

Python 資料科學應用開發

第十堂：特徵選擇與萃取

(Feature selection and extraction)

同學，歡迎你參加本課程

- ☑ 請關閉你的FB、Line等溝通工具，以免影響你上課。
- ☑ 考量頻寬、雜音，請預設關閉攝影機、麥克風，若有需要再打開。
- ☑ 隨時準備好，老師會呼叫你的名字進行互動，鼓勵用麥克風提問。
- ☑ 如果有緊急事情，你必需離開線上教室，請用聊天室私訊給老師，以免老師癡癡呼喚你的名字。
- ☑ 軟體安裝請在上課前安裝完成，未完成的同學，請盡快進行安裝。

課程檔案下載

The screenshot displays the homepage of the Juei Computer Online Live website. The header features the site's name and navigation links. A large banner on the left promotes Python programming. On the right, a dropdown menu is open, highlighting the 'Course Archive Download' option. An orange callout bubble points to this option.

巨匠電腦線上真人 開課查詢 免費體驗專區 課程總覽 ▾ 專業師資 學員專區 ▾ 講師專區 最新消息

360 f YouTube

您好! 登出

點數卡產品兌換
APCS檢測專區
公告專區
我的課表
IT真人課程劃位
電腦分校課程劃位
外語真人課程劃位
美語分校課程劃位
取消劃位
課程檔案下載
上課權益查詢
教學平台測試
學習諮詢
常見問題
個資維護
忘記密碼
登出

程式語言 **好難學?**
那是因為
你還沒學過 **Python!**
(線上老師 **LIVE** 直播教學 · 搶先看)

巨匠電腦真人課程

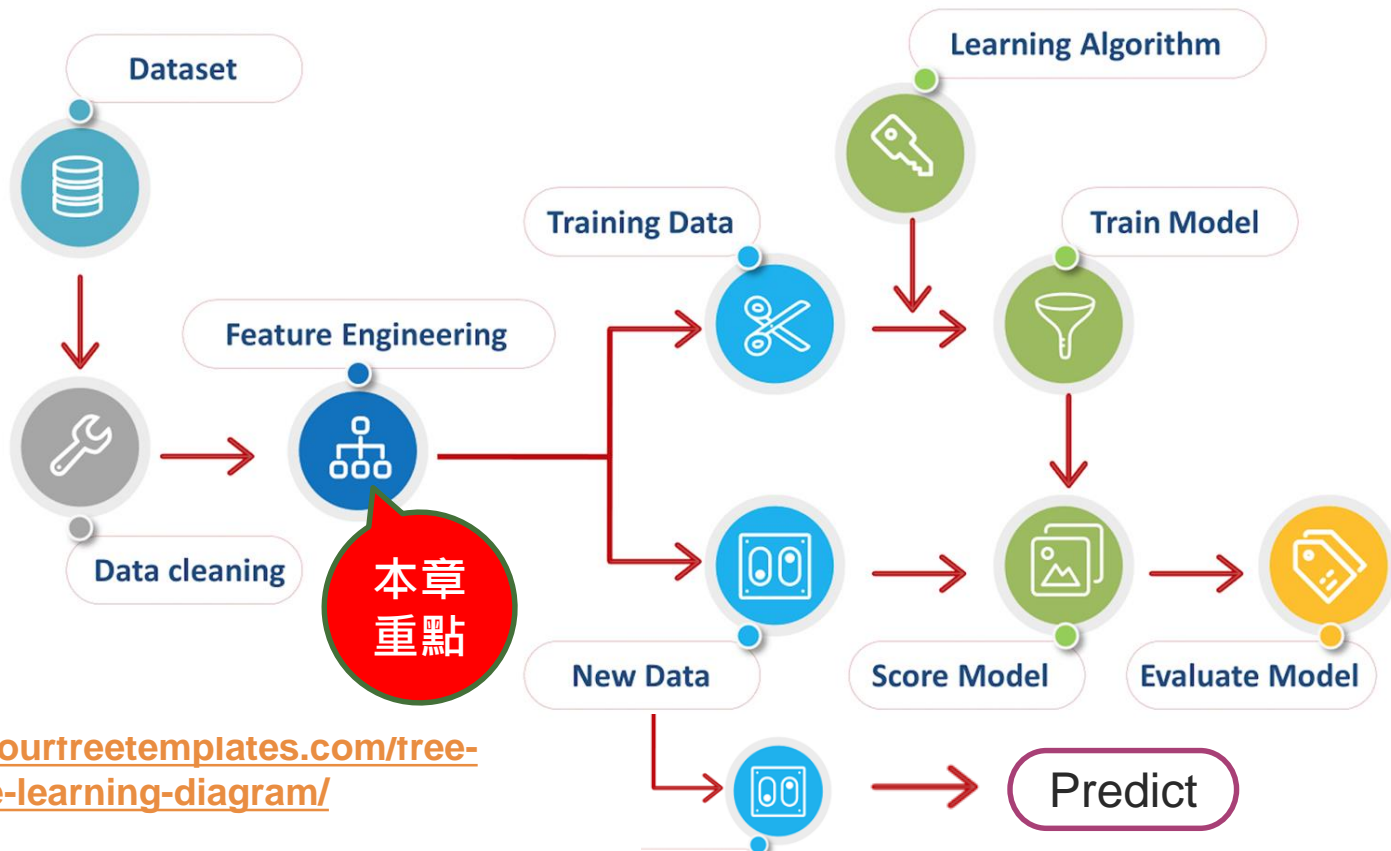
ZOOM 學員操作說明

The screenshot shows the Zoom interface with several callouts:

- 5 查看選項/共同註記/筆 (連連看)**: Points to the '共同註記' (Co-Annotate) option in the top right menu.
- 2 共享螢幕 (指導演練；點評作品)**: Points to the '共享螢幕' (Share Screen) button in the bottom toolbar. A sub-note says: '老師須先停止共享螢幕才能請學生共享螢幕' (The teacher must first stop sharing the screen before asking the student to share the screen).
- 1 聊天**: Points to the '聊天' (Chat) button in the bottom toolbar.
- 3 與會者/舉手**: Points to the '與會者' (Participants) button in the bottom toolbar.
- 4 解除靜音**: Points to the '解除靜音' (Unmute) button in the bottom toolbar.

Additional interface elements visible include the top bar with 'www.pcschool.com.tw', a toolbar with icons for '游鼠' (Cursor), '文字' (Text), '筆' (Pen), '橡皮' (Eraser), '格式' (Format), '撤銷' (Undo), '重做' (Redo), and '清除' (Clear). A '與會者 (15)' window is open, showing a list of participants and a '舉手' (Raise Hand) button.

機器學習流程



<https://yourtreetemplates.com/tree-machine-learning-diagram/>

偏差與差異

◆ 模型擬合

◆ **Overfitting**：過度擬合，找出來的模型受到訓練資料的影響太大，使得對預測的效果不佳。

◆ **Underfitting**：低度擬合，模型對於資料的描述能力太差，無法正確解釋資料。

◆ 可透過偏差和差異調整降低誤差

◆ 偏差或差異會導致模型過度擬合或低度擬合。

偏差與差異

◆ 偏差 Bias

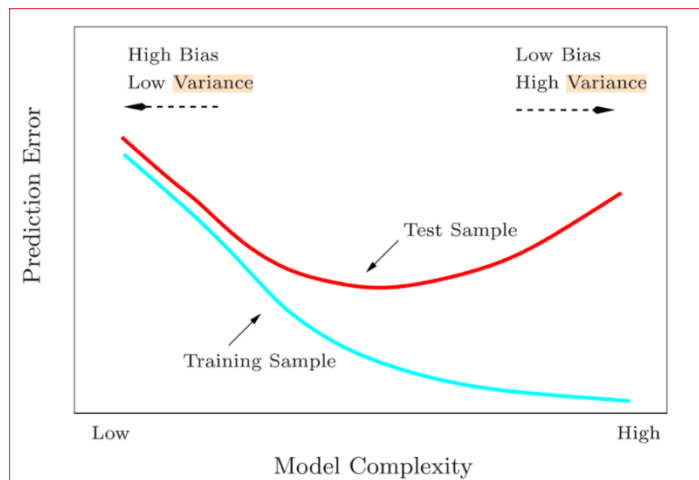
- ◆ 偏差是指實際值與預測值差異。
- ◆ 當模型具有高偏差時，代表模型過於簡單並且不能捕獲數據的複雜性，不適合用於這個數據上。

◆ 差異 Variance

- ◆ 當模型在訓練有素的資料集上表現良好，但在未經過訓練的資料集上表現不佳時，就會出現差異。
- ◆ 差異告訴我們實際值與預測值分散程度。
- ◆ 高差異導致過度擬合，遇到尚未學習的不同數據時無法做出正確的預測。

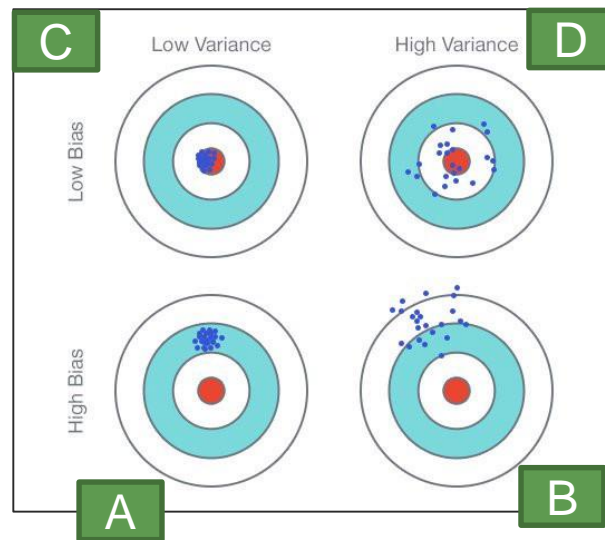
偏差與差異

- ◆ 高偏差的模型看起來非常簡單。具有高差異的模型試圖適合大多數數據，使得模型複雜且難以建模。



偏差與差異

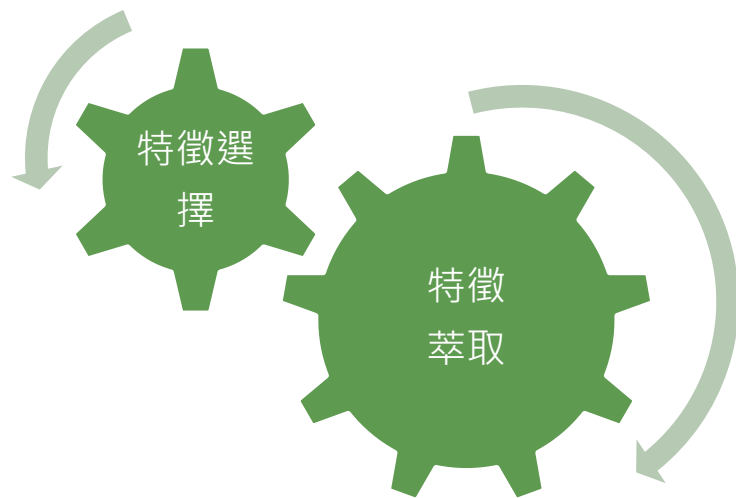
- ◆ A- 高偏差低差異：模型一致但平均不準確。
- ◆ B- 高偏差高差異：模型不準確，平均不一致。
- ◆ C- 低偏差低差異：模型在平均值上是準確且一致的。我們在模型中努力實現這一目標。
- ◆ D- 低偏差高差異：模型有些準確，但平均值不一致。數據的微小變化可能導致較大的錯誤。
- ◆ Lab (275x) :
 - ◆ Module5-275 / Bias-Variance-Trade-Off.ipynb



降維 (Dimensionality Reduction)

- ◆ 目的：reduce the complexity of the model and avoid overfitting
- ◆ 分兩類：
 - ◇ 特徵選擇 (Feature Selection)：只選擇部分特徵，作為訓練模型的輸入
 - ◇ 特徵萃取 (Feature Extraction)：從既有特徵導出新的特徵空間

降維之後的特徵處理



特徵選擇

◆ 兩種分析方式

- ◆ 透過 **SBS**，進行各種組合計算，循序的移除特徵，將最佳分數存入 **list**，最後只包含想要的特徵。
- ◆ 透過隨機森林分類方式，顯示出特徵的重要性。

循序向後選擇

(Sequential Backward Selection ; SBS)

- ◆ SBS循序的移除特徵，直到特徵空間只包含所想要的特徵個數。
- ◆ 目標：最小化『準則函數』（ Criterion Function ）
- ◆ 按以下步驟進行
 1. 以 $k=d$ 初始化演算法，其中 d 是全部「特徵空間」 \mathbf{X}_d 維數。
 2. 確認會令「準則」最大的「特徵」 $\mathbf{x}^- = \arg \max J(\mathbf{X}_k - \mathbf{x}) \quad \mathbf{x} \in \mathbf{X}_k$
 3. 從「特徵空間」中移除「特徵」： $\mathbf{X}_{k-1} := \mathbf{X}_k - \mathbf{x}^-; k := k - 1$
 4. 如果 k 等於所需「特徵」的個數，則停止，不然回到步驟 2 繼續。

實作

◆ 程式碼：ch04.ipynb

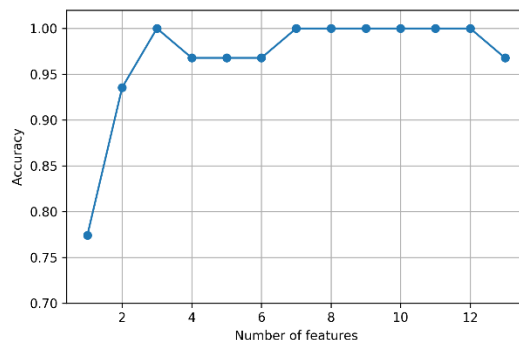
◆ 第45 ~ 49格，Sequential feature selection algorithms

◆ `from itertools import combinations`

會取各種組合計算分數，將最佳分數放入比較的List中

◆ 當特徵 = 3時，與特徵 = 12時準確率一樣好

◆ 第47格顯示哪三個特徵



優點

- ◆ 降低資料集的大小，可節省資料收集的成本，並降低資料調查的難度
- ◆ 避免過多特徵值造成的『維數災難』（The curse of dimensionality），見下頁說明

維數災難 (The curse of dimensionality)

維數災難

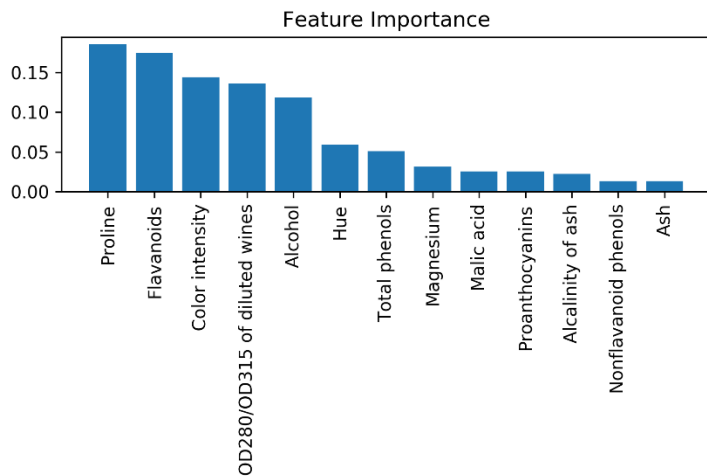
必須要注意的是，由於「維數災難」(curse of dimensionality)，「KNN」很容易產生「過度適合」現象。「維數災難」描述這樣一個現象：對一個固定大小的「訓練數據集」，當維度越多的時候，特徵空間變得越來越稀疏。直觀地說，我們可以這樣想，在一個高維度的空間中，即使是最近的鄰近樣本，也是太遙遠到不能提供一個合理的估計。



我們在介紹「邏輯斯迴歸」的小節中，討論過關於「正規化」(regularization)的觀念，以它來避免「過度適合」。然而，面對「正規化」不適用的模型（如：「決策樹」與「KNN」），我們可以使用「特徵選擇」或是「降維技術」，來幫助我們避免「維數災難」。這將在下一章中詳細的討論。

以隨機森林評估特徵的重要性

- ◆ 第50格，Assessing feature importance with Random Forests
- ◆ 利用 RandomForestClassifier 的 feature_importances_ 屬性可顯示特徵的重要性



降維 (Dimensionality Reduction)

- ◆ 特徵選擇 (Feature Selection)
 - ◇ 只選擇部分特徵，作為訓練模型的輸入
- ◆ 特徵萃取 (Feature Extraction)
 - ◇ 從既有特徵導出新的特徵空間

特徵萃取

主成分分析 PCA

- Principal Component Analysis
- 非監督式學習
- 可線性分離

線性判別分析 LDA

- Linear Discriminant Analysis
- 監督式學習
- 可線性分離

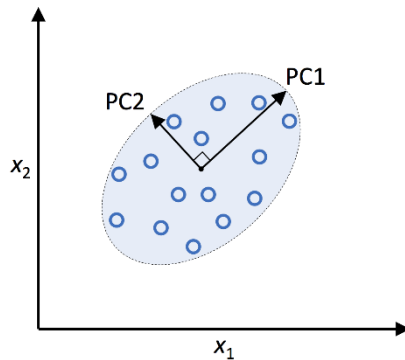
核主成分分析 Kernel PDA

- 非線性分離

主成分分析

(Principal Component Analysis ; PCA)

- ◆ 將數據轉換，投影到較低維的特徵空間，是一種以保留最多資訊為前提的數據壓縮方法。
- ◆ 新的特徵空間的『正交軸』（Orthogonal axes）即為『主成分』（Principal Component ; PC），如右圖。
 - ◆ x_1 、 x_2 ：原來的特徵座標軸
 - ◆ PC1、PC2：主成分



特徵萃取方法

- ◆ 從 d 維轉換為 k 維， $d > k$



$$\mathbf{x} = [x_1, x_2, \dots, x_d], \quad \mathbf{x} \in \mathbb{R}^d$$

$$\downarrow \mathbf{x}W, \quad W \in \mathbb{R}^{d \times k}$$

$$\mathbf{z} = [z_1, z_2, \dots, z_k], \quad \mathbf{z} \in \mathbb{R}^k$$

- ◆ 首先找變異數最大的特徵，為第一個主成分
- ◆ 其次，找次大的變異數最大的特徵，而且與第一個主成分最不相關（即正交）的特徵
- ◆ 依此類推，至全部特徵找到為止

主成分分析步驟

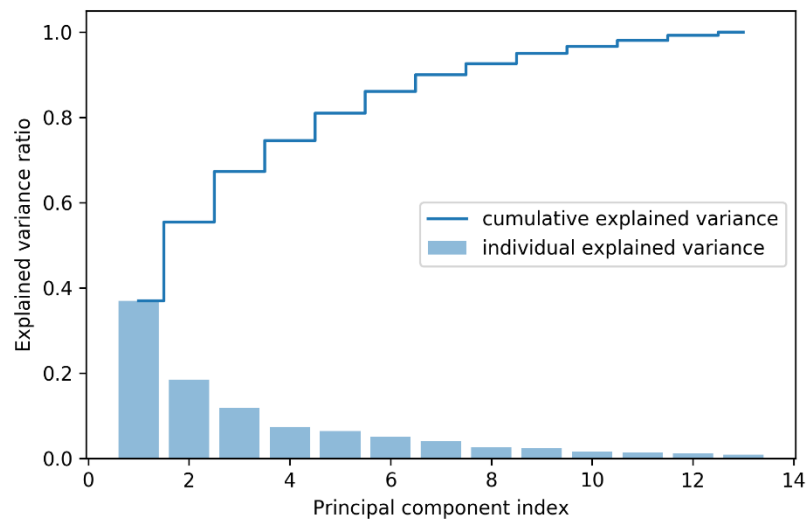
1. 標準化 d 維「數據集」。
2. 建立「共變異數矩陣」(covariance matrix)。
3. 分解「共變異數陣」為「特徵向量」(eigenvector)與「特徵值」(eigenvalues)。
4. 依照「特徵向量」相應的「特徵值」以遞減的方式對進行排序。
5. 選取 k 個最大「特徵值」相對應的 k 個「特徵向量」，其中 k 是新「特徵空間」的維數 ($k \leq d$)。
6. 用「最上面」的 k 個「特徵向量」，建立「投影矩陣」(project matrix) \mathbf{W} 。
7. 使用「投影矩陣」 \mathbf{W} ，轉換輸入是 d 維「數據集」，輸出是新的 k 維「特徵子空間」。

實作

◆ 程式碼：ch05.ipynb

◆ 第4 ~ 13格：Wine資料集 PCA 範例

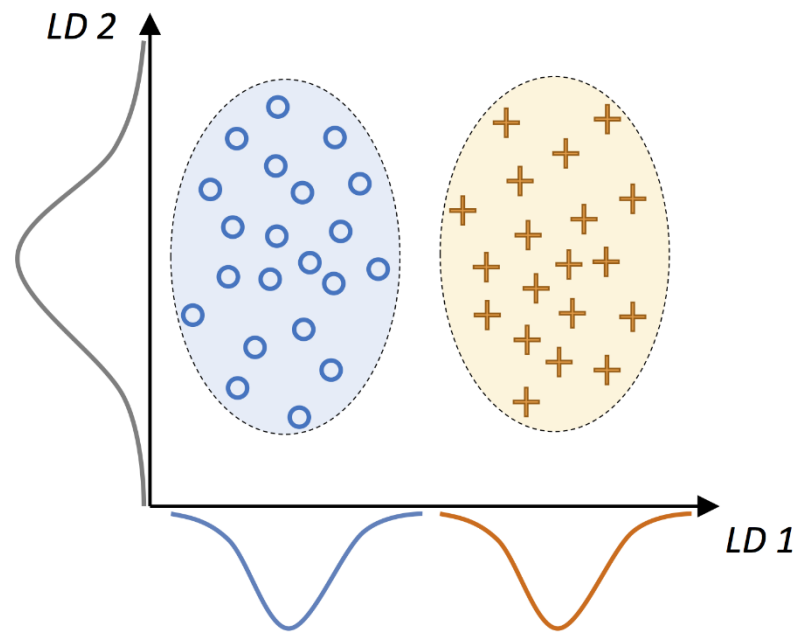
◆ 第14 ~ 22格：scikit-learn 作法



線性判別分析

(Linear Discriminant Analysis ; LDA)

- ◆ PCA 屬於『非監督演算法』，
LDA 屬於『監督演算法』
- ◆ LDA 求取類別內(s_w)的『散佈矩陣』
愈小愈好，類別間(s_b)的『散佈矩陣』
愈大愈好。類似 PCA，
依 $s_w^{-1} s_b$ 降冪排序，選取 N 個新
特徵



步驟

1. 「標準化」 d 維「數據集」（ d 為「特徵」的個數）。
2. 對於每個類別，計算 d 維的「平均值向量」（mean vector）。
3. 建立「類別間」（between-class）的「散佈矩陣」（scatter matrix） S_B 與「類別內」（within-class）的「散佈矩陣」 S_w 。
4. 從矩陣 $S_w^{-1}S_B$ 中計算「特徵向量」和相對應的「特徵值」。
5. 依照「特徵向量」相應的「特徵值」以遞減的方式對進行排序。
6. 選擇最大的 k 個「特徵值」的相對應的 k 個「特徵向量」。並依此建立 $d \times k$ 維的「轉換矩陣」 W ；「特徵向量」包含在「轉換矩陣」的「行」中。
7. 使用「轉換矩陣」 W ，將樣本「投影」到新「特徵子空間」中。

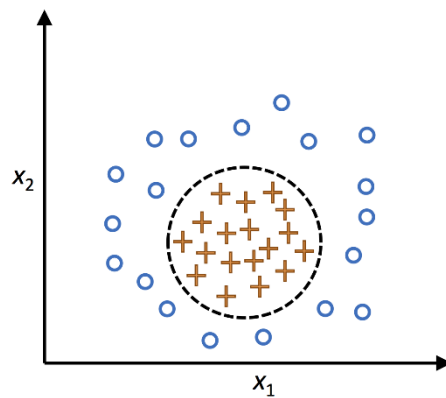
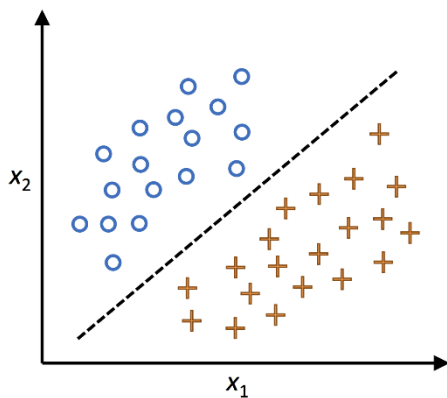
實作

- ◆ 程式碼：ch05.ipynb
 - ◆ 第23 ~ 33格：LDA python 實作
 - ◆ 第34 ~ 36格：scikit-learn 作法

核主成分分析

(Kernel Principal Component Analysis ; Kernel PCA)

- ◆ PCA 及 LDA 均假設輸入的數據是可以『線性分離』的，因此降維時，採取線性轉換，投影到新的特徵空間
 - ◆ Kernel PCA 則支援將『非線性分離』的數據，降維成可『線性分離』的新特徵空間
- ◆ 左圖：線性分離
 - ◆ 右圖：非線性分離



實作

- ◆ 程式碼：ch05.ipynb
 - ◇ 第38 ~ 39格：Kernel PCA python 實作
 - ◇ 第40 ~ 41格：scikit-learn 作法
- ◆ Lab (275x)
 - ◇ Module5-275 / DimensionalityReduction.ipynb

作品：鐵達尼資料集資料清理

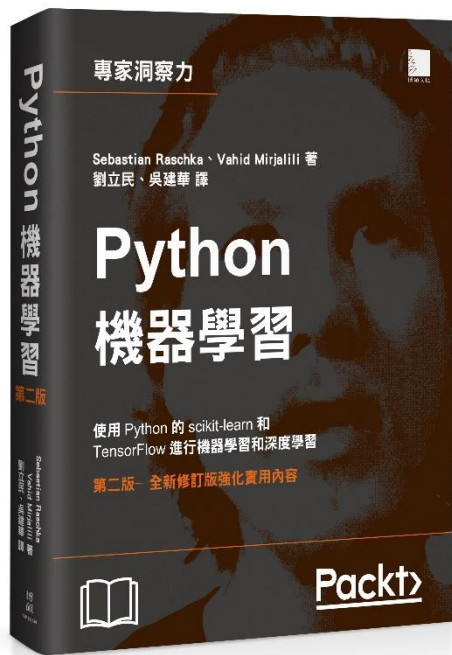
- ◆ 資料集：鐵達尼 (Titanic)
- ◆ `import seaborn as sns`
- ◆ `titanic = sns.load_dataset("titanic")`
- ◆ 參考

展示



『[Titanic survival - python solution _ Kaggle](#)』 by zhenqi_liu

參考用書



- ◆ 書名：Python機器學習（第二版）

<http://www.drmaster.com.tw/bookinfo.asp?BookID=MP11804>

- ◆ 作者：Sebastian Raschka, Vahid Mirjalili ISBN
- ◆ 譯者：劉立民、吳建華
- ◆ 出版社：博碩

問卷

<http://www.pcschoolonline.com.tw>

開課查詢

免費體驗專區

課程總覽

專業師

1

學員專區

講師專區



➤ 課程檔案下載：

學員的「上課教材」，下載檔案為壓縮檔 ([解壓縮操作步驟](#))。
如無法觀看上課教材，請安裝 [PDF閱讀軟體](#)。

公告專區

我的課表

課程劃位

取消劃位

2

課程檔案下載

自107年1月1日起，課程錄影檔由180天改為365天(含)內無限次觀看 (上課隔日18:00起)。

問
卷

上課日期	課程名稱	課程節次	教材下載		
2017/12/27 2000 ~ 2200	線上真人-ZBrush 3D動畫造型設計	18	上課教材	錄影檔	課堂問卷
2017/12/20 2000 ~ 2200	線上真人-ZBrush 3D動畫造型設計	17	上課教材	錄影檔	
2017/12/18 2000 ~ 2200	線上真人-ZBrush 3D動畫造型設計	16	上課教材	錄影檔	



巨匠線上真人

www.pcschoolonline.com.tw