

國立臺中科技大學應用統計系 -輔導學生考照講座

大數據分析

- R/Python/Julia/SQL 程式設計與應用
(R/Python/Julia/SQL Programming and Application)
- 資料視覺化 (Data Visualization)
- 機器學習 (Machine Learning)
- 統計品管 (Statistical Quality Control)
- 最佳化 (Optimization)



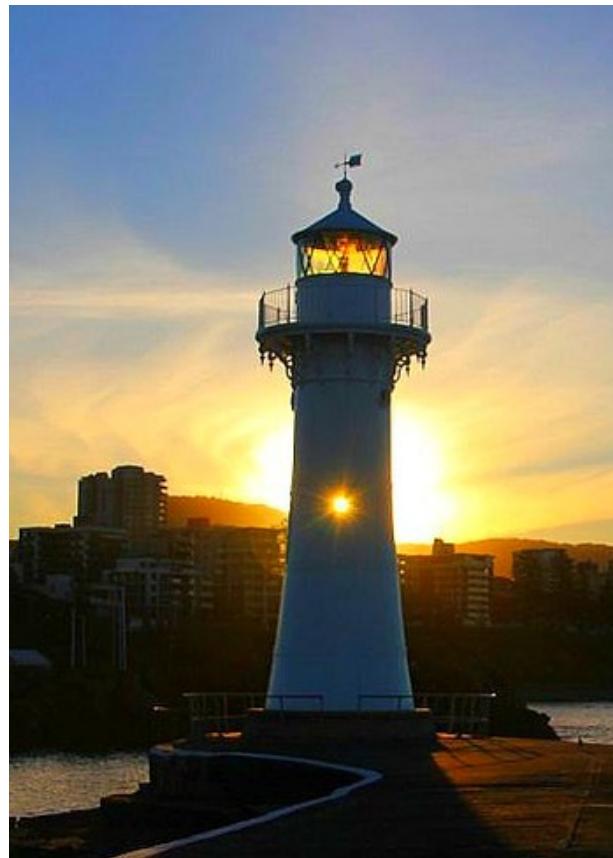
李明昌博士

alan9956@gmail.com

<http://rwepa.blogspot.com/>

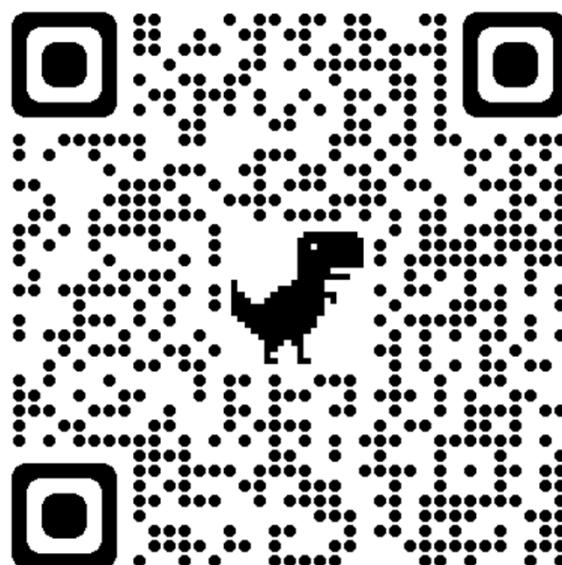
大綱

1. 學習方向與技能
2. 資料科學家技能
3. 認證考試
4. Q & A



RWEPA簡介 <http://rwepa.blogspot.com/>

- 姓名：李明昌 (ALAN LEE)
- 現職：中華R軟體學會 常務理事
臺灣資料科學與商業應用協會 常務理事
- 大專院校、資策會、工業技術研究院、國家發展委員會、中央氣象局、公平交易委員會、各縣市政府與日本名古屋產業大學等公民營單位演講達290餘場，2600小時以上。

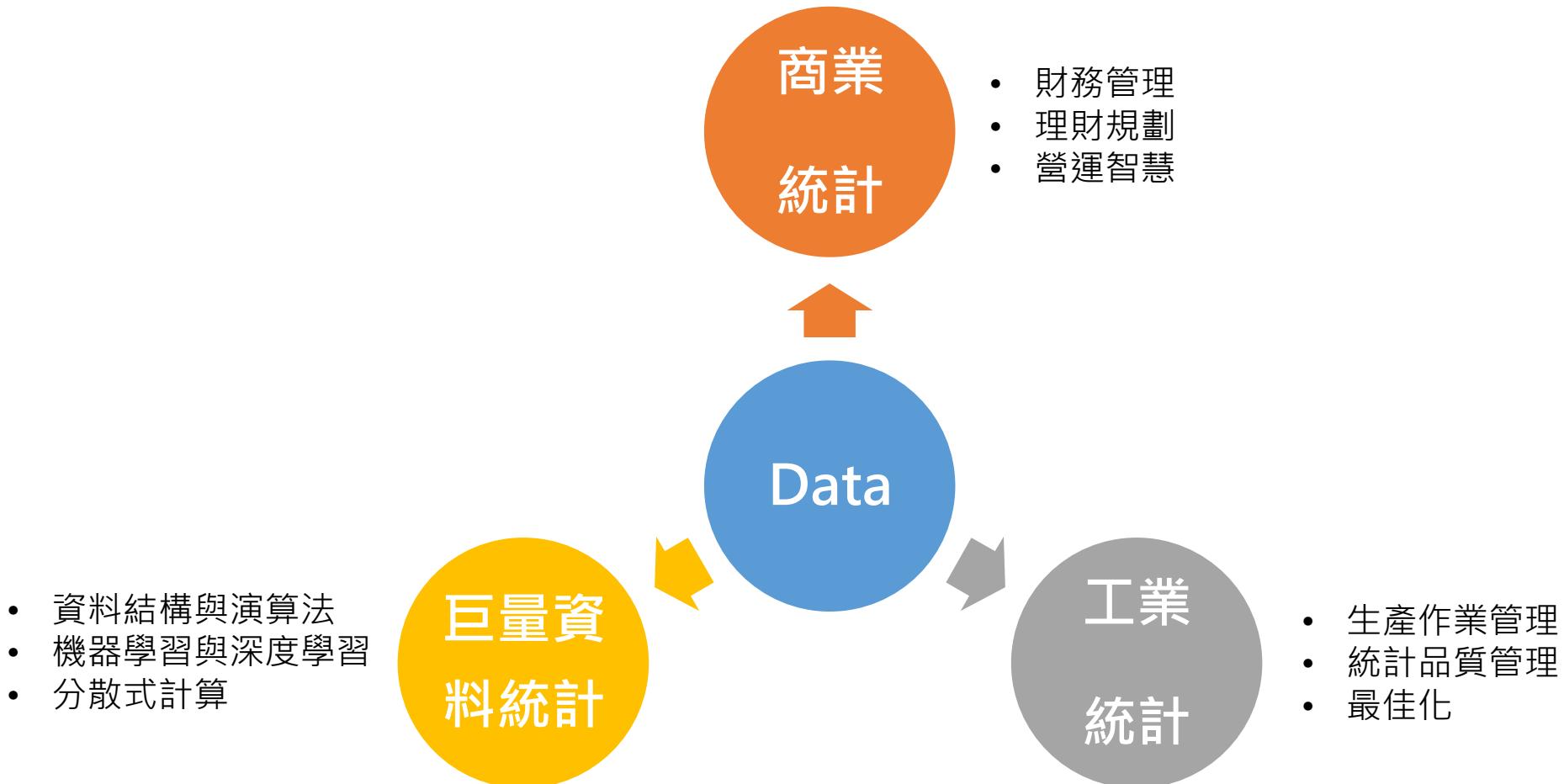


- iPAS 巨量資料分析師 證照推廣
- iPAS 營運智慧分析師 證照推廣



1.學習方向與技能

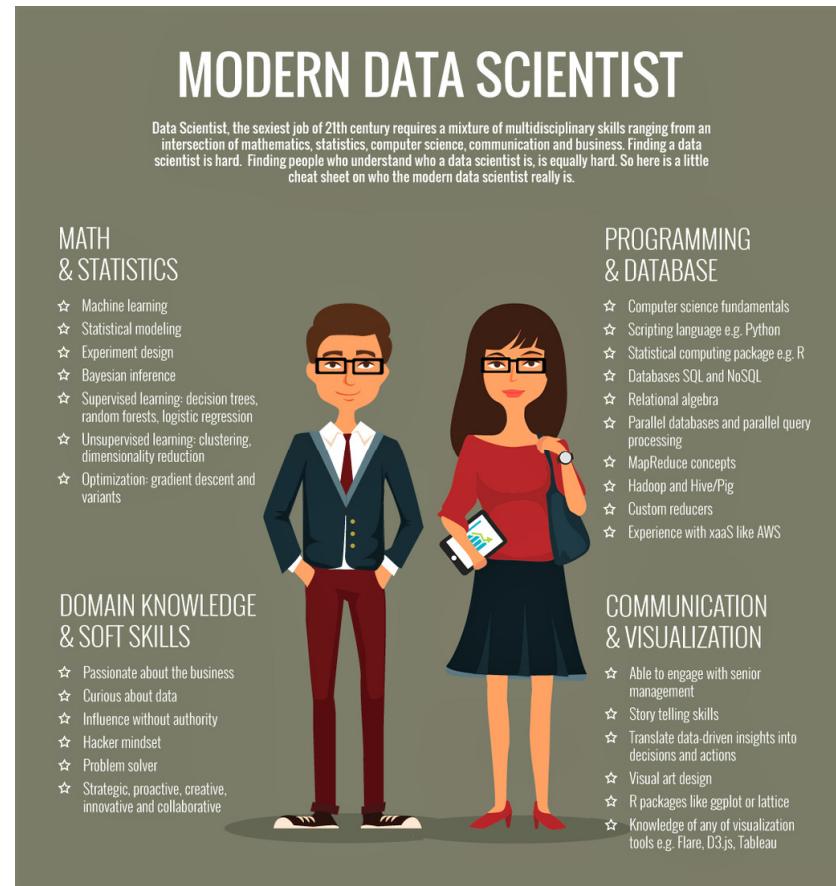
國立臺中科技大學應用統計系





2. 資料科學家技能

資料科學家四大技能



The 10 Algorithms Data Scientist must have to Know:

https://medium.com/@_moazzemhossain/the-10-algorithms-data-scientist-must-have-to-know-97a2c478ce94

數學,統計

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning
- Unsupervised learning
- Optimization

專業知識與軟體技術

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative
(戰略性、主動性、創造性、創新性和協同性) → 老闆

程式與資料庫

- Computer science fundaments
- Scripting language (Python, R, Julia)
- Statistical computing package
- Database SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- Hadoop , MapReduce, Spark
- AWS, Azure, Google Cloud



溝通與視覺化

- Able to engage with senior management
- Story telling skills (Tableau, Power BI)
- Translate data-driven insights into decisions and actions
- Visual art design
- R package: `ggplot2`, `plotly`, `shiny`
- Python package: `plotnine`, `plotly`, `dash`
- Knowledge for visualization

資料科學的心法

Python vs. R

- <https://www.python.org/>

程式語言

Python is a **programming language** that lets you work quickly and integrate systems more effectively. [»» Learn More](#)

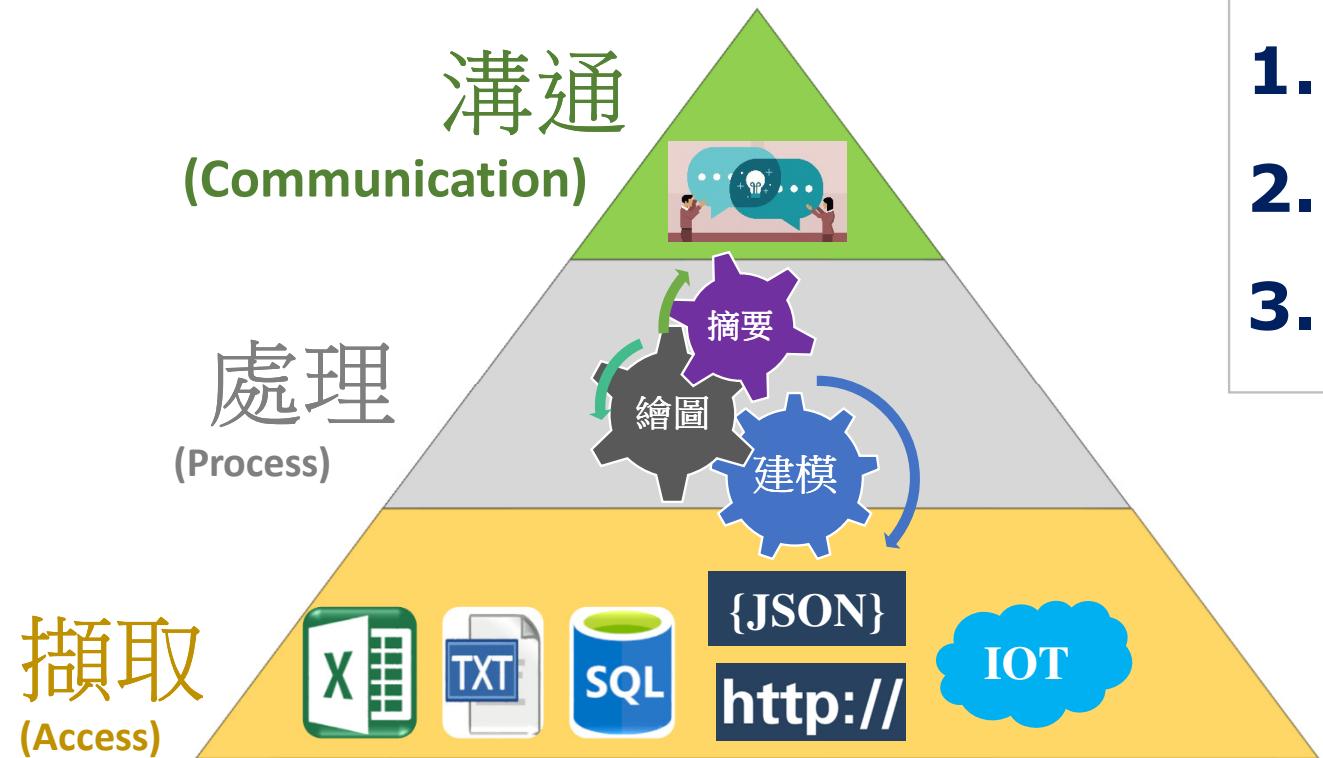
- <https://www.r-project.org/>

統計運算

R is a free software environment for **statistical computing** and **graphics**. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose

繪圖

★★★資料分析架構→APC方法



如何學習 Python/R?

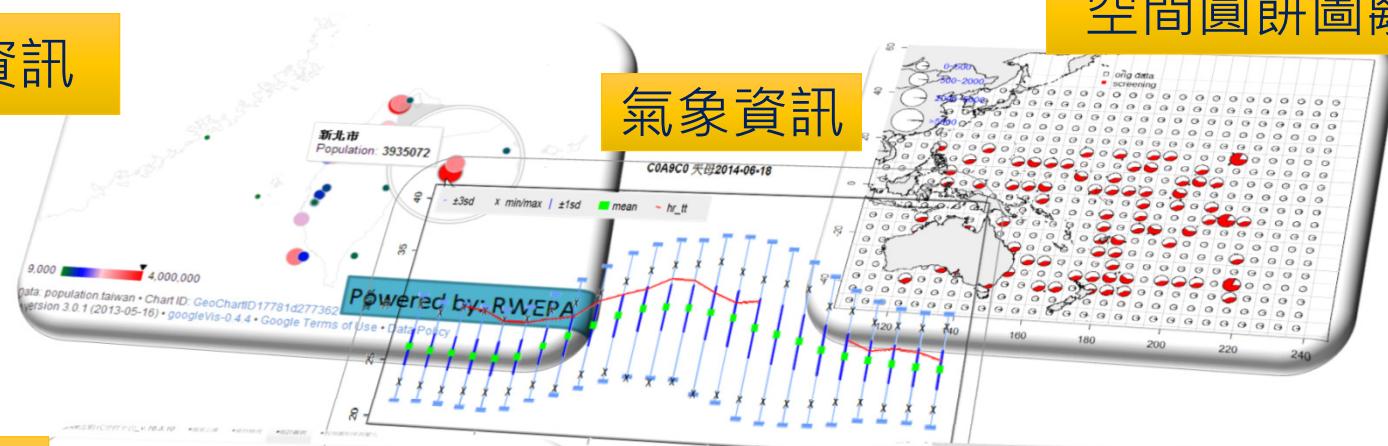
- 熟悉教材內容
- 將教材內容的資料集改為工作(學術)資料集
- 遇到問題時, 想辦法尋找答案
- 掌握 APC方法
- 掌握 摘要, 繪圖, 建模
- 參考網路應用文章 (進階) & 學術論文



資料分析與視覺化應用

R + shiny → 互動式網頁

地理資訊

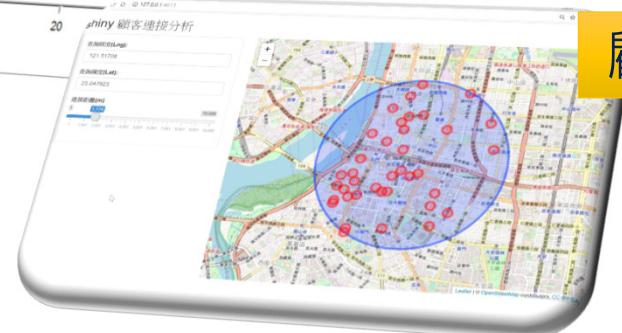


空間圓餅圖離群值分析

氣象資訊



保險預測



顧客連結分析

中央氣象局 1,600萬筆資料

網頁呈現



客製化選單

R統計運算

保險預測模型

機率模型閾值調整

預測結果

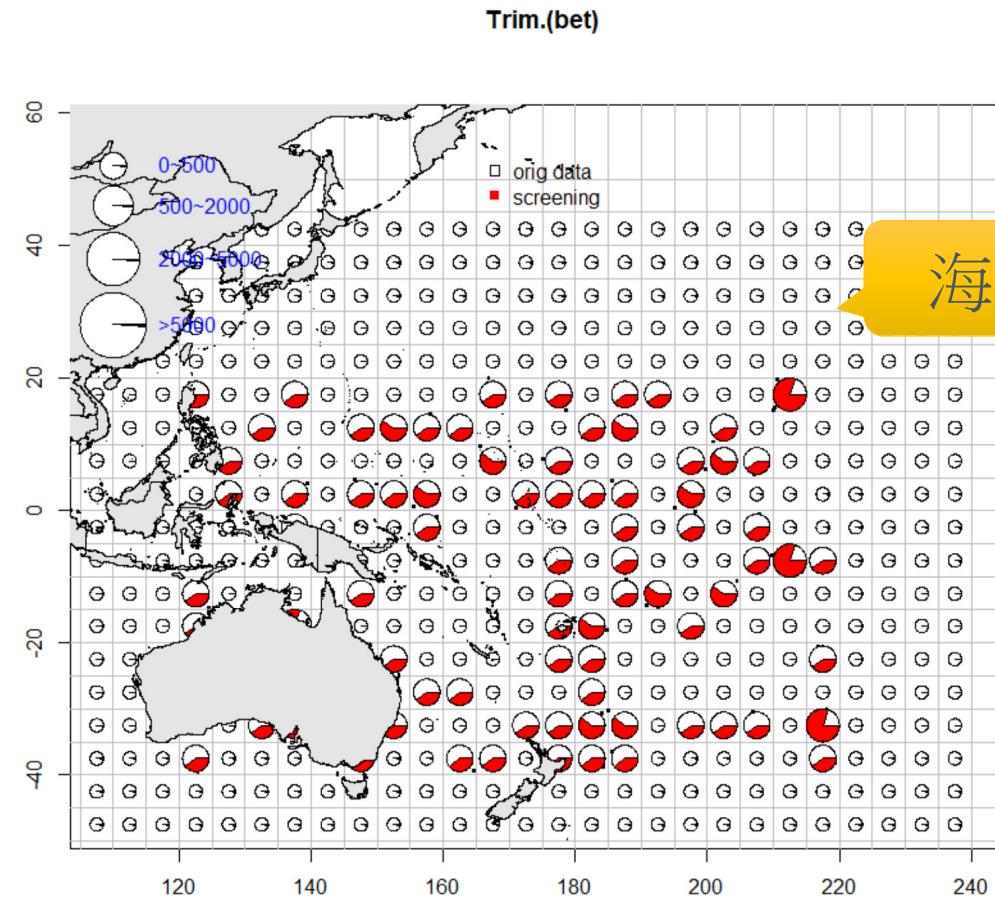
The screenshot shows the iinsurance interactive analysis platform version v.16.3.24. The top navigation bar includes links for '檔案上傳' (File Upload), '資料處理' (Data Processing), '統計圖表' (Statistical Charts), '模型評估' (Model Evaluation), and '預測模型' (Prediction Model). A red box highlights the '機率模型閾值' (Probability Model Threshold) input field, which is set to 0.1. Another red box highlights the '檢視結果' (View Results) button in the '預測資料上傳' (Upload Prediction Data) section. A yellow callout points to the '預測結果' column in the table below. A red arrow points from the '機率模型閾值' field to the '檢視結果' button.

性別	女性	車輛種類	私家車	曝露風險		無索償折扣	被保險人年齡	私家車 一車齡 0	私家車 一車齡 1	私家車 一車齡 2	私家車 -車齡 0_1_2 組合	車齡 0_1_2 組合	預測機率	理賠		
				曝露風險對數	對數											
M	0	A	1	0.9144422	-0.08944106	50	4	1	0	0	1	0	2	0.1069	有	
M	0	A	1	0.8158795	-0.20348856	20	4	0	0	1	1	2	2	0.1441	有	
3	M	0	A	1	0.8377823	-0.17699695	50	3	0	0	1	1	2	2	0.1866	有
4	M	0	A	1	0.4325804	-0.83798702	50	6	0	1	0	1	1	2	0.0944	無
5	M	0	A	1	0.7173169	-0.33223755	50	4	0	0	1	1	2	2	0.1218	有
6	M	0	A	1	0.8377823	-0.17699695	50	4	0	0	1	1	2	2	0.1495	有
7	M	0	A	1	0.8487337	-0.16400975	50	5	0	0	1	1	2	2	0.1422	有
8	F	1	A	1	0.8268309	-0.19015503	10	3	0	0	1	1	2	2	0.1733	有
9	M	0	A	1	0.7145791	-0.33606164	0	5	1	0	0	1	0	2	0.0694	無
10	M	0	A	1	0.3340178	-1.09656101	0	3	0	0	1	1	2	2	0.0783	無

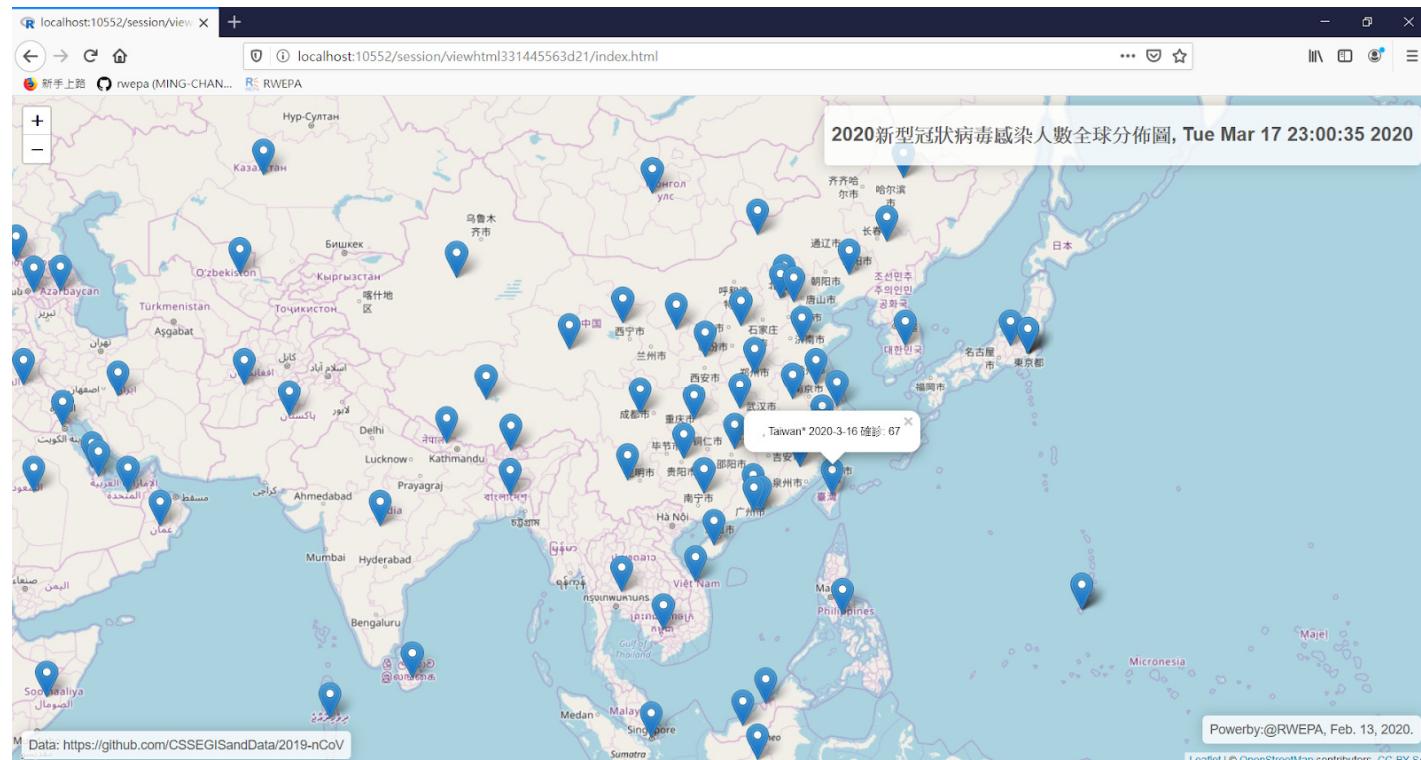
Showing 1 to 10 of 12 entries

127.0.0.1:6177/#tab-9487-2

空間圓餅圖離群值分析



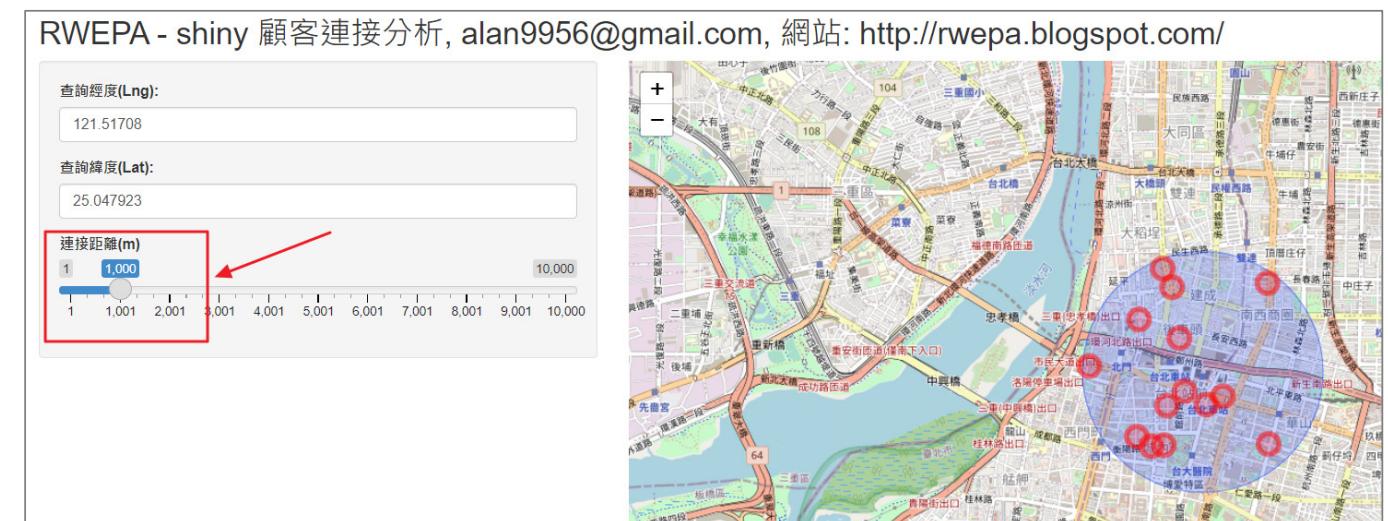
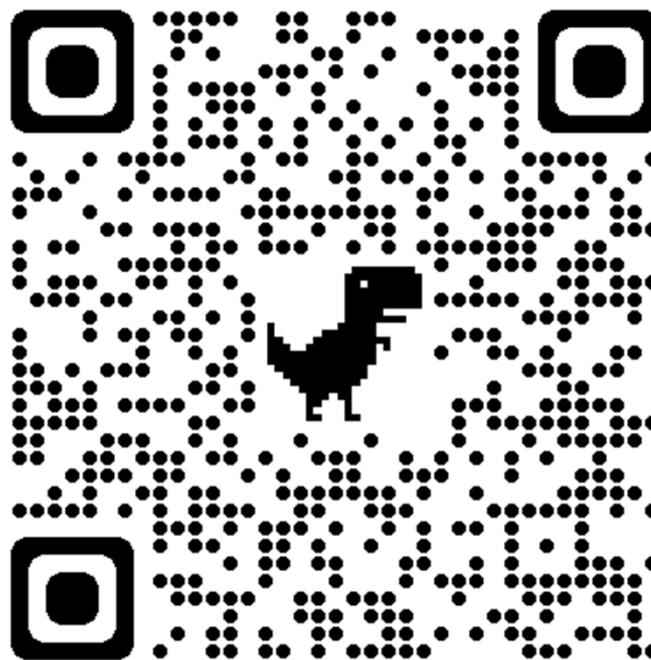
2020新型冠狀病毒視覺化



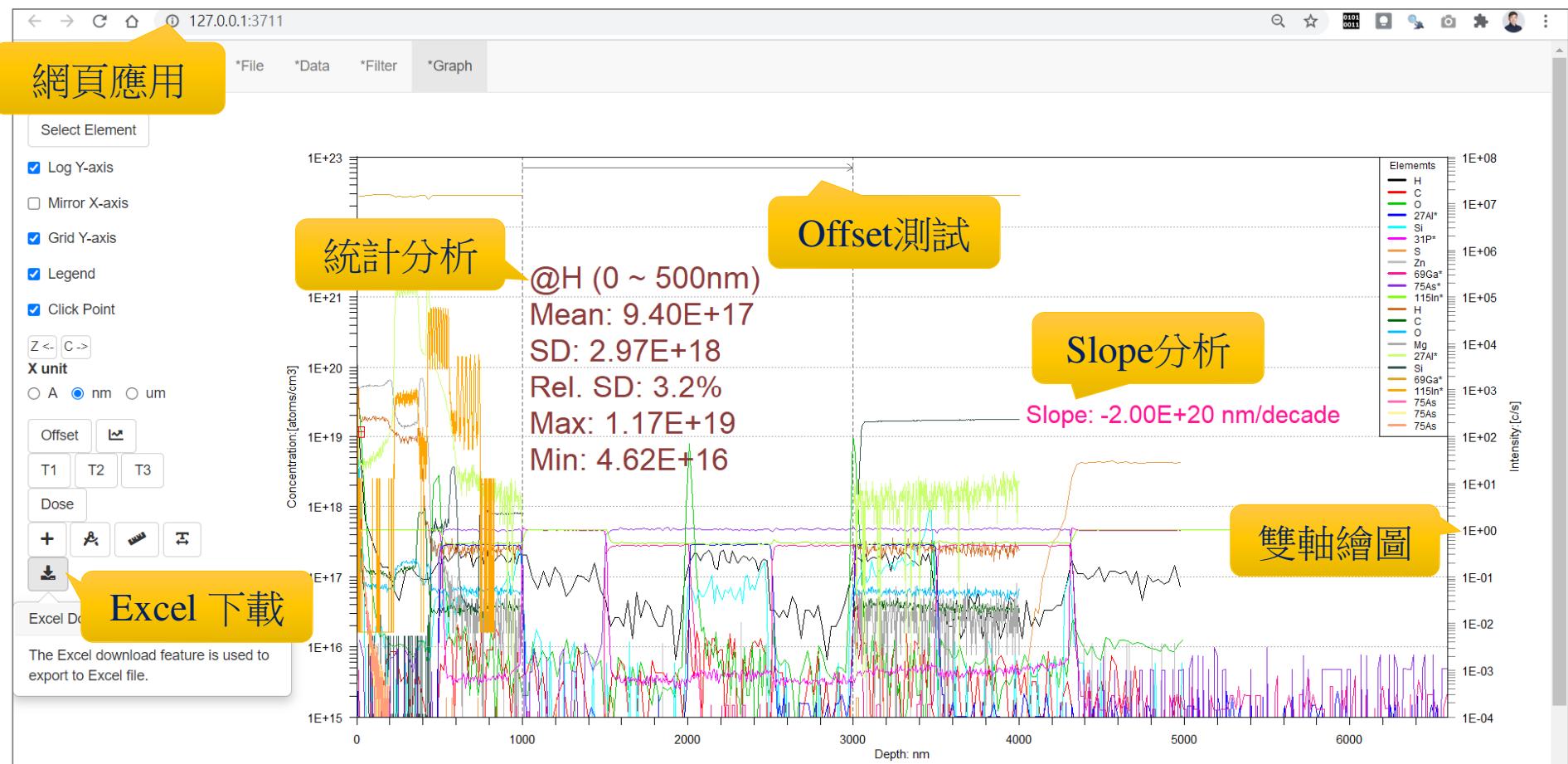
<http://rwepa.blogspot.com/2020/02/2019nCoV.html>

RWEPA - shiny 顧客連接分析

- <https://rwepa.shinyapps.io/shinyCustomerConnect/>



離子資料分析與視覺化應用



CRISP-DM標準流程

資料探勘生命週期 - CRISP-DM

- 跨產業資料探勘標準作業流程 (Cross Industry Standard Process for Data Mining)
- 資料探勘方法論
- CRISP-DM是於1990年起，由SPSS以及NCR兩大廠商在合作戴姆克萊斯勒-賓士(Daimler Benz)的資料倉儲以及資料探勘過程中發展出來的。

CRISP-DM 資料探勘流程(續)

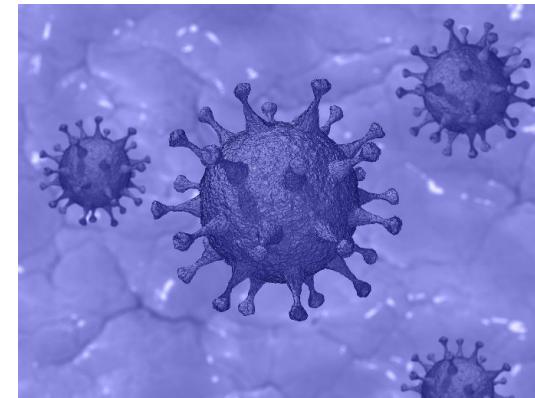
- 步驟 1：商業理解
- 步驟 2：資料理解
- 步驟 3：資料準備
- 步驟 4：模式建立
- 步驟 5：評估與測試
- 步驟 6：佈署應用

} 佔整專案時間的
70%~80%

- 訓練資料70%
- 測試資料30%

商業理解

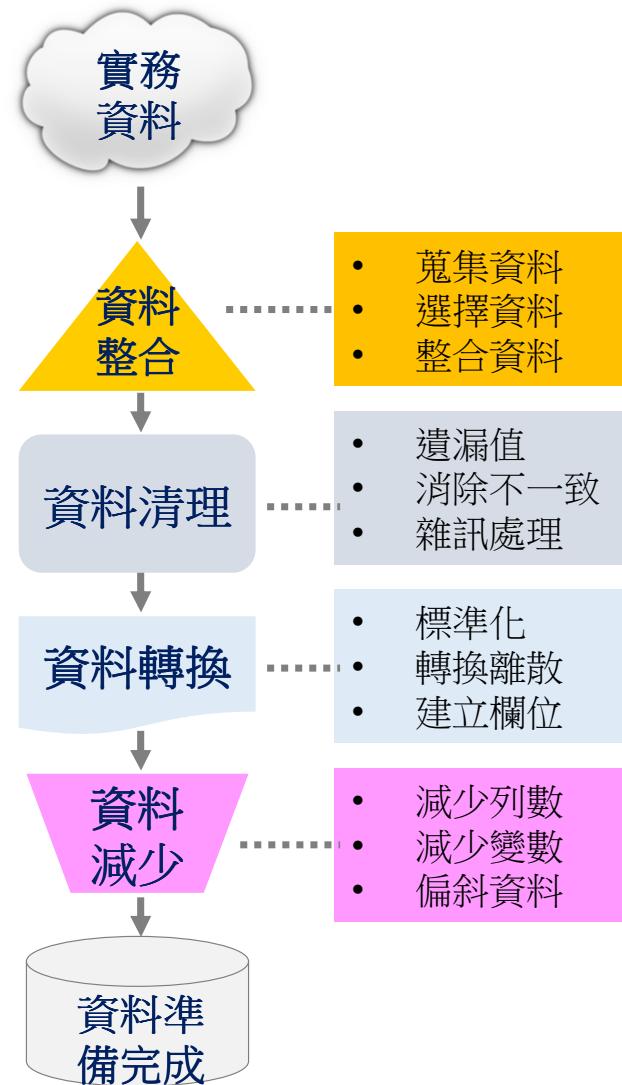
- 終極目標是要解決具體的產業問題，諸如提高購買率、找出詐欺交易、銷售預測與異常偵測等，因此以專業知識 (domain knowledge)進行商業理解是重要的第一步，處理重點：
 - 擬定商業目標
 - 進行當前處境評估
 - 決定資料探勘目標/成本
 - 產生專案計劃
 - 解決顧客問題



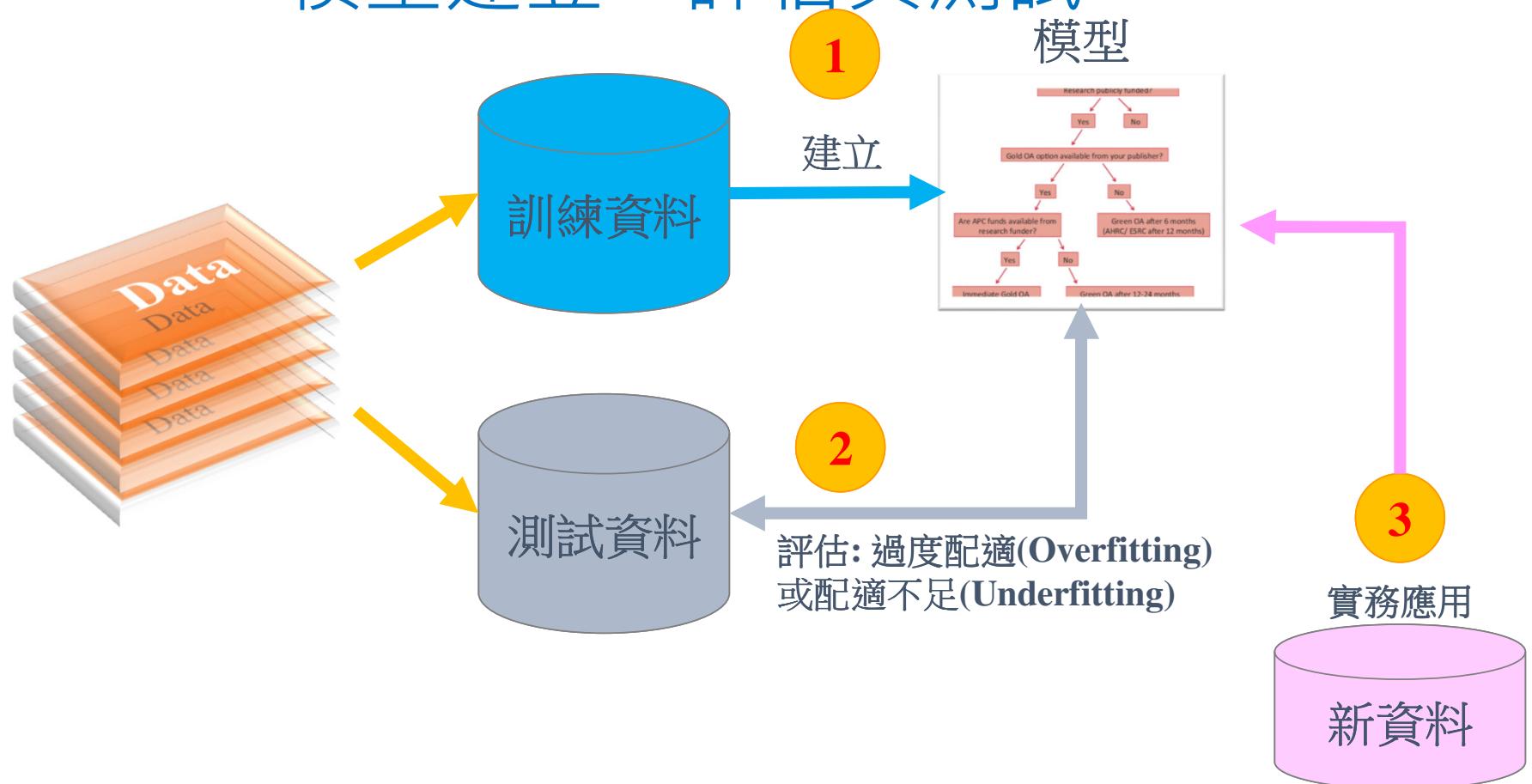
資料理解

- 包括描述資料、探索資料、核驗資料品質
- 敘述統計分析
 - 六力分析(summary函數)
- 繪圖
 - 依群組特性
 - 依時間特性
 - 新增評估欄位
- 趨勢
- 離群值 (outlier)
- 散佈圖、散佈圖矩陣
- 盒鬚圖

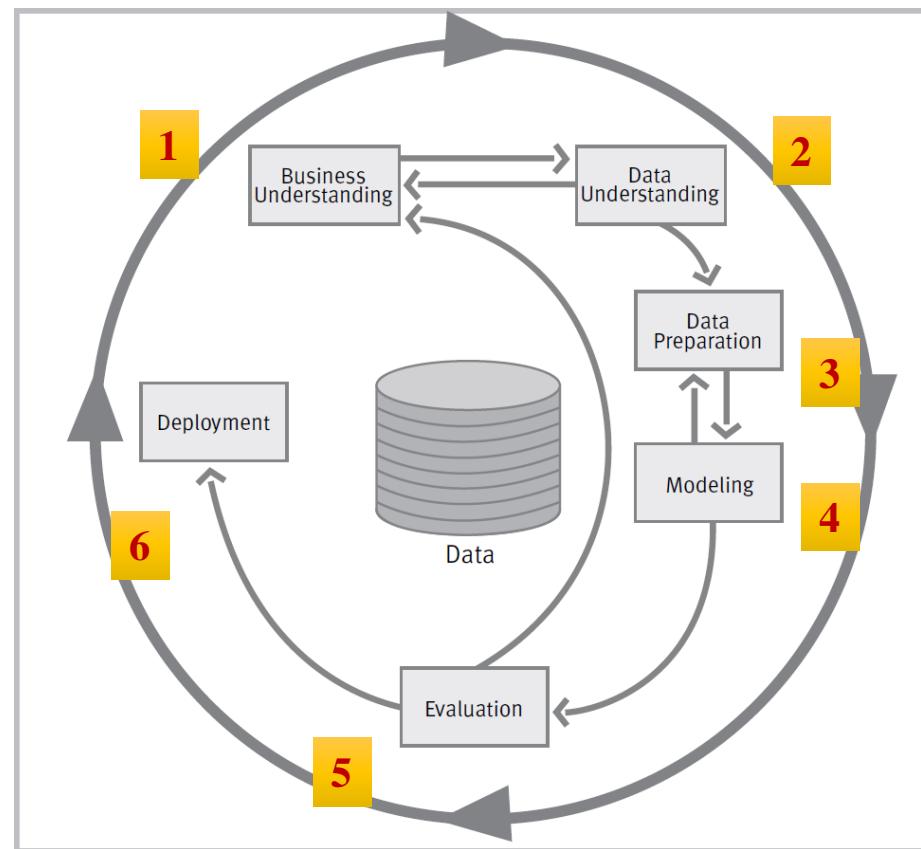
資料準備



模型建立、評估與測試



CRISP-DM 資料探勘流程(續)



參考 https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

數值模型績效指標

- 不可直接使用誤差的算術平均!

$$\text{Total error} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$$



- 均方誤差 (Mean Squared Error, MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- 均方根誤差 (Root Mean Squared Error, RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- 平均絕對誤差 (Mean Absolute Error, MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

類別模型績效指標

- <http://rwepa.blogspot.com/2013/01/rocr-roc-curve.html>

```
#          | 真實P類別 真實N類別
# ****|*****
# 預測P類別 | TP真陽數 FP假陽數
# 預測N類別 | FN假陰數 TN真陰數
# ****|*****
#          | P         N

# 1.TPR (True positive rate) 真陽性率，愈大愈好 -----
# =TP / (TP+FN)
# =TP / P
# =Sensitivity 積敏度
# =Recall 召回率
# =Probability of detection
# =Power
# 實際為陽性的樣本中，判斷為陽性的比例。
# 例如真正有生病的人中，被醫院判斷為有生病者的比例。
```

混淆矩陣
(Confusion Matrix)

Python 參考資料

RWEPA 搜尋此網誌 (例: task)

- GitHub DataDemo
- 關於作者
- R與實驗設計應用影片(6)
- ★★★R入門資料分析與視覺化(付費,中文字幕)
- ★★★R商業預測與應用(付費,中文字幕)
- iPAS-R-tutorial(繪圖中文字型solved)
- iPAS-Python-tutorial
- R教學-基礎篇/程式碼(免費)
- Python程式設計PDF(免費)

Python 程式設計-李明昌 免費電子書 - PDF 分享, 220頁



R 入門資料分析與視覺化應用(7小時28分鐘)

- <https://mastertalks.tw/products/r?ref=MCLEE>

課程提供教學範例的原始程式檔案與資料集



- **主題**
 1. R, RStudio簡介與套件使用
 2. 認識資料物件
 3. 資料處理與分析
 4. 資料視覺化應用
- **特色**
 1. 資料分析的**關鍵八步**
 2. 提供必備**ggplot2**套件的應用知識與使用情境
 3. 提供日期時間**zoo, xts**套件的整合應用操作
 4. 提供**人力資源**資料與**銷售**資料，強化**實務資料**操作能力

R 商業預測應用(8小時53分鐘)

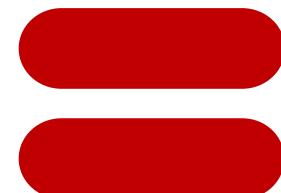
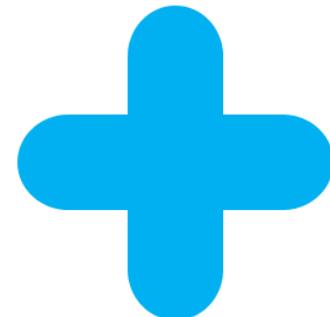
- <https://mastertalks.tw/products/r-2?ref=MCLEE>



課程提供教學範例的原始程式檔案與資料集

- **主題**
 1. R , RStudio工具操作
 2. 非監督式學習商業預測
 3. 監督式學習商業預測
 4. 財金資料預測應用
- **特色**
 1. 採用**最有效率**方式學習大數據R語言，並應用於**職場資料分析與商業預測應用**
 2. 提供**多元線性迴歸**的必備知識
 3. 提供**財金資料商業預測應用**的基礎與進階必學技能
 4. 提供學員人力資源資料與**台指期tick資料**預測演練

學習目標





機器學習

機器學習 Machine learning

- 監督式學習 (Supervised learning)
 - Telling the algorithm what to predict
- 非監督式學習 (Unsupervised learning)
 - No label or target value given for the data
- 半監督學習 (Semi-supervised learning)
 - 具有少量標記資料
- 強化學習 (Reinforcement learning)
 - 為了達成目標，隨著環境的變動，而逐步調整其行為，並評估每一個行動之後所到的回饋是正向的或負向的。
- 深度學習 (Deep learning)



監督式學習 vs. 非監督式學習

- 監督式學習 Supervised learning - 執行 $X \rightarrow$ 預測 $\rightarrow Y$
 - 迴歸分析 Regression analysis
 - 廣義線性模型 General linear model (GLM)
 - 天真貝氏法 Naïve-Bayes
 - K近鄰法 k-nearest neighbors (KNN)
 - 決策樹 Decision tree
 - 支持向量機 Support vector machine (SVM)
 - 類神經網路 Neural network (NN)
 - 集成學習 Ensemble learning: 使用多種學習算法來獲得比單獨使用演算法更好預測結果
- 非監督式學習 Unsupervised learning
 - 集群法 Clustering
 - 關聯規則 Association rule
 - 主成分分析 Principal Component Analysis





3.認證考試

iPAS巨量資料分析師簡介

<https://www.ipas.org.tw/bda/>

<<<一定要具備 Python, R 技能>>>

科目一：資料導向程式設計

評鑑主題	評鑑內容	建議命題內容
1. 資料架構 (50%)	1-1. 資料類型與物件 (15%)	<ul style="list-style-type: none"> R原生資料結構(向量、矩陣、陣列、資料框、串列)，及衍生資料結構等。 Python原生的串列(list)、值組(tuple)、字典(dict)、集合(set)及衍生的套件。 常見存放數據的交換格式等。
	1-2. 資料庫概念(含NoSQL) (20%)	<ul style="list-style-type: none"> 比較NoSQLDB與傳統關聯式資料庫概念上的差別。 MySQL、MongoDB在處理資料匯入及處理運算基礎指令等。
	1-3. 資料匯入與匯出 (15%)	<ul style="list-style-type: none"> Python與R，匯入匯出CSV, XML, JSON, SQL，以及1-1所列出的各式資料格式需注意的觀念及方法。
2. 程式實作基礎 (50%)	2-1. 程式設計類型 (15%)	<ul style="list-style-type: none"> R, Python等程式語言概念，含：合法的識別符號、R物件導向、Python物件導向。 R S3物件導向概念。 R與Python語言的向量化及隱式迴圈用法。
	2-2. 自訂函數與控制敘述 (20%)	<ul style="list-style-type: none"> R, Python比較自訂函數與預存程序的差異、純量函數、行內資料集函數、多敘述資料集函數、建立/修改/刪除自訂函數指令語法。
	2-3. 程式除錯與效能提升方法 (15%)	<ul style="list-style-type: none"> R語言與Python程式優化技巧、除錯方法。

科目二：資料處理與分析概論

評鑑主題	評鑑內容	建議命題內容
1. 資料處理 (40%)	1-1. 資料組織與清理 (10%)	資料正規化、編碼、歸戶、勾稽、資料交換格式、遺漏值/缺失處理、字串處理及正則表達式
	1-2. 資料摘要與彙總 (10%)	資料排序、資料群組與摘要等
	1-3. 屬性轉換與萃取 (10%)	類別型與數值型資料轉換處理、連續型與間斷型資料處理、衍生性欄位資料處理、資料合併等
	1-4. 巨量資料處理概念 (10%)	分散式運算概念、分散式儲存概念、HDFS特色MapReduce程式基本概念等
2. 資料分析 (60%)	2-1. 統計分析基礎 (20%)	隨機誤差建模概念、數值變數與類別變數的相關/獨立/共變異數等觀念與差異、抽樣方法與抽樣分配、估計與檢定、相似性與距離等
	2-2. 探索式資料分析與非監督式學習 (20%)	資料繪圖與製表、集群、頻繁型態分析、離群值分析等
	2-3. 線性模型與監督式學習 (20%)	各種線性迴歸方法適用情況、廣義線性模型、羅吉斯迴歸分類、集群方法與應用、關聯規則等



科目一：資料導向程式設計

1-1 資料類型與物件

1-1 資料類型與物件

2. 在 Python 語言中，已知「`a=numpy.array([[1, 3],[2, 4]])`」，則`3*a` 意義為何？

- (A) $a \cdot a \cdot a$
- (B) 3 乘以每個元素
- (C) 3 乘以第 1 行元素
- (D) 3 乘以第 3 行元素

```
In [1]: import numpy  
  
In [2]: a=numpy.array([[1, 3],[2, 4]])  
  
In [3]: a  
Out[3]:  
array([[1, 3],  
       [2, 4]])  
  
In [4]: 3*a  
Out[4]:  
array([[ 3,  9],  
       [ 6, 12]])
```



ROCR package - ROC curve

- RWEPA → rocr
- <http://rwepa.blogspot.com/2013/01/rocr-roc-curve.html>

2020.3.23 中英文對照,Kappa統計量更新:

分類預測模型評估-混淆矩陣 (Confusion matrix)

混淆矩陣可用於監督式分類模型評估

	真實P類別	真實N類別
--	-------	-------

預測P類別 | TP真陽數 FP假陽數

預測N類別 | FN假陰數 TN真陰數

	P	N
--	---	---

1. TPR (True positive rate) 真陽性率, 愈大愈好
 $=TP / (TP+FN)$

$=TP/P$

$=Sensitivity$ 瞩敏度

$=Recall$ 召回率

$=Probability\ of\ detection$

$=Power$

實際為陽性的樣本中，判斷為陽性的比例。

例如真正有生病的人中，被醫院判斷為有生病者的比例。

1-1 資料類型與物件

1. 下方的混淆矩陣為影像分類模型對 100 張動物相片進行辨識後的結果，試問該分類模型的準確率（Accuracy）為何？

預測 實際	狗	狼
狗	44	6
狼	8	42

- (A) 0.82
- (B) 0.84
- (C) 0.86
- (D) 0.88

- $Accuracy = \frac{44+42}{44+6+8+42} = \frac{86}{100} = 0.86$
- $Recall = \frac{44}{44+6} = \frac{88}{100} = 0.88$

1-1資料類型與物件

5. 參考附圖，R 語言中，關於 mydf 資料物件，下列敘述何者「不」正確？

```
> mydf <- head(iris)
> mydf
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

- (A) class(mydf)的結果為"data.frame"
- (B) mydf[1]的結果為取出第 1 列元素值
- (C) mydf[, 1]的結果為取出第 1 行的元素值
- (D) mydf[6, 4]的結果為 0.4

1-1 資料類型與物件

7. 參考附圖，Python 語言中，關於 pandas 套件，下列敘述何者正確？

```
In [1]: import pandas as pd  
  
In [2]: import numpy as np  
  
In [3]: df = pd.DataFrame(np.random.randn(3,4),  
...:                      index=pd.date_range('20200101', periods=3),  
...:                      columns=list('ABCD'))  
  
In [4]: df  
Out[4]:  
          A         B         C         D  
2020-01-01  0.317269 -0.256608  0.440586  0.309337  
2020-01-02  0.835061 -1.025774 -1.546098  1.860604  
2020-01-03 -0.889885  1.135232  0.176671  0.904192
```

- R: summary(df)
- Python: df.describe()

- (A) type(df)的結果會顯示 pandas.core.series.Series
- (B) df.summary()的結果會顯示 count, mean, std 等統計值
- (C) df.values 的結果會取出 df 資料物件的指標
- (D) df.sort_values(by='C')的結果是依照 C 欄數值大小，由小至大排序整個表格



科目一：資料導向程式設計

1-2資料庫概念(含NoSQL)

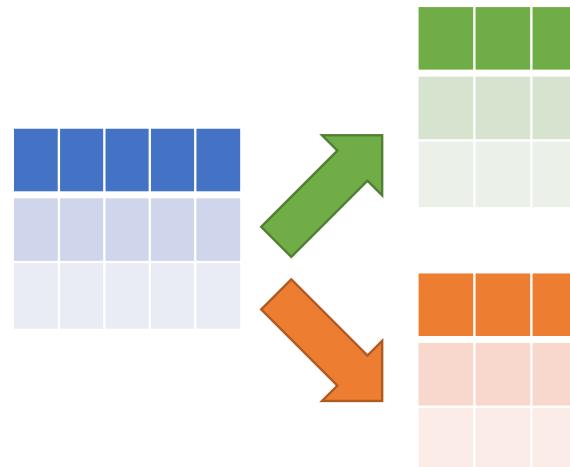
1-2資料庫概念(含NoSQL)

13. 在資料表正規化的過程（1NF 到 BCNF）中，每一個階段都是以欄位的「相依性」作為分割資料表的依據之一，關於正規化步驟之敘述，下列何者不正確？
- (A) 第一正規化型式：除去重覆群
 - (B) 第二正規化型式：除去部分相依
 - (C) 第三正規化型式：除去遞移相依
 - (D) Boyce-Codd 正規化型式：除去多值相依

• 第四正規化型式：
除去多值相依

資料庫的正規化 (Normalization)

- 第一階正規化的規則
 - 1. 資料表中必須有 Primary Key
 - 2. 每個欄位中都只儲存單一值
 - 3. 資料表中沒有意義相同的多個欄位
- 第二階正規化的規則
 - 1. 必須符合 1NF 的格式
 - 2. 各欄位與 Primary Key 間沒有『部分相依』的關係
- 第三階正規化的規則
 - 1. 符合 2NF 的格式
 - 2. 各欄位與 Primary Key 間沒有『間接相依』的關係, 即沒有『遞移相依』的關係
- Boyce-Codd 正規化
 - 1. 符合 2NF 的格式
 - 2. 各欄位與 Primary Key 沒有『間接相依』的關係
 - 3. Primary Key 中的各欄位不可以相依於其他非Primary Key 的欄位



1-2資料庫概念(含NoSQL)

15. 您是一位巨量資料分析師，關於 NoSQL 資料庫的設計特性，下列敘述何者不正確？
- (A) 使用 Key-Value 資料模式
 - (B) 資料查詢透過 API 查詢
 - (C) Google Big Table 為 NoSQL 類型
 - (D) 需預先設計固定的 Schema 欄位

1-2資料庫概念(含NoSQL)

16. 以下何者不是 NoSQL 資料庫？
- (A) MongoDB , CouchDB
 - (B) Apache Cassandra , Apache HBase
 - (C) Redis , Memcached
 - (D) MariaDB , SQLite

Memcached:

- distributed memory object caching system
- an in-memory key-value store

1-2資料庫概念(含NoSQL)

- 鍵值資料庫: Apache Cassandra, Redis, Aerospike, BigTable
- 記憶體資料庫: Memcached, IBM – DB2 BLU, Oracle Berkeley DB, Oracle TimesTen
- 文件導向資料庫 CouchDB, MongoDB
- 圖學資料庫: Neo4j (以圖為資料結構儲存和查詢, 不是圖片)

SQL - SELECT

SELECT *select_list* [INTO *new_table*]

[**FROM** *table_source*] [**WHERE** *search_condition*]

[**GROUP** BY *group_by_expression*]

WHERE 搜尋條件

[**HAVING** *search_condition*]

[**ORDER BY** *order_expression* [ASC | DESC]]

ORDER BY 排序

參考: <https://docs.microsoft.com/zh-tw/sql/t-sql/queries/select-transact-sql?view=sql-server-2019>

1-2資料庫概念(含NoSQL)

15. 參考附圖之資料表，請問執行各選項中之語句，何者「無法」得到含有'Spencer'的結果？

id	height	weight	userName	gender	age
1	158	52	Cathy	female	33
2	180	78	Spencer	male	40
3	155	49	Jen	female	28
4	145	40	Rudy	female	16
5	173	70	Johnson	male	26

- (A) SELECT userName FROM userTable WHERE age > 30 AND LENGTH(userName) > 5
- (B) SELECT userName FROM userTable WHERE height >= 173
- (C) SELECT userName FROM userTable WHERE (lower(userName) LIKE 's%')
- (D) SELECT userName FROM userTable WHERE weight >= 50 and gender != 'male'

(A) AND : 表示二個條件皆須為真
 (B) 身高大於或等於173
 (C) 函數 lower : 轉換成小寫
 LIKE 's%' : 相似於小寫s開頭字串
 (D) 運算子 != : 不等於



科目一：資料導向程式設計

1-3資料匯入與匯出

1-3資料匯入與匯出

29. 在 Python 語言中可用 `open()`函數開啟檔案，若僅要`讀取檔案`，mode 引數需使用下列何項設定值？
- (A) a
 - (B) r
 - (C) w
 - (D) wb

檔案處理 – open (Python內建69個標準函數)



- myfile = open(filename, mode)

mode feature

r 只能讀取, 例: df = open('data.csv', mode = 'r'), r 表示 read, mode預設為r

w 新建檔案寫入資料(檔案可以不存在, 若存在則清空)

a 將資料附加到舊檔案最後面位置

r+ 讀取與寫入(檔案需存在且游標指在開頭)

w+ 清空檔案內容, 寫入資料並可讀出(檔案如果不存在, 會自行新增)

a+ 資料附加到原檔案後面, 可讀取資料

1-3資料匯入與匯出

31. 關於 Python 語言讀取檔案，下列敘述何者不正確？

- (A) 使用 `open("file", "r")`若 file 不存在，會創建 file
- (B) 使用 `open("file", " w+ ")`若 file 不存在，會創建 file
- (C) 使用 `open("file", " w+ ")`為可讀寫的檔案模式
- (D) 使用 `open("file", " a+ ")`為附加讀寫檔案模式，寫入時添加於後，亦可以讀檔

1-3資料匯入與匯出

17. 關於R語言與Python語言，下列敘述何者正確？

- (A) R語言中，無法使用 `write.table()`函數匯出.csv逗號分隔檔案
- (B) Python語言中，可使用 `numpy` 的`to_csv()`方法匯出.csv逗號分隔檔案
- (C) R語言中，可使用 `write.csv()`函數匯出.csv逗號分隔檔案
- (D) Python語言中，可使用 `open("file.csv", "r")`讀寫.csv逗號分隔檔案

- (B) Python: `pandas.to_csv()`
- (D) 參考 `help(open)`



科目一：資料導向程式設計

2-1 程式設計類型

2-1 程式設計類型

25. 下列為 Python 程式的物件導向概念：

```
class C1:
```

```
    a = 10
```

```
class C2(C1):
```

```
    a = 20
```

```
c2 = C2()
```

Python 用法：

class 子類別(父類別)

請問 c2.a 的值為下列何者？

(A) 0

(B) 10

(C) 20

(D) 40

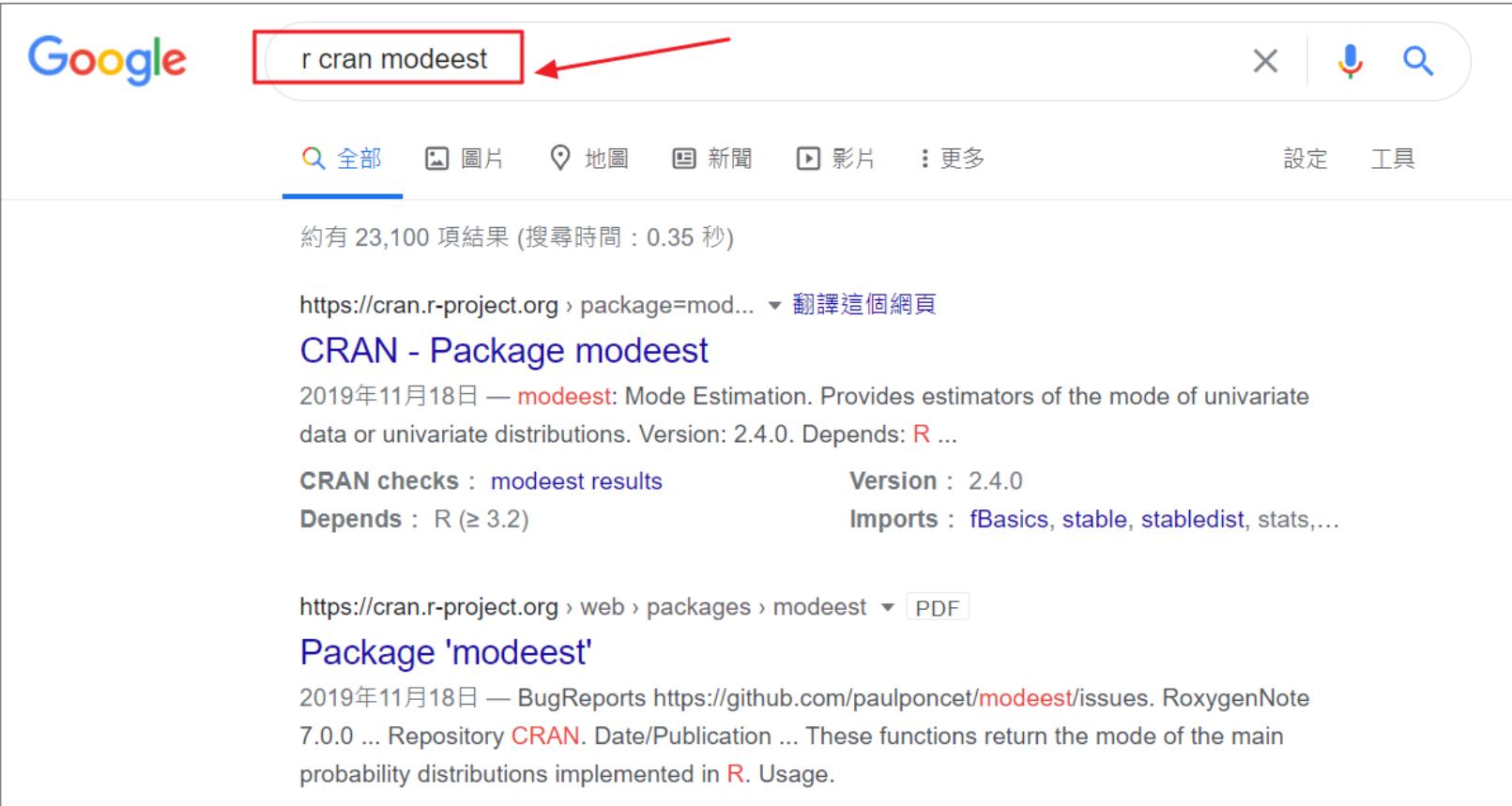
2-1 程式設計類型

22. 在R語言的{modeest}套件中，下列何者為具眾數估計功能的函數？

- (A) var()
- (B) sd()
- (C) mlv()
- (D) quantile()

- 技巧: 本題使用消除法
- (A) var : variance 變異數
- (B) sd: Standard Deviation 標準差
- (D) quantile 百分位數

r cran modeest



Google r cran modeest ← X |  

全部 圖片 地圖 新聞 影片 更多 設定 工具

約有 23,100 項結果 (搜尋時間 : 0.35 秒)

[https://cran.r-project.org › package=mod... ▾ 翻譯這個網頁](https://cran.r-project.org/package=modeest)

CRAN - Package modeest

2019年11月18日 — **modeest**: Mode Estimation. Provides estimators of the mode of univariate data or univariate distributions. Version: 2.4.0. Depends: R ...

CRAN checks : modeest results **Version :** 2.4.0
Depends : R (≥ 3.2) **Imports :** fBasics, stable, stabledist, stats,...

[https://cran.r-project.org › web › packages › modeest ▾ PDF](https://cran.r-project.org/web/packages/modeest)

Package 'modeest'

2019年11月18日 — BugReports <https://github.com/paulponcet/modeest/issues>. RoxygenNote 7.0.0 ... Repository CRAN. Date/Publication ... These functions return the mode of the main probability distributions implemented in R. Usage.

modeest 單變數之眾數估計

CRAN - Package modeest

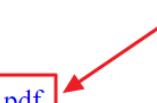
cran.r-project.org/web/packages/modeest/index.html

modeest: Mode Estimation

Provides estimators of the mode of univariate data or univariate distributions.

Version: 2.4.0
Depends: R (≥ 3.2)
Imports: [fBasics](#), [stable](#), [stabledist](#), stats, [statip](#) ($\geq 0.2.3$)
Suggests: [evd](#), [knitr](#), [mvtnorm](#), [testthat](#), [VGAM](#)
Published: 2019-11-18
Author: Paul Poncet [aut, cre]
Maintainer: Paul Poncet <paulponcet at yahoo.fr>
BugReports: <https://github.com/paulponcet/modeest/issues>
License: [GPL-3](#)
URL: <https://github.com/paulponcet/modeest>
NeedsCompilation: no
Materials: [README](#) [NEWS](#)
In views: [Distributions](#)
CRAN checks: [modeest results](#)

Downloads:

Reference manual: [modeest.pdf](#) 
Package source: [modeest_2.4.0.tar.gz](#)

mlv {modeest}

16 / 32 | - 138% + | ☰ ⚡

```
mlv(x, method = "meanshift", par = mean(x))
```

mlv

Estimation of the Mode(s) or Most Likely Value(s)

Description

mlv is a generic function for estimating the mode of a univariate distribution. Different estimates (or methods) are provided:

- **mfv**, which returns the most frequent value(s) in a given numerical vector,

28. Python 語言中，執行附圖程式碼後，請問選項中何者之執行結果為布林值 False？

```
class Car:
```

```
    def __init__(self):
```

```
        self.name = 'Normal Car'
```

```
    def run(self):
```

```
        return 'Car is running'
```

```
CarA = Car()
```

```
CarB = Car()
```

(A) type(CarA) == type(CarB)

(B) CarA.name == CarB.name

(C) CarA == CarB

(D) CarA.run() == CarB.run()

a) True

b) True

c) False

d) True



科目一：資料導向程式設計

2-2自訂函數與控制敘述

2-2自訂函數與控制敘述

41. 關於 Python，下列敘述何者不正確？
- (A) range(10)可以產生 0, 1, ...9 的數字序列
 - (B) continue 語法可以進入下一次循環
 - (C) pass 語法可以產生判斷邏輯
 - (D) 使用 if...else 進行條件判斷

Python 流程控制：

- **break** : 強制離開整個迴圈
- **continue** : 強制離開本次迴圈，繼續進入下一個圈
- **pass** : 不做任何事情，沒有回傳值，程式繼續執行

2-2自訂函數與控制敘述

33. Python 語言中，選項中何者呼叫函數 myfunc 最可能發生錯誤？

- (A) myfunc(0)
- (B) myfunc(0,b=1)
- (C) myfunc(0,1,2,3)
- (D) myfunc(0,b=1,2)

2-2自訂函數與控制敘述

35. 關於Box-Cox 轉換，下列敘述何者正確？

- (A) 是一種幕次方轉換，適用於當變數值恆正的時候
- (B) 是一種倒數轉換，適用於當變數值恆負的時候
- (C) 將對稱分布轉為偏斜分佈
- (D) 適用於對稱分佈

- 參考: https://en.wikipedia.org/wiki/Power_transform
- 功能: 將資料轉換為常態分配

2-2自訂函數與控制敘述

40. 關於 R 語言的自訂函數，下列敘述何者不正確？

- (A) 定義自訂函數使用關鍵字 define
- (B) 自訂函數的參數個數可以為零
- (C) 自訂函數中有回傳值時可使用 return
- (D) 自訂函數支援遞迴 (recursion)

自訂函數

- R : function { ... }
- Python : def
- Julia : function ... end

2-2自訂函數與控制敘述

41. 下列 R 語言 addLog 函數的敘述，何者正確？

```
1 addLog <- function(number1, number2) {  
2   number1 + log(number2)  
3 }
```

- (A) 在主控台執行 `addLog(number2=exp(4), number1=1)` 的結果會顯示 Error
- (B) 在主控台執行 `addLog(exp(4), number1=1)` 的結果會顯示 5
- (C) 在主控台執行 `addLog(1, exp(4))` 的結果會顯示 Error
- (D) 在主控台執行 `addLog(number2=exp(4), n1=1)` 的結果會顯示 5

R – function 函數

```
> f <- function() {  
+   u <- 1  
+   v <- 2  
+   u+v  
+ }  
> # (A)  
> f  
function() {  
  u <- 1  
  v <- 2  
  u+v  
}  
> # (B) (C)  
> u  
Error: object 'u' not found  
> # (D)  
> u + v  
Error: object 'u' not found  
> f()  
[1] 3  
>
```

- 函數內宣告的變數是區域變數.
- 執行函數須加上左右括號 f()



科目一：資料導向程式設計

2-3 程式除錯與效能提升方法

2-3 程式除錯與效能提升方法

37. 關於程式碼錯誤處理，下列敘述何者不正確？

- (A) 撰寫程式時讓錯誤可以重製，可便於偵查
- (B) 可執行的程式即為無誤的程式
- (C) 可先要求程式碼的邏輯正確，再設法提升執行效率
- (D) 可藉由函數輔助處理例外狀況與瞭解錯誤訊息

- A. 重製(reproduction): 方便程式碼除錯
- B. 可執行的程式，**有可能**特殊情形沒考慮，例：輸入
數值欄位進行運算，改為字元輸入時，會有錯誤
- C. 商業的**邏輯正確性**應優先考量
- D. 使用函數處理例外，方便理解錯誤之情形

2-3 程式除錯與效能提升方法

39. 在 R 語言中，程式碼執行時會產生訊息，下列敘述何者不正確？

- (A) 產生錯誤（errors）訊息時，將繼續執行無錯誤的程式
- (B) 警告（warnings）訊息會說明潛在的問題
- (C) 一般傳回的訊息在於說明代碼輸出的結果
- (D) 產生警告（warnings）訊息時，可繼續執行程式

- 程式錯誤 → 中止
- 程式警告 → 繼續執行

2-3程式除錯與效能提升方法

43. 「例外狀況處理函數」之目的為允許程式設計師在例外狀況發生時，採取某些行動，例如：結合 `stop()`函數，以讓程式設計師在必要的時候傳回訊息，瞭解程式碼執行的狀況。下列何者不為 R 語言的例外狀況處理函數？
- (A) `withCallingHandlers()`
 - (B) `suppressMessages()`
 - (C) `tryCatch()`
 - (D) `try()`

R: suppressMessages 隱藏(壓制)訊息

message {base}

R Documentation

Diagnostic Messages

Description

Generate a diagnostic message from its arguments.

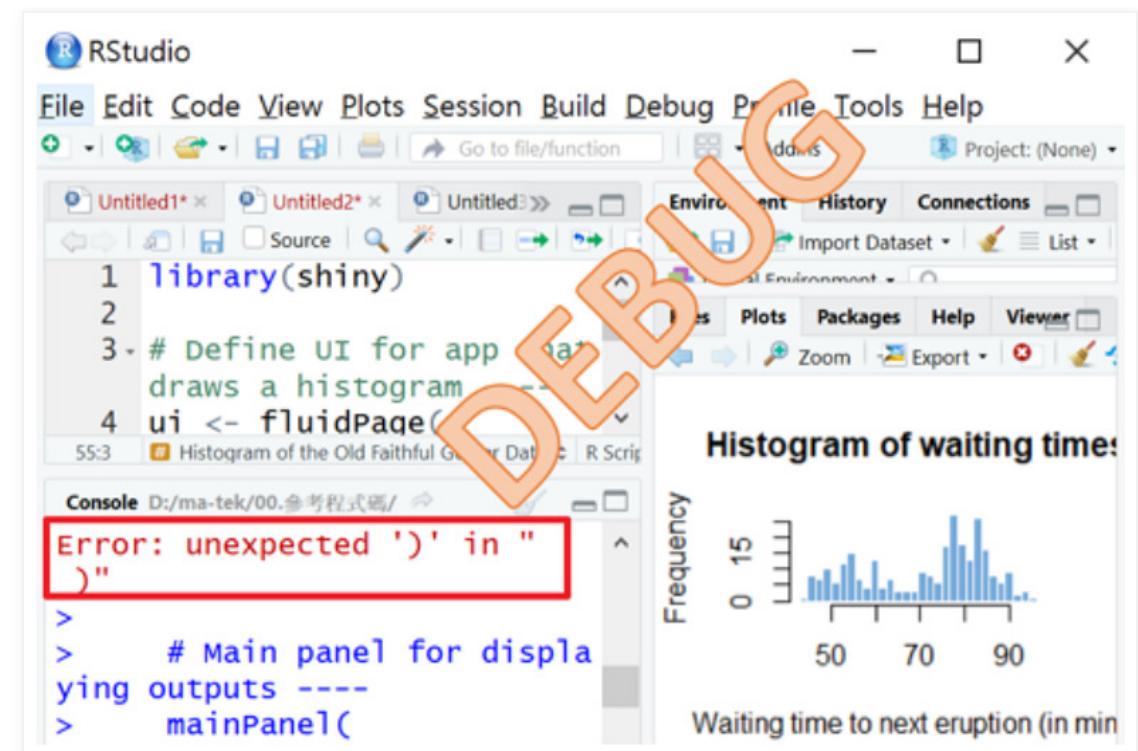
Usage

```
message(..., domain = NULL, appendLF = TRUE)
suppressMessages(expr, classes = "message")
```

```
packageStartupMessage(..., domain = NULL, appendLF = TRUE)
suppressPackageStartupMessages(expr)
```

```
.makeMessage(..., domain = NULL, appendLF = FALSE)
```

R程式除錯與效能提升設計



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Untitled1* Untitled2* Untitled3... Environment History Connections

Source Go to file/function

1 library(shiny)
2
3 # Define UI for app that
4 ui <- fluidPage(

Error: unexpected ')' in "
)"
>
> # Main panel for displaying outputs ----
> mainPanel(

Histogram of waiting times

Frequency

Waiting time to next eruption (in min)



科目二：資料處理與分析概論

1-1 資料組織與清理

1-1 資料組織與清理

5. 關於資料之遺缺值處理，下列何者不正確？

- (A) 無須考慮遺缺值比例，全部刪除
- (B) 類別資料補上眾數之值
- (C) 利用模型補上估計產生之值
- (D) 透過差值法（interpolation method）補上該值

1-1 資料組織與清理

1. 在分析資料前，通常需要先清理資料。當數字與文字混合在一起時，但我們僅需要提取出數字時，若以逐筆資料提取十分曠日廢時，在 Python 語法中的套件 `re` 可以處理大部分的此類問題，例如語法：`re.findall(pattern, string)`，當 `pattern = '\d'`，可以提取出 `string` 中所有單一數字；`pattern = '\d\d'`，可以提取出 `string` 中所有 2 個相連數字；`pattern = '\d+'`，可以提取出 `string` 中所有任意相連個數的數字。請問當 `string = '王大明手機號碼:0912334567,地址...'` 時，下列何者語法無法提取出 0912334567？

- (A) `re.findall('\d+', string)`
- (B) `re.findall('\d\d\d\d\d\d\d\d\d\d', string)`
- (C) `re.findall('\d\d+', string)`
- (D) `re.findall('\d\d\d\d\d\d\d+', string)`

1-1 資料組織與清理

2. 下列何者「不」是資料前處理該進行的程序？

- (A) 資料清理 (data cleaning)
- (B) 資料轉換 (data transform)
- (C) 屬性挑選 (feature selection)
- (D) 資料建模 (data modeling)

1-1 資料組織與清理

4. 關於資料具有離群值（outlier），進行資料標準化時，下列敘述何者較為適合？

- (A) 可採用 Z-分數法 (Z-score)
- (B) 可採用最小最大正規化法 (min-max normalization)
- (C) 可採用穩健縮放法 (robust scaler)
- (D) 可採用最大絕對值縮放法 (maximum absolute scaler)

$$x' = \frac{x - \text{medina}}{IQR}$$

- Robust scalar: 可以有效的縮放帶有outlier的數據，透過Robust如果數據中含有異常值在縮放中會捨去。
- Scale features using statistics that are robust to outliers. This Scaler removes the median and scales the data according to the quantile range (defaults to IQR: Interquartile Range). The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile).
- <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html>

1-1 資料組織與清理

5. 參考附圖，Python 語言中，關於使用 numpy 套件處理遺缺值（missing value），下列敘述何者「不」正確？

```
import numpy as np
```

- (A) np.nan 執行結果為 nan
- (B) np.isnan(np.nan) 執行結果為 True
- (C) np.NaN 執行結果為 NaN
- (D) np.isnan(np.NaN) 執行結果為 True

• (C) 正確答案為 nan

資料正規化

資料正規化(Normalization)

- 監督式學習的特性：
 - 最小化誤差: 使模型擬合訓練資料, 如: cost function- error term
 - 正規化參數: 目的是為了防止過度擬合, 如: cost function- penalty term
 - 參數過多會導致模型複雜度上升，產生過度擬合，即訓練集誤差很小，但測試集誤差很大。
 - 目標是在簡單模型的基礎上，最小化訓練誤差，使模型具有更好的泛化能力（即測試誤差也很小）。
 - 正規化可使用 L1 norm, L2 norm 方法

L1 Norm vs. L2 Norm

- **L1-norm** is also known as least absolute deviations (LAD) or least absolute errors (LAE).
 - It is basically minimizing the sum of the **absolute** differences (S) between the target value (Y_i) and the estimated values $f(x_i)$

$$S_{L1} = \sum_{i=1}^n |y_i - f(x_i)|$$

- **L2-norm** is also known as least squares.
 - It is basically minimizing the sum of the **square** of the differences (S) between the target value (Y_i) and the estimated values $f(x_i)$

$$S_{L2} = \sum_{i=1}^n (y_i - f(x_i))^2$$

L1 vs. L2 norm

- L1 norm: 使用 β 純絕對值

Cost function for lasso regression = $(y - X\beta)^T(y - X\beta) + \lambda|\beta|$

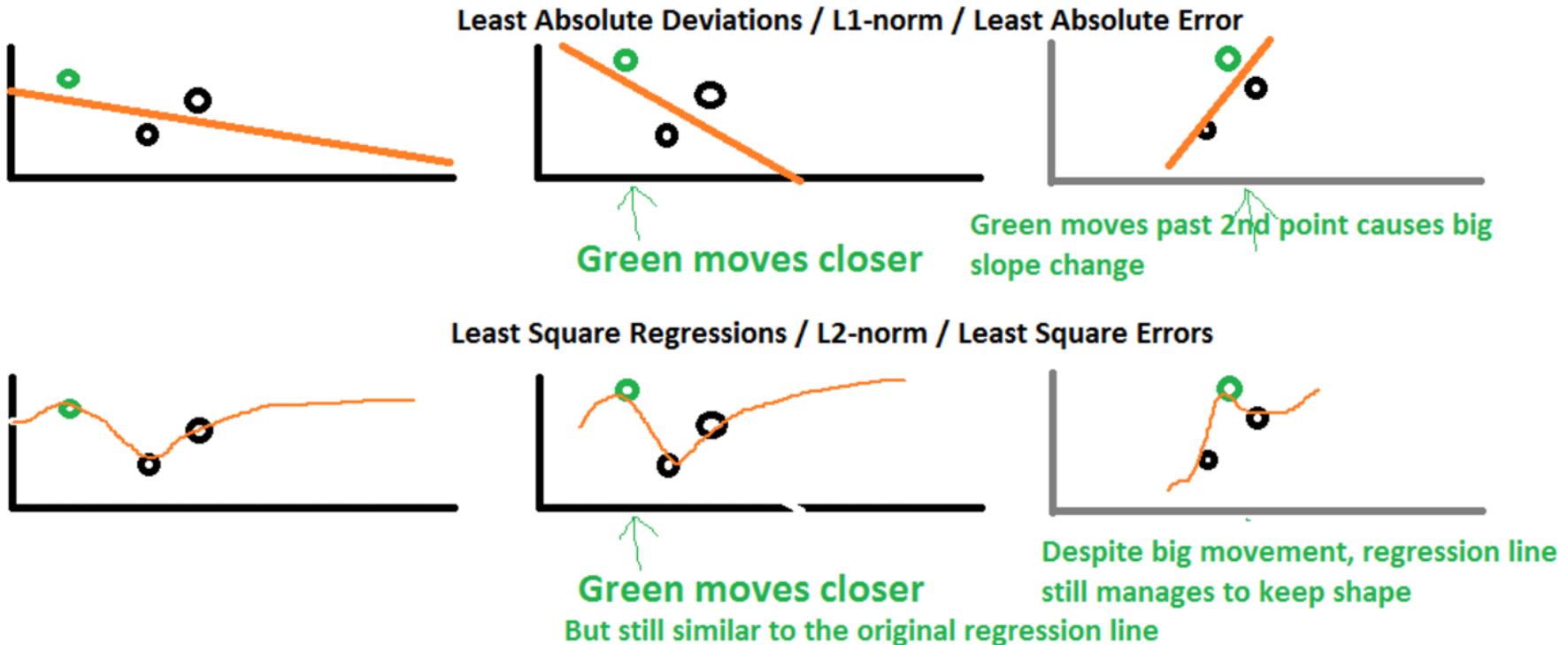
- L2 norm: 使用 β 平方

Cost function for logistic regression = $(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$

特性	L1 norm	L2 norm
Robustness 穩健性	佳	
Stability 穩定性(horizontal adjustments)		佳
Computational difficulty 計算複雜度		大
Sparsity 稀疏性	大 (讓模型不重要的參數變成 0)	

參考資料 <https://www.kaggle.com/residentmario/l1-norms-versus-l2-norms>

L1 Norm vs. L2 Norm (續)



參考: <http://www.chioka.in/differences-between-the-l1-norm-and-the-l2-norm-least-absolute-deviations-and-least-squares/>

資料標準化

資料標準化(Standardization)

- 資料標準化

- 將資料按比例進行線性轉換
- 將資料進行非線性轉換
- 使資料落在某一特定的區間,例如: $[0, 1]$ 之間
- 適用於資料有差異範圍

- 資料標準化之目的:

- 提升模型的收斂速度
- 提高模型的精準度
- 適用於主成分分析, 集群法, KNN, SVM, Logistic regression

非監督式學習

監督式學習

資料標準化

- (0,1)標準化: 將資料轉換至[0, 1]區間

$$X_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- 最小最大標準化(min-max normalization)

$$X_{new} = x_{new_min} + \frac{(x - x_{min})}{(x_{max} - x_{min})} \times (x_{new_max} - x_{new_min})$$

- Z-score標準化: 將任意資料轉換為趨近平均值為0, 標準差為1的分配, **結果一定是常態分配?**

$$X_{new} = \frac{x - \bar{x}}{\sigma}, \bar{x} \text{ 平均值, } \sigma \text{ 標準差}$$

Z-score標準化結果一定是常態分配 → 假新聞

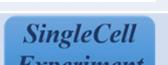
- GitHub DataDemo
- ★★★R入門資料分析與視覺化(付費,中文字幕)
- ★★★R商業預測與應用(付費,中文字幕)
- **iPAS-R-tutorial**
- iPAS-Python-tutorial
- R教學-基礎篇/程式碼(免費)
- Python程式設計PDF(免費)

```
set.seed(168)
x <- runif(10000)
hist(x-mean(x) / sd(x))
```

Python 模組

模組	功能	
Numpy	Large, multi-dimensional arrays and matrices	
Scipy	Optimization, linear algebra, integration, interpolation, FFT, signal and image processing	
Pandas	DataFrame object for data manipulation	
Matplotlib	Static, animated, and interactive visualizations	
Statsmodels	Statistical models	
Scikit-learn	Machine learning library	
Tensorflow	Deep learning	
Biopython	Biological computation	
Scanpy	Single-cell analysis	

R 套件

模組	功能	
dplyr	A grammar of data manipulation	
data.table	Extension of data.frame	
ggplot2	Create Elegant Data Visualizations Using the Grammar of Graphics	
shiny	Web application framework	
caret	Classification and Regression Training	
mlr3	Provides R6 objects for efficient, object-oriented programming on the building blocks of machine learning	
Tensorflow	R Interface to 'TensorFlow'	
Bioconductor	Tools for the analysis genomic data	
SingleCellExperiment	Orchestrating Single-Cell Analysis with Bioconductor	

參考資料

- RWEPA
 - <http://rwepa.blogspot.com/>
- Python 程式設計-李明昌 <免費電子書>
 - <http://rwepa.blogspot.com/2020/02/pythonprogramminglee.html>
- R入門資料分析與視覺化應用教學(付費)
 - <https://mastertalks.tw/products/r?ref=MCLEE>
- R商業預測與應用(付費)
 - <https://mastertalks.tw/products/r-2?ref=MCLEE>

謝謝您的聆聽

Q & A



李明昌

alan9956@gmail.com

<http://rwepa.blogspot.tw/>