:
a linear mixture of the sources

$x_i(t) = a_{i1}*s_1(t) + a_{i2}*s_2(t) + a_{i3}*s_3(t) + a_{i4}*s_4(t)$

$$\Rightarrow \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \\ s_4(t) \end{bmatrix}$$

x A s

- ICA 的工作原理是假設組成信號源的子成分是非高斯的，並且在統計上相互獨立。

二、 參考資料

1. Class Handout, Lee, Chia-Jung professor, MDM64001, School of Big Data Management, Soochow University
2. 機器/統計學習:主成分分析(Principal Component Analysis, PCA) https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8-%E7%B5%B1%E8%A8%88%E5%AD%B8%E7%BF%92-%E4%B8%BB%E6%88%90%E5%88%86%E5%88%86%E6%9E%90-principle-component-analysis-pca-58229cd26e71
3. 機器學習: 降維(Dimension Reduction)- 線性區別分析(Linear Discriminant Analysis) https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E9%99%8D%E7%B6%AD-dimension-reduction-%E7%B7%9A%E6%80%A7%E5%8D%80%E5%88%A5%E5%88%86%E6%9E%90-linear-discriminant-analysis-d4c40c4cf937
4. Cocktail Party Problem - Georgia Tech - Machine Learning https://www.youtube.com/watch?v=T0HP9cxri0A