資料科學概論筆記

Author: Uriah Xuan (109 NTUT CSIE)

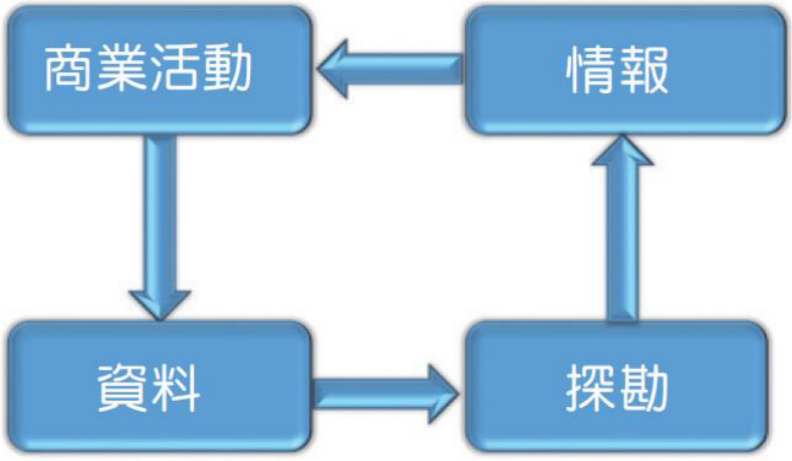
大數據的 5V 特徵

英文名稱	中文名稱	概念
Volume	大量化	大數據通常有著極大量的資料，因此需要注重保存與安全性
Variety	多樣化	大數據的資料形式其實是不限的，文字、圖片等等
Velocity	快速化	大數據的資料處理需要在一定時限內處理完成
Veracity	真實性	大數據的資料需要有一定的真實性。
Value	價值化	資料帶有價值意義，需要合理的運用，以低價值創造高價值

經營管理及數據導向決策問題

1. 決策時缺乏適合與即時有效的資訊參考。
2. 報表不夠齊全，格式與決策層需要的不同。
3. 花費大量人物力，製作使用率低與無效的報表。
4. 報表與管理規範結合困難，經常發生異常狀況。
5. 無法從報表分析產生新管理觀念，找出成功關鍵因素與核心競爭力。

商業智慧與資料探勘的循環



Business Intelligence Development Model

（待補）

辨識模式

模式有助於解析複雜事務與展露趨勢、種類為時間、空間、功能。

帕金森定理（Parkinson's Law）

在工作能被完成的時限內，工作量會一直被增加，直到所有可用時間都被填充為止。

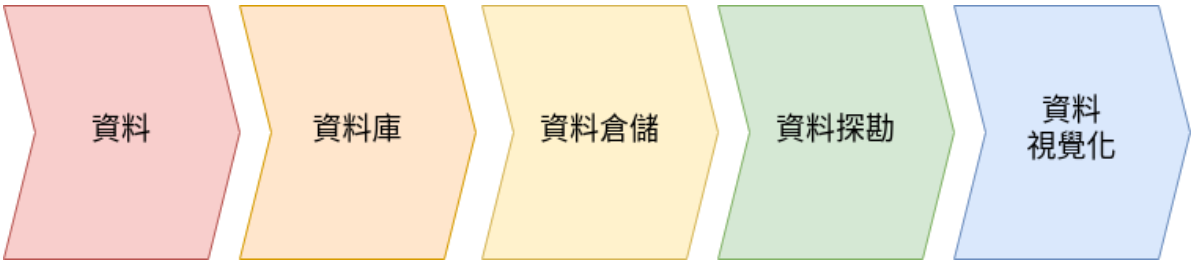
八二法則（Pareto Principle）

約僅有 20% 的變因操控著 80% 的局面，

也就是說：「所有變量中，最重要的僅有 20%，雖然 80% 佔多數，但控制的範圍卻遠低於關鍵的少數」。

資料處理鍊

> 注意：這個條目還在建立與理解中，在服用時可能會承受一定的薛丁格錯誤風險。



資料與變數

TABLE 1.1 Data Set for 60 Nations in the World Trade Organization

Nation	WTO Status	Per Capita GDP ($)	Fitch Rating	Fitch Outlook
Armenia	Member	3,615	BB−	Stable
Australia	Member	49,755	AAA	Stable
Austria	Member	44,758	AAA	Stable
Azerbaijan	Observer	3,879	BBB−	Stable
Bahrain	Member	22,579	BBB	Stable
Belgium	Member	41,271	AA	Stable
Brazil	Member	8,650	BBB	Stable
Bulgaria	Member	7,469	BBB−	Stable
Canada	Member	42,349	AAA	Stable
Cape Verde	Member	2,998	B+	Stable
Chile	Member	13,793	A+	Stable
China	Member	8,123	A+	Stable
Colombia	Member	5,806	BBB−	Stable
Costa Rica	Member	11,825	BB+	Stable
Croatia	Member	12,149	BBB−	Negative
Cyprus	Member	23,541	B	Negative
Czech Republic	Member	18,484	A+	Stable
Denmark	Member	53,579	AAA	Stable
Ecuador	Member	6,019	B−	Positive
Egypt	Member	3,478	B	Negative
El Salvador	Member	4,224	BB	Negative
Estonia	Member	17,737	A+	Stable
France	Member	36,857	AAA	Negative
Georgia	Member	3,866	BB−	Stable
Germany	Member	42,161	AAA	Stable
Hungary	Member	12,820	BB+	Stable
Iceland	Member	60,530	BBB	Stable
Ireland	Member	64,175	BBB+	Stable
Israel	Member	37,181	A	Stable
Italy	Member	30,669	A−	Negative
Japan	Member	38,972	A+	Negative
Kazakhstan	Observer	7,715	BBB+	Stable
Kenya	Member	1,455	B+	Stable
Latvia	Member	14,071	BBB	Positive
Lebanon	Observer	8,257	B	Stable
Lithuania	Member	14,913	BBB	Stable
Malaysia	Member	9,508	A−	Stable
Mexico	Member	8,209	BBB	Stable
Peru	Member	6,049	BBB	Stable
Philippines	Member	2,951	BB+	Stable
Poland	Member	12,414	A−	Positive
Portugal	Member	19,872	BB+	Negative
South Korea	Member	27,539	AA−	Stable
Romania	Member	9,523	BBB−	Stable
Russia	Member	8,748	BBB	Stable
Rwanda	Member	703	B	Stable
Serbia	Observer	5,426	BB−	Negative
Singapore	Member	52,962	AAA	Stable
Slovakia	Member	16,530	A+	Stable

名詞	意義
資料	經由蒐集、分析及彙總所得到，作為說明與解釋之用的事實與數值。
資料集	為特定研究目的蒐集的所有資料，由許多元素所組成。
元素	資料蒐集的實體，包含很多變數，例如上方表格的每個國家即為一個元素
變數	元素的某一特性，例如上列表格的每個元素有以下四個變數：WTO狀態、GDP、Fitch Rating、Fitch Outlook
觀察值	對特定元素蒐集的一組衡量值就是觀察值，例如上表的第1個觀察值(Armenia)包含了一組衡量值：Member、3615、BB-及Stable

衡量尺度

資料蒐集需要以下衡量尺度之一：名目尺度、順序尺度、區間尺度及比例尺度。

衡量尺度決定資料包含的資訊量，也指出資料彙整的或統計分析時的最適方法。

名目尺度(nominal scale)

用來表示元素屬性的標記或名稱，比較等於或不等於。

例如上表的國家WTO狀態可以分成「是WTO會員國」與「是WTO觀察員」，因此我們可以以數字1表示這個國家是WTO會員國，2表示這個國家是WTO觀察員，就能夠方便把資料輸入電腦，兩個國家的WTO狀態只能用相同與否來區分。

也因為名目尺度的意義是比較等於或不等於，因此詢問「WTO會員國與WTO觀察員哪個比較大」或者「兩個國家的WTO狀態相加等於多少」是完全毫無意義的行為。

順序尺度(ordinal scale)

與名目尺度不同，順序尺度的類別有一定的大小或順序，比起名目尺度只能比較相等，順序尺度能夠比較大小。

例如上表的Fitch Rating，其中AAA代表最好，F代表最差，因此可以根據評等排出高低，所以是順序尺度。

區間尺度(interval scale)

若變數具有順序資料的特性，且觀察值可以相加或相減，其結果仍有意義，這個變數的衡量尺度就是區間尺度，且一定以數值表示。

例如統測成績就是一個區間尺度，假設有三位學生的統測成績為699、560、350，則我們可以由高到低依序排序來衡量出成績表現的優劣，而他們的差距也存在意義，例如699的學生比560的學生高出139分。

比例尺度(ratio scale)

若變數具有順序資料的特性，且觀察值可以加減乘除，其結果仍有意義，這個變數的衡量尺度就是比例尺度，且一定以數值表示。

與區間尺度的差別在於，**比例尺度要求絕對零點，也就是值必須要大於等於0且在0上必須要是自然的不存在**。

例如年齡不存在0歲，而高度不存在0公分，而可以描述20歲比5歲大4倍。

百分位數

百分位數可以瞭解資料在最小值與最大值間的分布狀況，以pth百分位數可以把資料分成兩個部份。

大約pth百分比的觀察值會小於pth百分位數，而大約有(100-p)百分比的觀察值會大於pth百分位數。

計算時需要先「非嚴格遞增排序」，計算公式如下： $L_p = \frac{p}{100}(n+1)$

四分位數

四分位數將整筆資料分成四個部份，每個部份大概含有25%的資料個數，定義如下：

1. Q_1為第一四分位數或 25th percentile
2. Q_2為第二四分位數或 50th percentile，也就是中位數
3. Q_3為第三四分位數或 75th percentile

四分位距(IQR)

四分位距較能克服極端值的影響，定義如下 $\text{IQR} = Q_3 - Q_1$ 也就是中間資料50的全距。

離群偵測

有時在一個資料集中會有極大與極小的數值，這些數值稱為離群。

我們可以利用z分數去偵測離群值，一般來說，約有99.7的資料會落在標準差± 3內，我們會希望資料的標準差與中心的距離不超過3。

另一種方式是使用第一分位、第三分位與四分位距來做偵測，能夠給定一個區間來要求值必須要在這個區間內，定義如下：

$\text{Lower Limit} = Q_1 - 1.5(\text{IQR})$ $\text{Upper Limit} = Q_3 + 1.5(\text{IQR})$

資料模型

資料模型是為了能有組織有效率地，將我們需求的資料儲存於資料庫系統中，以及有一個適當的表達方式。

主要分成三種模式的**演變**：

1. 階層式資料模型
2. 網路式資料模型
3. 關聯式資料模型

階層式資料模型

主要使用樹狀結構，將一筆一筆的紀錄組織起來，適合一對多的資料組成關係，但無法直接表達多對多的關係。

網路式資料模型

改善了階層式資料模型，可以供給多對多的關係。

在資料呈現上非常複雜、且無法確實表達資料與資料之間的網路連結關係。

關聯式資料模型

以表格來表達關係，每一列即為一個紀錄，通常會實體關聯圖作為輔助設計的依據。

SQL 查詢

SQL 全名為 結構化查詢語言（Structured Query Language），對於資料庫來說，我們可以用 SQL 語法來進行查詢。

例如要從一張名為 Teacher 的資料表中，顯示出所有結果屬性，寫法可以寫成這樣。

```
SELECT * FROM Teacher
```

若要從這堆老師中選出其中一個名為王小軒的老師，可以寫成這樣。

```
SELECT * FROM Teacher WHERE name = '王小軒'
```

若只要得到王小軒老師的 ID，可以寫成這樣。

```
SELECT ID FROM Teacher WHERE name = '王小軒'
```

非結構化資料

非結構化資料的介紹

非結構化資料形式涵蓋了聲音、圖像、影像、文字等等。

> 結構化資料 vs. 非結構化資料 | Pure Storage

非結構化資料的特性

數位化生成

大部分由機器生成的資料，並不一定能夠整齊的被資料庫的欄位所對應，因此會使得資料庫變成多維度，且非常難以預測。

多模式

資料經過蒐集之後，具有大量不同類別的資料，例如 e-mail、文檔、圖檔等等。

持續變動

大量資料被生成、處理、分析與即時運算。

地理分散

不同資料被儲存在不同的地方，來達成資安特性。

> 結構化資料 vs. 非結構化資料 | Pure Storage
>
> Unified Fast File and Object: A New Category of Storage | Pure Storage
>
> Types and Examples of NoSQL Databases - Big Data Analytics News
>
> 淺談資料格式 — 結構化與非結構化資料. 進入大數據時代，資料成為挖掘商機的礦脈，對資料的管理不夠，想要利用大數據來開創新… | by 行銷資料科學 | Marketingdatascience | Medium

結構化與非結構資料化的比較

	結構化資料	非結構化資料
呈現方面	表格	無法呈現
處理方面	需正規化	不須正規化
形式	有限的資料形式	不限
儲存空間需求	較少	較大
存取	較簡單	較困難

資料探勘

從一大群的資料中，利用技術（人工智慧、機器學習、統計學等等）探勘出有意義的資料。

資料的探勘任務

主要分成六種常見的任務。

1. 異常檢測：辨識不尋常的資料，或針對錯誤資料進一步調查。
2. 關聯規則學習：搜尋變數之間的關係。
3. 聚類：在未知資料的結構下，發現資料的類別與結構，利用演算法將資料分成更多子集，讓子集的資料都有相似的一些屬性。
4. 分類：對新的資料推廣成已知結構的任務，例如將一封新郵件分類成「正常郵件」與「垃圾郵件」，可利用決策樹來分析數據或輔助預測。
5. 迴歸：試圖找到最小誤差的建模函式。
6. 匯總：提供一個更緊湊的資料集表示，來生成視覺化或報表。