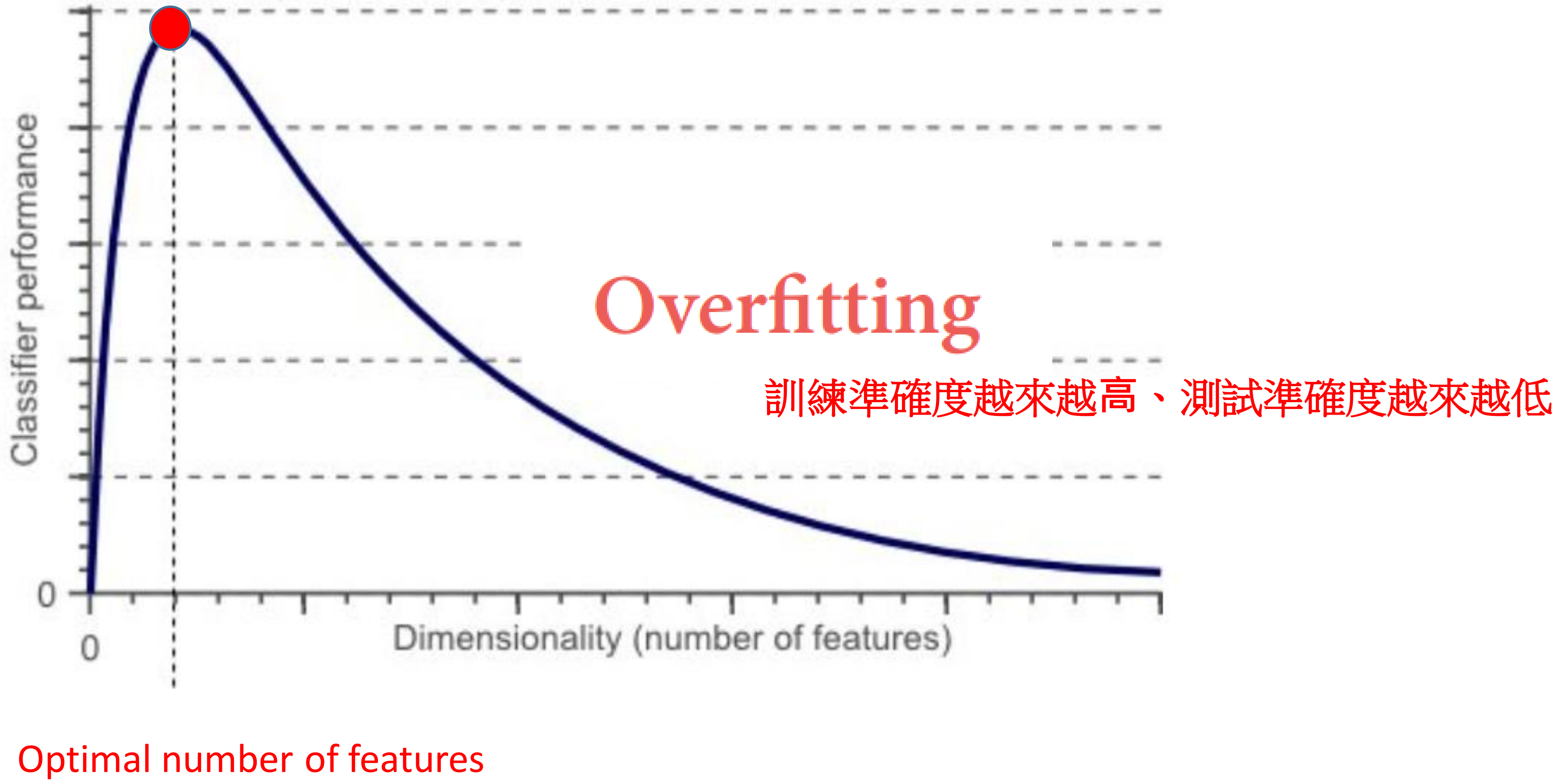
主成分分析

Principal Components Analysis

維度災難 (Curse of Dimensionality)

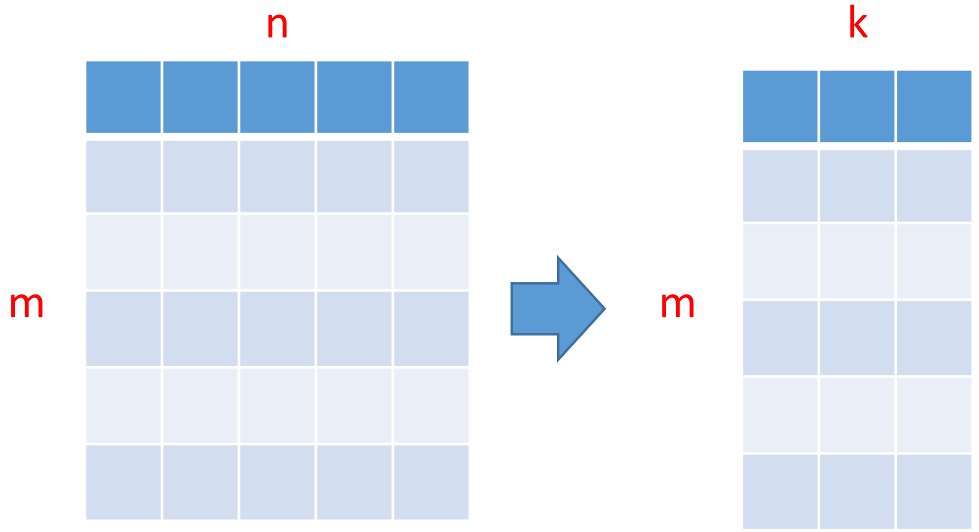


如何解決overfitting

- 正規化 (Regularization)
  - 維持現有的features，但降低部分不重要feature的影響力
  - 對過於複雜模型引進一個懲罰(penalty)
- 降低features的數量 - feature彼此間有較高的相關性
  - 人工選擇(利用domain knowledge)
  - PCA (特徵提取(Feature Extraction))
    - 對非監督式數據壓縮
- 增加資料量

PCA algorithm (m x n → m x k) n維降到k維

- Step 1: Normalize the data
- Step 2: Calculate the covariance matrix
- Step 3: Calculate the eigenvalues and eigenvectors
- Step 4: Choosing components and forming a feature vector
- Step 5: Forming Principal Components



https://www.dezyre.com/data-science-in-python-tutorial/principal-component-analysis-tutorial

Step 1: Normalize the data

- 標準化 features
  - First step is to normalize the data that we have so that PCA works properly.

  - 因為PCA對於不同feature的scale比較敏感

Step 2: Calculate the covariance matrix

- 求共變異數矩陣 (Covariance Matrix) (n*n)

$$Cov(X,Y) = E((X-\mu_x)(Y-\mu_y))$$
$$= E(XY - X\mu_y - Y\mu_x + \mu_x\mu_y)$$
$$= E(XY) - \mu_y E(X) - \mu_x E(Y) + \mu_x\mu_y$$
$$= E(XY) - \mu_y\mu_x - \mu_x\mu_y + \mu_x\mu_y$$
$$= E(XY) - \mu_x\mu_y$$

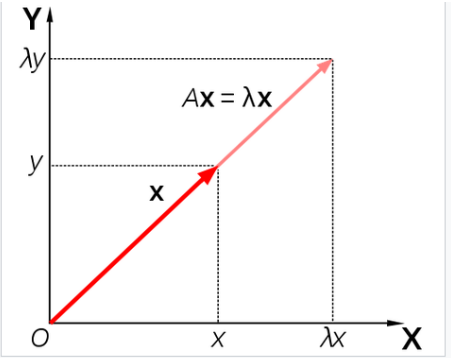
Since the dataset we took is 2-dimensional, this will result in a 2x2 Covariance matrix.

$$Matrix(Covariance) = \begin{bmatrix} Var[X_1] & Cov[X_1, X_2] \\ Cov[X_2, X_1] & Var[X_2] \end{bmatrix}$$

共變異數愈大，線性相關性越高

- Please note that $Var[X_1] = Cov[X_1,X_1]$ and $Var[X_2] = Cov[X_2,X_2]$.

Step 3: Calculate the eigenvalues and eigenvectors



Matrix *A* acts by stretching the vector *x*, not changing its direction, so *x* is an eigenvector of *A*.

- 矩陣分解
  - 特徵向量(eigenvector)，特徵值(eigenvalue)

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$$Av = w = \lambda v$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = 4 \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix}$$

eigenvalue

eigenvector of A

then *v* is an **eigenvector** of the linear transformation *A* and the scale factor *λ* is the **eigenvalue** corresponding to that eigenvector.

Step 3: Calculate the eigenvalues and eigenvectors

例子:

$$Av = w = \lambda v$$

$$A = \begin{bmatrix} -4 & -6 \\ 3 & 5 \end{bmatrix}$$

$$(A - \lambda I)v = 0,$$

where I is the n by n identity matrix.
單位矩陣

$$A - \lambda I = \begin{bmatrix} -4 & -6 \\ 3 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -4-\lambda & -6 \\ 3 & 5-\lambda \end{bmatrix}$$

$$|A - \lambda I| = 0$$

$$|A - \lambda I| = (-4-\lambda)(5-\lambda) + 18 = \lambda^2 - \lambda - 2$$

$$\lambda^2 - \lambda - 2 = 0 \Rightarrow (\lambda - 2)(\lambda + 1) = 0 \Rightarrow \boxed{\lambda = 2 \text{ or } -1}$$

eigenvalues

Step 3: Calculate the eigenvalues and eigenvectors

- λ = 2

$$(A-2I\ )x=\begin{bmatrix}-6 & -6\\ 3 & 3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\mathbf{0}$$

$$-6x_1-6x_2=0$$
$$3x_1+3x_2=0$$

eigenvector $\mathbf{v}_1=r\begin{bmatrix}-1\\ 1\end{bmatrix}$

- λ = -1

eigenvector

$$\mathbf{v}_2=s\begin{bmatrix}-2\\ 1\end{bmatrix}$$

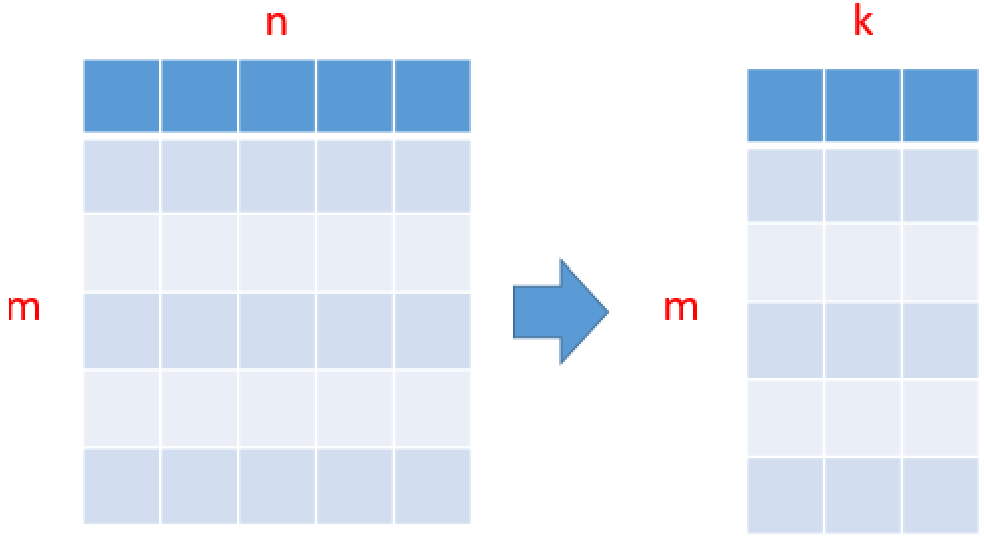
*r*與*s*為純量

Step 4: Choosing components and forming a feature vector

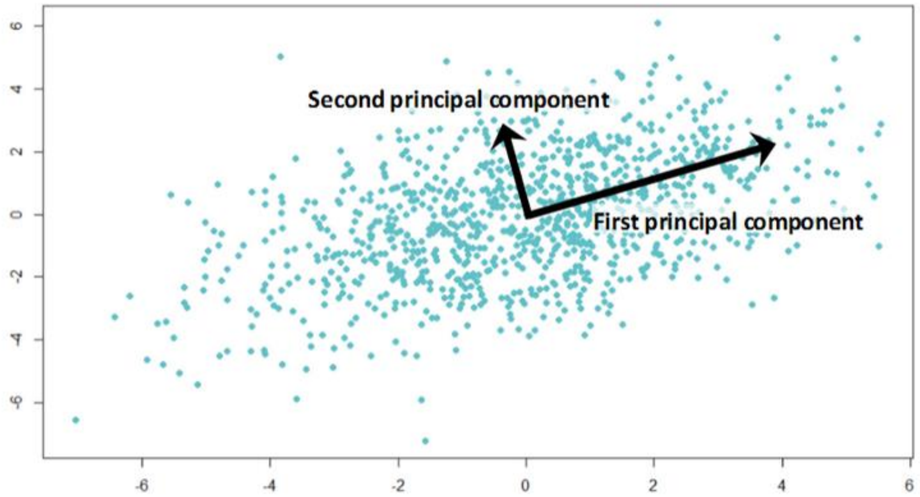
- 選取最大的k個eigenvalues和對應k個eigenvectors
  - 最多有n個eigenvalues: n*1 (k<=n)
- 合併k個eigenvectors成為「投影矩陣(project matrix)」 (W) (W: n*k)

Step 5: Forming Principal Components

- 使用W投影矩陣, 將n維數據集, 輸出為新的矩陣 (m*k)

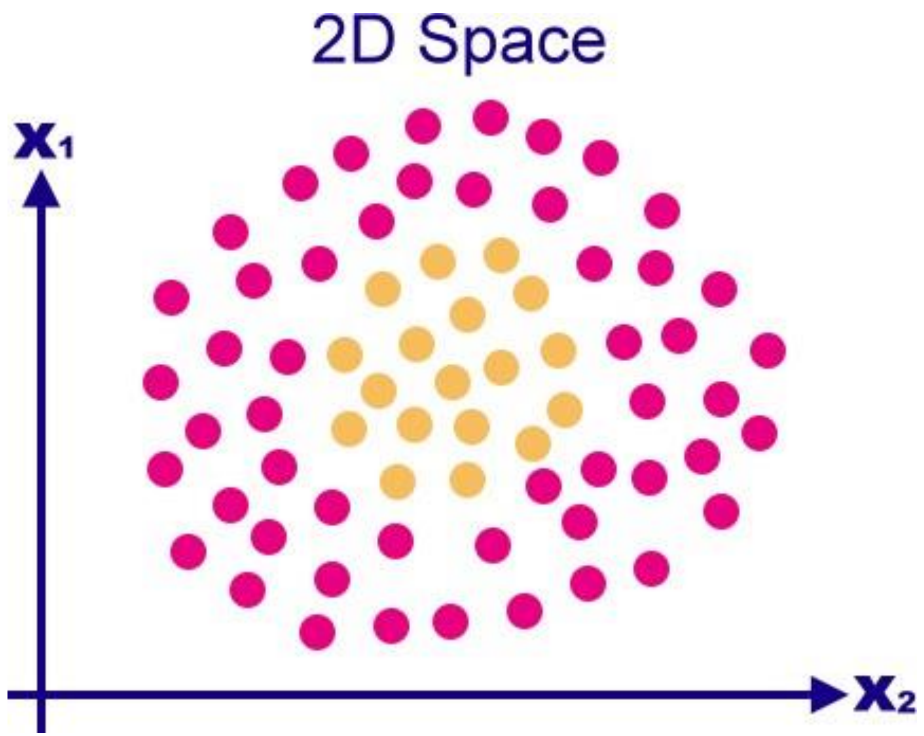


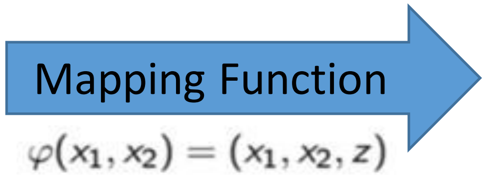
n → k (i.e., k=2)

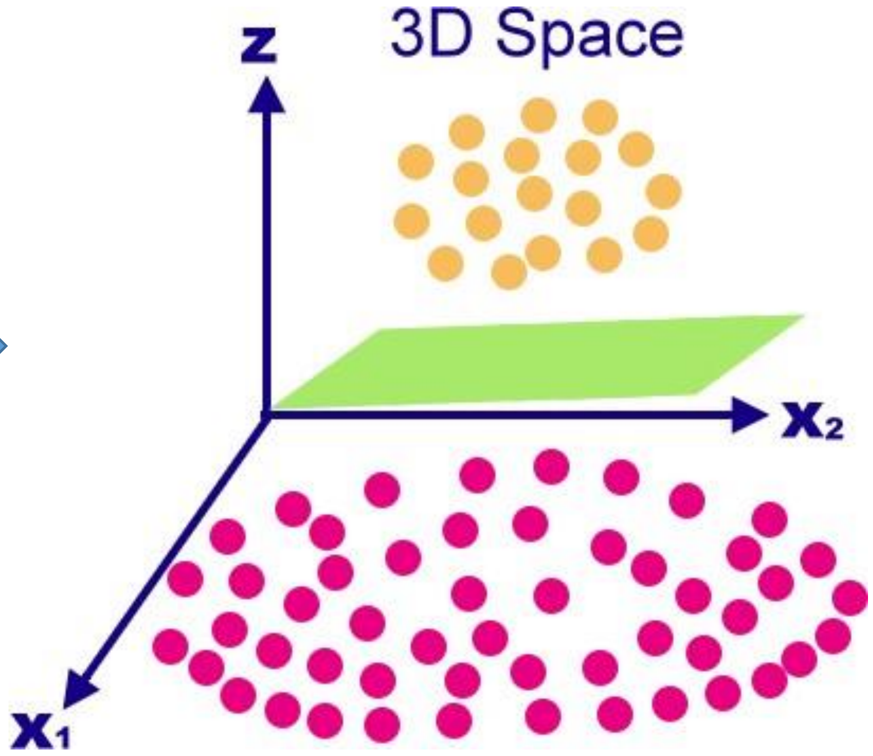


https://en.wikipedia.org/wiki/Principal_component_analysis

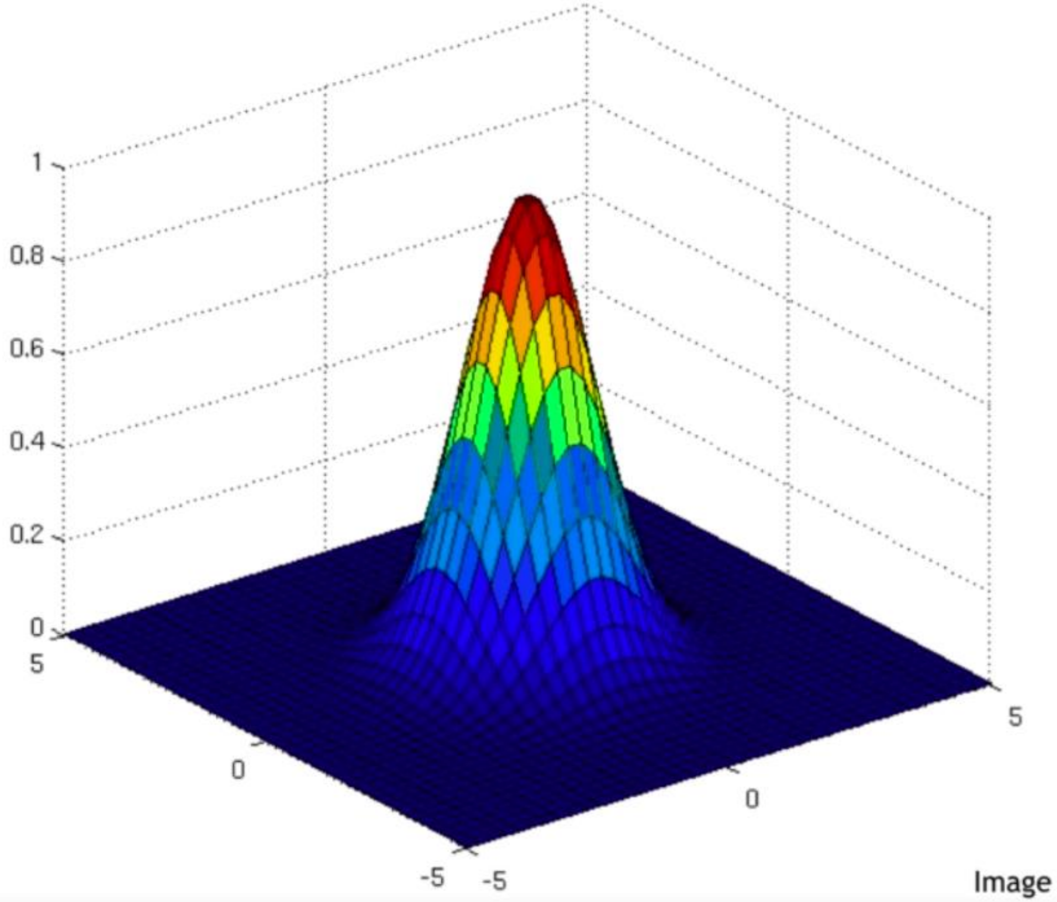
Kernel PCA 線性不可分轉換成線性可分







The Gaussian RBF Kernel



$$K(\vec{x}, \vec{l}^{i}) = e^{-\frac{\left\|\vec{x}-\vec{l}^{i}\right\|^{2}}{2\sigma^{2}}}$$

Image source: http://www.cs.toronto.edu/~duvenaud/cookbook/index.htm

實作時間

- 程式
  - PCA-iris.ipynb
  - PCA-Wine.ipynb
  - Kernel_PCA_SNA.ipynb
- 資料
  - Iris
  - PCA-Wine.ipynb
  - Social_Network_Ads

版權聲明

- 本講義所使用之圖片, 表格, 文字, 內容, 書籍資料, 引用統計資料與程式碼及數據集資料等, 除自製外，其智慧財產權為原網站, 作者, 公司所擁有。

- 講義投影片, 程式碼與數據集僅供教學使用, 請同學勿將課程所使用之講義投影片, 程式碼與數據集放在網路上供人下載及分享, 也請勿做商業用途。