從零開始的關聯式學習

Pandas 與 MIxtend

張家瑋 博士

副教授

國立臺中科技大學資訊工程系

Work Experience

- 2022/2~ Now
  Associate Professor
  National Taichung University of Science and Technology
- 2018/2 ~ Now
  Adjunct Assistant Professor
  National Cheng Kung University
- 2015/8 ~ 2017/11
  Project Manager & Data Scientist
  NEXCOM International Co., Ltd.

About Me

- [Since Jan. 2019] Young Professionals Chair, IET Taipei Local Network.
- [Since Dec. 2017] Consultant, NEXCOM Industry 4.0 Center.
- [Jan. 2017] Ph.D. degree, National Cheng Kung University.

Research Topics

a) Natural Language Processing
  - ✓ Natural Language Understanding
  - ✓ Chatbot
  - ✓ Text Summarization / Classification

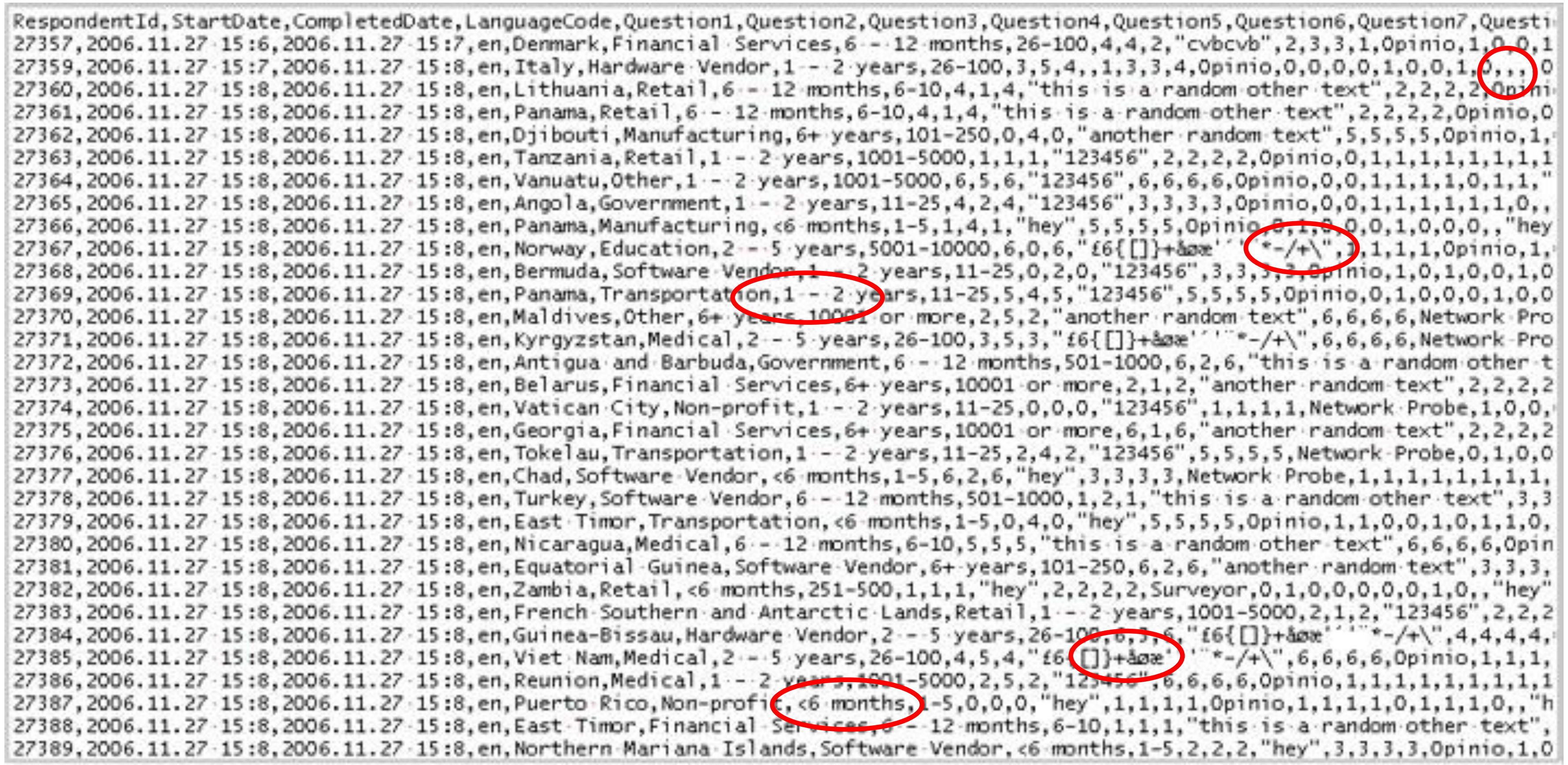
b) Deep Learning

c) Data Mining

d) Internet of Things
  - ✓ Smart Speaker

Basic Background of Data Science

Data?



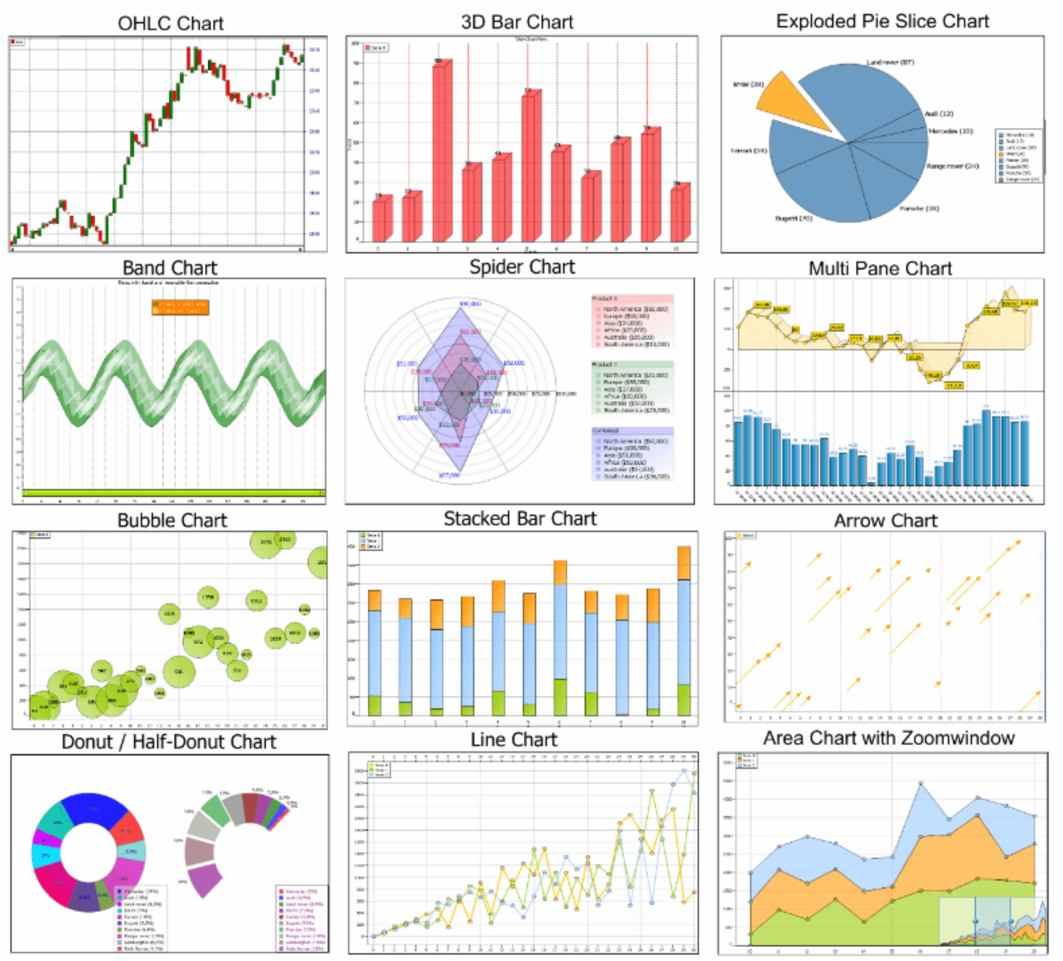
```
RespondentId,StartDate,CompletedDate,LanguageCode,Question1,Question2,Question3,Question4,Question5,Question6,Question7,Questi
27357,2006.11.27 15:6,2006.11.27 15:7,en,Denmark,Financial Services,6 - 12 months,26-100,4,4,2,"cvbcvb",2,3,3,1,Opinio,1,0,0,1
27359,2006.11.27 15:7,2006.11.27 15:8,en,Italy,Hardware Vendor,1 - 2 years,26-100,3,5,4,,1,3,3,4,Opinio,0,0,0,0,1,0,0,1,0,,,0
27360,2006.11.27 15:8,2006.11.27 15:8,en,Lithuania,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2,Opini
27361,2006.11.27 15:8,2006.11.27 15:8,en,Panama,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2,Opinio,0
27362,2006.11.27 15:8,2006.11.27 15:8,en,Djibouti,Manufacturing,6+ years,101-250,0,4,0,"another random text",5,5,5,5,Opinio,1,
27363,2006.11.27 15:8,2006.11.27 15:8,en,Tanzania,Retail,1 - 2 years,1001-5000,1,1,1,"123456",2,2,2,2,Opinio,0,1,1,1,1,1,1,1,1
27364,2006.11.27 15:8,2006.11.27 15:8,en,Vanuatu,Other,1 - 2 years,1001-5000,6,5,6,"123456",6,6,6,6,Opinio,0,0,1,1,1,1,0,1,1,"
27365,2006.11.27 15:8,2006.11.27 15:8,en,Angola,Government,1 - 2 years,11-25,4,2,4,"123456",3,3,3,3,Opinio,0,0,1,1,1,1,1,1,0,,
27366,2006.11.27 15:8,2006.11.27 15:8,en,Panama,Manufacturing,<6 months,1-5,1,4,1,"hey",5,5,5,5,Opinio,0,1,0,0,0,1,0,0,0,,"hey
27367,2006.11.27 15:8,2006.11.27 15:8,en,Norway,Education,2 - 5 years,5001-10000,6,0,6,"£6{[]}+åøæ'´'¨*-/+\",1,1,1,1,Opinio,1,
27368,2006.11.27 15:8,2006.11.27 15:8,en,Bermuda,Software Vendor,1 - 2 years,11-25,0,2,0,"123456",3,3,3,3,Opinio,1,0,1,0,0,1,0
27369,2006.11.27 15:8,2006.11.27 15:8,en,Panama,Transportation,1 - 2 years,11-25,5,4,5,"123456",5,5,5,5,Opinio,0,1,0,0,0,1,0,0
27370,2006.11.27 15:8,2006.11.27 15:8,en,Maldives,Other,6+ years,10001 or more,2,5,2,"another random text",6,6,6,6,Network Pro
27371,2006.11.27 15:8,2006.11.27 15:8,en,Kyrgyzstan,Medical,2 - 5 years,26-100,3,5,3,"£6{[]}+åøæ'´'¨*-/+\",6,6,6,6,Network Pro
27372,2006.11.27 15:8,2006.11.27 15:8,en,Antigua and Barbuda,Government,6 - 12 months,501-1000,6,2,6,"this is a random other t
27373,2006.11.27 15:8,2006.11.27 15:8,en,Belarus,Financial Services,6+ years,10001 or more,2,1,2,"another random text",2,2,2,2
27374,2006.11.27 15:8,2006.11.27 15:8,en,Vatican City,Non-profit,1 - 2 years,11-25,0,0,0,"123456",1,1,1,1,Network Probe,1,0,0,
27375,2006.11.27 15:8,2006.11.27 15:8,en,Georgia,Financial Services,6+ years,10001 or more,6,1,6,"another random text",2,2,2,2
27376,2006.11.27 15:8,2006.11.27 15:8,en,Tokelau,Transportation,1 - 2 years,11-25,2,4,2,"123456",5,5,5,5,Network Probe,0,1,0,0
27377,2006.11.27 15:8,2006.11.27 15:8,en,Chad,Software Vendor,<6 months,1-5,6,2,6,"hey",3,3,3,3,Network Probe,1,1,1,1,1,1,1,1,
27378,2006.11.27 15:8,2006.11.27 15:8,en,Turkey,Software Vendor,6 - 12 months,501-1000,1,2,1,"this is a random other text",3,3
27379,2006.11.27 15:8,2006.11.27 15:8,en,East Timor,Transportation,<6 months,1-5,0,4,0,"hey",5,5,5,5,Opinio,1,1,0,0,1,0,1,1,0,
27380,2006.11.27 15:8,2006.11.27 15:8,en,Nicaragua,Medical,6 - 12 months,6-10,5,5,5,"this is a random other text",6,6,6,6,Opin
27381,2006.11.27 15:8,2006.11.27 15:8,en,Equatorial Guinea,Software Vendor,6+ years,101-250,6,2,6,"another random text",3,3,3,
27382,2006.11.27 15:8,2006.11.27 15:8,en,Zambia,Retail,<6 months,251-500,1,1,1,"hey",2,2,2,2,Surveyor,0,1,0,0,0,0,0,1,0,,"hey"
27383,2006.11.27 15:8,2006.11.27 15:8,en,French Southern and Antarctic Lands,Retail,1 - 2 years,1001-5000,2,1,2,"123456",2,2,2
27384,2006.11.27 15:8,2006.11.27 15:8,en,Guinea-Bissau,Hardware Vendor,2 - 5 years,26-100,6,3,6,"£6{[]}+åøæ'´'¨*-/+\",4,4,4,4,
27385,2006.11.27 15:8,2006.11.27 15:8,en,Viet Nam,Medical,2 - 5 years,26-100,4,5,4,"£6{[]}+åøæ'´'¨*-/+\",6,6,6,6,Opinio,1,1,1,
27386,2006.11.27 15:8,2006.11.27 15:8,en,Reunion,Medical,1 - 2 years,1001-5000,2,5,2,"123456",6,6,6,6,Opinio,1,1,1,1,1,1,1,1,1
27387,2006.11.27 15:8,2006.11.27 15:8,en,Puerto Rico,Non-profit,<6 months,1-5,0,0,0,"hey",1,1,1,1,Opinio,1,1,1,1,0,1,1,1,0,,"h
27388,2006.11.27 15:8,2006.11.27 15:8,en,East Timor,Financial Services,6 - 12 months,6-10,1,1,1,"this is a random other text",
27389,2006.11.27 15:8,2006.11.27 15:8,en,Northern Mariana Islands,Software Vendor,<6 months,1-5,2,2,2,"hey",3,3,3,3,Opinio,1,0
```

- 未經過處理的原始記錄。
- 資料**缺乏組織及分類**，無法明確的表達事物代表的意義。
- 資料是關於事件的一組**離散且客觀**的事實描述，是構成資訊和知識的原始材料。

Information?

- 資訊是經過處理後的資料。
- 資訊是有用的或有意義的資料。
- 對接受者**有意義的資料能使接受者產生資訊**。

Information?



Knowledge?

- 知識是資訊、文化脈絡以及經驗的**整合**。
- 知識是對**某個主題**確信的認識，並且這些認識擁有潛在的能力為**特定目的**而使用。
- 藉由**專業技能**或**豐富經驗**用以分析資訊的結果。

Intelligence?

- 以知識為基礎，**運用**個人能力，**實踐**能力來開創價值。
- 分析、判斷、創造、思考的能力。
- 智慧具有反應能力與價值判斷。

人工智慧？

- 指由人製造出來的機器所表現出來的智慧。(Wiki)
- 弱人工智慧 → 專家系統
  - 處理特定的問題
- 強人工智慧
  - 通用人工智慧

過去

- Small Data
  - 針對某一個問題，只能獲得**小量數據**。
  - 數百筆到數萬筆。
  - 花費**大量人工編碼**。

過去

- Small Data 統計分析
  - 樣本推論母體**(抽樣)**
  - 在小樣本中，需要發展一系列理論來解釋事物的原理**(學說)**
  - [啟示] 1936 羅斯福與藍頓 的民調

過去

- 小數據
- Rule-based AI
- 類神經網路 (1980s)

現在與未來

- Big Data
  - 由"**母體**"來分析數據
  - 數萬筆到幾近無限
  - **雜亂**的原始資料

上一世代

- 大數據
- 分類：SVM → 機器學習
- 分群：Kmeans
- 關聯式法則：Apriori

What's difference?

- Small Data vs Big Data
  - 都有目的或待解的問題

But

- 減少假設
- 力求呈現真實世界

What's difference?

- 資料可重組與檢視關聯。
- 接受「**數據的雜亂性**」，不再追求「精確」的數據。
- **重「相關」**而輕「因果」。

現在

- 大數據
- 運算力的提升
- 深度學習 (強AI的可能性)
  - 類神經網路的文藝復興
- 演進趨勢
  - 腦神經科學
  - 認知科學
  - 認知心理學

Big Data 的沿革 (1/3)

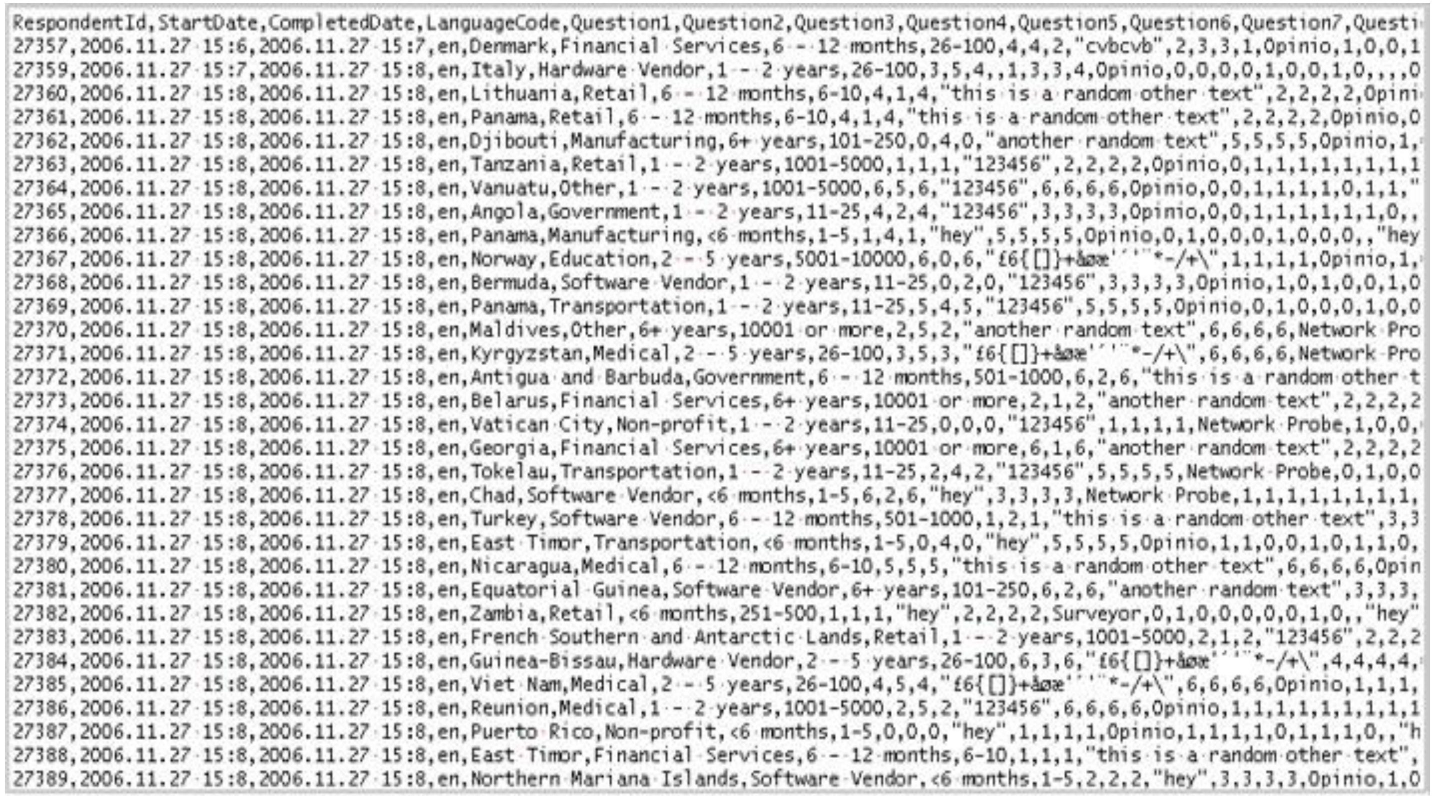
- Data Mining
  - 資料探勘是利用分析技術來發掘資料間**未知的關聯性與規則**。
  - 少女未婚懷孕　購物商場比老爸還早知道？！
    - https://www.nownews.com/news/20120223/42676

Data Mining

- ✓ 分群
  - 用於沒有標籤的資料，又通常為非監督式演算法。
- ✓ 分類
  - 用於有標籤的資料，又通常為監督式演算法。
- ✓ 關聯式法則
  - 有序性規則的資料

Data Mining

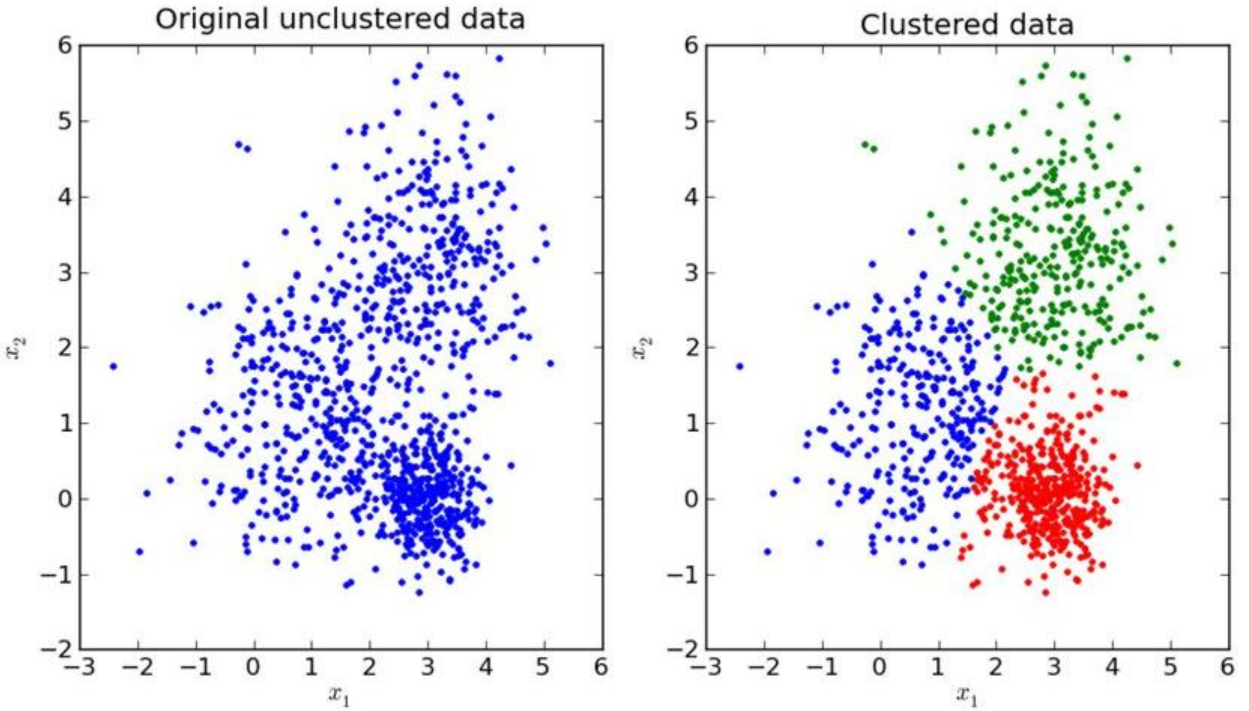
- ✓ 分群
  - – 用於沒有標籤的資料，又通常為非監督式演算法。



```
RespondentId,StartDate,CompletedDate,LanguageCode,Question1,Question2,Question3,Question4,Question5,Question6,Question7,Questi
27357,2006.11.27 15:6,2006.11.27 15:7,en,Denmark,Financial Services,6 - 12 months,26-100,4,4,2,"cvbcvb",2,3,3,1,Opinio,1,0,0,1
27359,2006.11.27 15:7,2006.11.27 15:8,en,Italy,Hardware Vendor,1 - 2 years,26-100,3,5,4,,1,3,3,4,Opinio,0,0,0,0,1,0,0,1,0,,,,0
27360,2006.11.27 15:8,2006.11.27 15:8,en,Lithuania,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2,Opini
27361,2006.11.27 15:8,2006.11.27 15:8,en,Panama,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2,Opinio,0
27362,2006.11.27 15:8,2006.11.27 15:8,en,Djibouti,Manufacturing,6+ years,101-250,0,4,0,"another random text",5,5,5,5,Opinio,1,
27363,2006.11.27 15:8,2006.11.27 15:8,en,Tanzania,Retail,1 - 2 years,1001-5000,1,1,1,"123456",2,2,2,2,Opinio,0,1,1,1,1,1,1,1,1
27364,2006.11.27 15:8,2006.11.27 15:8,en,Vanuatu,Other,1 - 2 years,1001-5000,6,5,6,"123456",6,6,6,6,Opinio,0,0,1,1,1,1,0,1,1,"
27365,2006.11.27 15:8,2006.11.27 15:8,en,Angola,Government,1 - 2 years,11-25,4,2,4,"123456",3,3,3,3,Opinio,0,0,1,1,1,1,1,1,0,,
27366,2006.11.27 15:8,2006.11.27 15:8,en,Panama,Manufacturing,<6 months,1-5,1,4,1,"hey",5,5,5,5,Opinio,0,1,0,0,0,1,0,0,0,,"hey
27367,2006.11.27 15:8,2006.11.27 15:8,en,Norway,Education,2 - 5 years,5001-10000,6,0,6,"£6{[]}+åøæ'´'¨*-/+\",1,1,1,1,Opinio,1,
27368,2006.11.27 15:8,2006.11.27 15:8,en,Bermuda,Software Vendor,1 - 2 years,11-25,0,2,0,"123456",3,3,3,3,Opinio,1,0,1,0,0,1,0
27369,2006.11.27 15:8,2006.11.27 15:8,en,Panama,Transportation,1 - 2 years,11-25,5,4,5,"123456",5,5,5,5,Opinio,0,1,0,0,0,1,0,0
27370,2006.11.27 15:8,2006.11.27 15:8,en,Maldives,Other,6+ years,10001 or more,2,5,2,"another random text",6,6,6,6,Network Pro
27371,2006.11.27 15:8,2006.11.27 15:8,en,Kyrgyzstan,Medical,2 - 5 years,26-100,3,5,3,"£6{[]}+åøæ'´'¨*-/+\",6,6,6,6,Network Pro
27372,2006.11.27 15:8,2006.11.27 15:8,en,Antigua and Barbuda,Government,6 - 12 months,501-1000,6,2,6,"this is a random other t
27373,2006.11.27 15:8,2006.11.27 15:8,en,Belarus,Financial Services,6+ years,10001 or more,2,1,2,"another random text",2,2,2,2
27374,2006.11.27 15:8,2006.11.27 15:8,en,Vatican City,Non-profit,1 - 2 years,11-25,0,0,0,"123456",1,1,1,1,Network Probe,1,0,0,
27375,2006.11.27 15:8,2006.11.27 15:8,en,Georgia,Financial Services,6+ years,10001 or more,6,1,6,"another random text",2,2,2,2
27376,2006.11.27 15:8,2006.11.27 15:8,en,Tokelau,Transportation,1 - 2 years,11-25,2,4,2,"123456",5,5,5,5,Network Probe,0,1,0,0
27377,2006.11.27 15:8,2006.11.27 15:8,en,Chad,Software Vendor,<6 months,1-5,6,2,6,"hey",3,3,3,3,Network Probe,1,1,1,1,1,1,1,1,
27378,2006.11.27 15:8,2006.11.27 15:8,en,Turkey,Software Vendor,6 - 12 months,501-1000,1,2,1,"this is a random other text",3,3
27379,2006.11.27 15:8,2006.11.27 15:8,en,East Timor,Transportation,<6 months,1-5,0,4,0,"hey",5,5,5,5,Opinio,1,1,0,0,1,0,1,1,0,
27380,2006.11.27 15:8,2006.11.27 15:8,en,Nicaragua,Medical,6 - 12 months,6-10,5,5,5,"this is a random other text",6,6,6,6,Opin
27381,2006.11.27 15:8,2006.11.27 15:8,en,Equatorial Guinea,Software Vendor,6+ years,101-250,6,2,6,"another random text",3,3,3,
27382,2006.11.27 15:8,2006.11.27 15:8,en,Zambia,Retail,<6 months,251-500,1,1,1,"hey",2,2,2,2,Surveyor,0,1,0,0,0,0,0,1,0,,"hey"
27383,2006.11.27 15:8,2006.11.27 15:8,en,French Southern and Antarctic Lands,Retail,1 - 2 years,1001-5000,2,1,2,"123456",2,2,2
27384,2006.11.27 15:8,2006.11.27 15:8,en,Guinea-Bissau,Hardware Vendor,2 - 5 years,26-100,6,3,6,"£6{[]}+åøæ'´'¨*-/+\",4,4,4,4,
27385,2006.11.27 15:8,2006.11.27 15:8,en,Viet Nam,Medical,2 - 5 years,26-100,4,5,4,"£6{[]}+åøæ'´'¨*-/+\",6,6,6,6,Opinio,1,1,1,
27386,2006.11.27 15:8,2006.11.27 15:8,en,Reunion,Medical,1 - 2 years,1001-5000,2,5,2,"123456",6,6,6,6,Opinio,1,1,1,1,1,1,1,1,1
27387,2006.11.27 15:8,2006.11.27 15:8,en,Puerto Rico,Non-profit,<6 months,1-5,0,0,0,"hey",1,1,1,1,Opinio,1,1,1,1,0,1,1,1,0,,"h
27388,2006.11.27 15:8,2006.11.27 15:8,en,East Timor,Financial Services,6 - 12 months,6-10,1,1,1,"this is a random other text",
27389,2006.11.27 15:8,2006.11.27 15:8,en,Northern Mariana Islands,Software Vendor,<6 months,1-5,2,2,2,"hey",3,3,3,3,Opinio,1,0
```

Data Mining

- ✓ 分群
  - – 用於沒有標籤的資料，又通常為非監督式演算法。

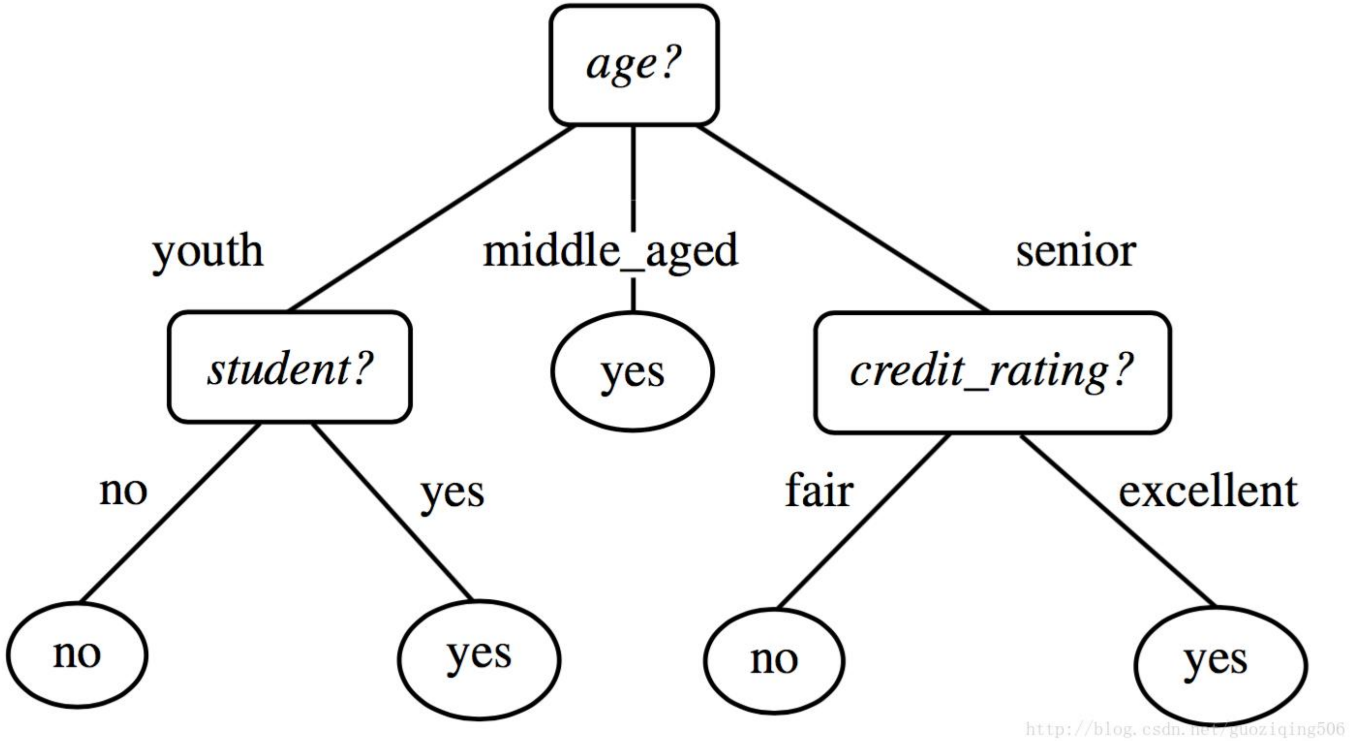


Original unclustered data

Clustered data

Data Mining

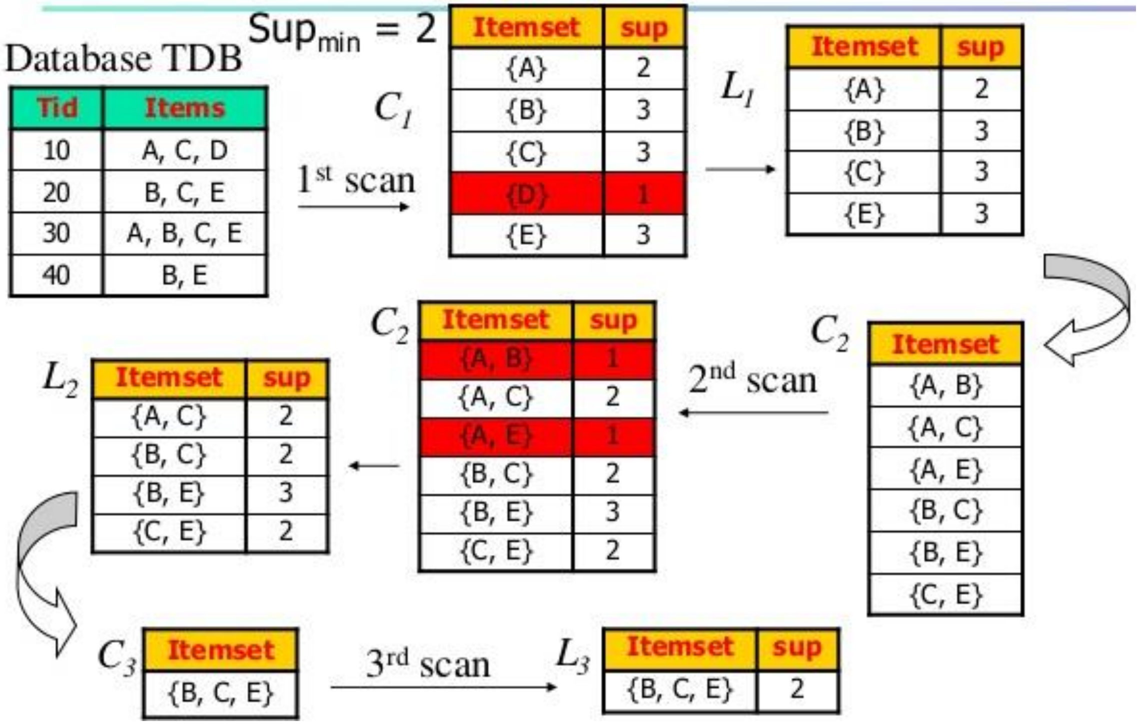
- ✓ 分類
  - – 用於有標籤的資料，又通常為監督式演算法。



Data Mining

✓ 關聯式法則

  – 有序性 (尿布與啤酒)



$\text{Sup}_{\text{min}} = 2$

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan → C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

→ L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan → C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

→ L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan → L_3

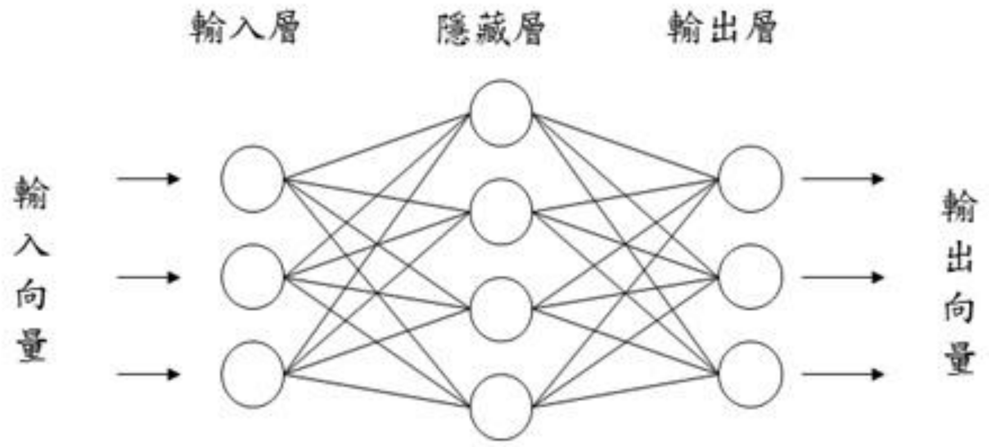
Itemset	sup
{B, C, E}	2

Big Data 的沿革 (2/3)

- Machine Learning
  - 人工智慧的分支，可用於資料探勘。
  - 讓機器可以**自動學習**、從巨量資料中找到規則，進而有能力做出**分類或預測**。
    - 判斷出類別
    - 估計出數值

Big Data 的沿革 (3/3)

- Deep Learning
  - 是機器學習的分支
  - **類神經網路**的文藝復興
  - 從**大規模未標記資料**中建立更好的預測模型
  - 建立強 AI 的可能性



資料分析的基本步驟

1. 資料清除：去除極端、遺失值資料、不重要的屬性
2. 資料整合：因應用目的或特性，整合不同來源的資料
3. 資料選擇：揀選重要的屬性來逼近目的之最佳成效
4. 資料轉換：基於領域知識進行特徵縮放、數值類別轉換等
5. 資料探勘：選用合適的分析演算法得到目的之結果
6. 樣式評估：評估結果的樣式，是否如預期
7. 知識表示：因應目的將樣式轉換成合適的表達方法

資料分析的演算法重點

- 預處理（Preprocessing）
- 降維（Dimensionality Reduction）
- 模型選擇（Model Selection）
  - 監督式學習（Supervised learning）
    - 分類（Classification）：機器給出一個類別
    - 迴歸（Regression）：機器給出一個數值
  - 非監督式學習（Unsupervised learning）
    - 分群（Clustering）
  - 關聯式學習（Association Rule Learning）

資料分析的常見角色

- 資料產品經理人：將真實世界的問題轉換成資料可以解決的問題，通常是該問題領域的專業人士
- 資料工程師：蒐集、整理、清理資料，通常是具備程式技術能力的工程師
- 資料分析師：負責資料建模和分析，通常由擅長找出資料關聯的統計人擔當
- 資料視覺化設計師：將報表變得簡明易懂

資料驅動創新應用

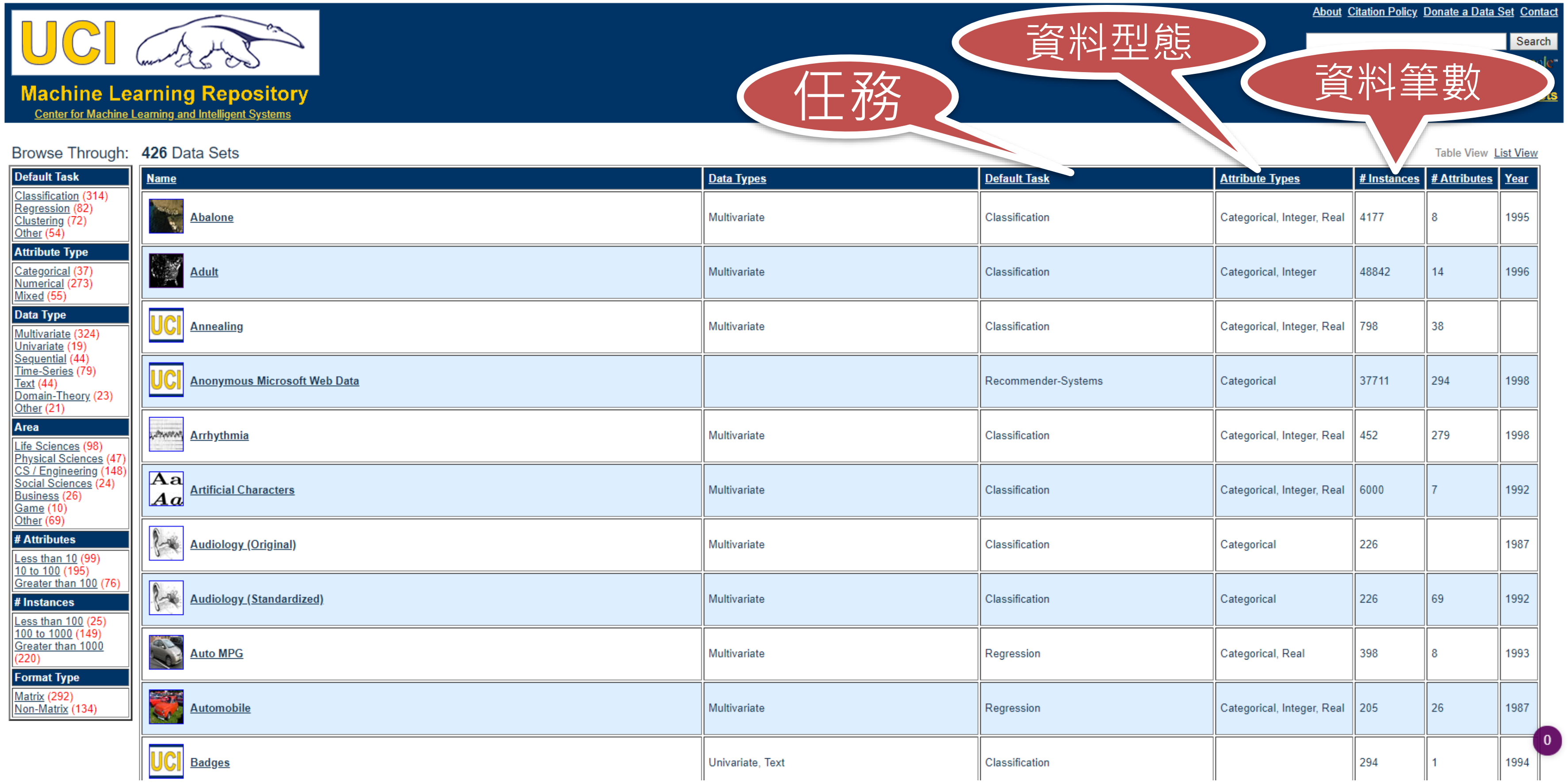
- 文字、聲音、影像
  - 自然語言處理
  - 語音辨識
  - 影像辨識

- 數值與非數值
  - 連續性
  - 離散性、類別

AI Open Datasets

UC Irvine Machine Learning Repository

http://archive.ics.uci.edu/ml/datasets.html



任務

資料型態

資料筆數

UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
About Citation Policy Donate a Data Set Contact
Search
Table View List View

Browse Through: **426** Data Sets

Default Task
- Classification (314)
- Regression (82)
- Clustering (72)
- Other (54)

Attribute Type
- Categorical (37)
- Numerical (273)
- Mixed (55)

Data Type
- Multivariate (324)
- Univariate (19)
- Sequential (44)
- Time-Series (79)
- Text (44)
- Domain-Theory (23)
- Other (21)

Area
- Life Sciences (98)
- Physical Sciences (47)
- CS / Engineering (148)
- Social Sciences (24)
- Business (26)
- Game (10)
- Other (69)

# Attributes
- Less than 10 (99)
- 10 to 100 (195)
- Greater than 100 (76)

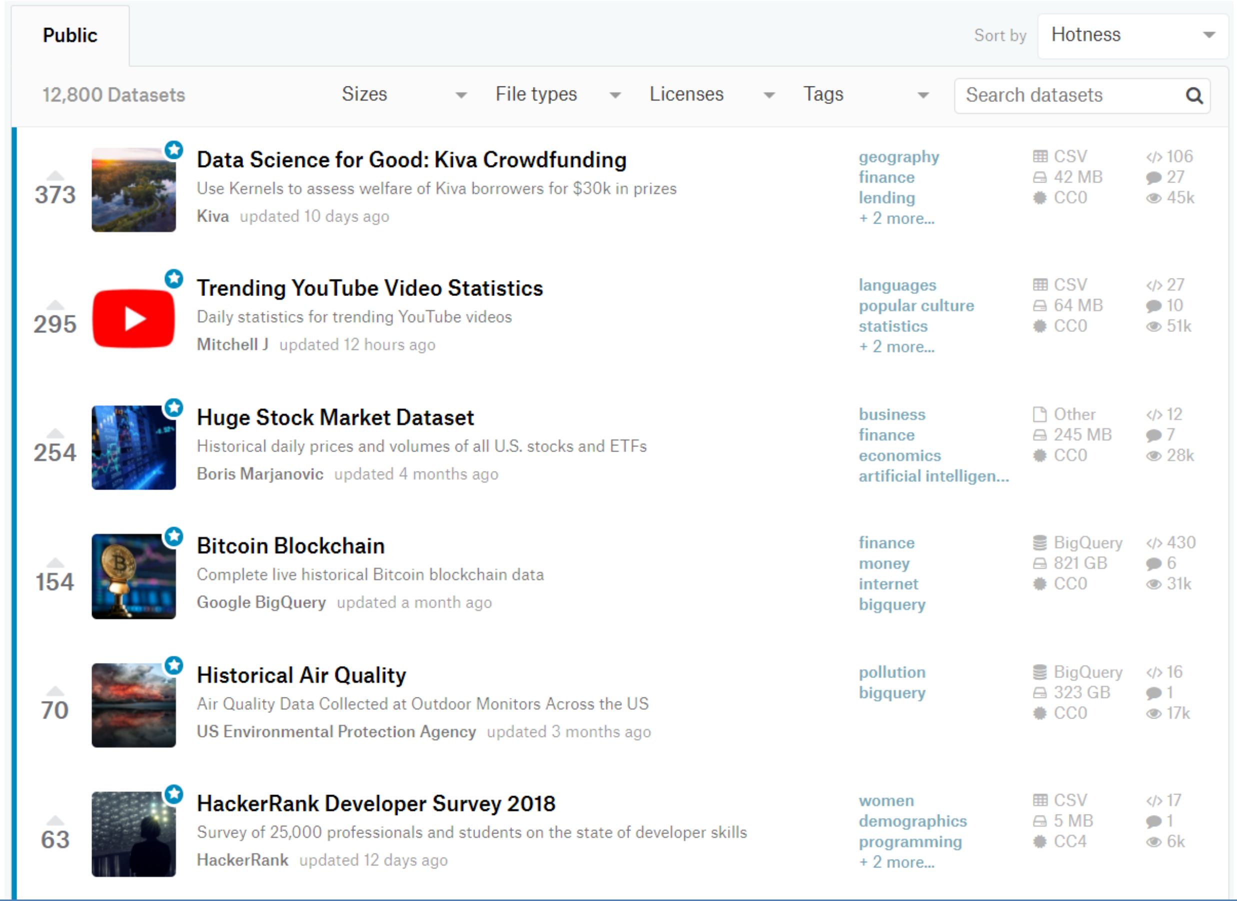
# Instances
- Less than 100 (25)
- 100 to 1000 (149)
- Greater than 1000 (220)

Format Type
- Matrix (292)
- Non-Matrix (134)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Audiology (Original)	Multivariate	Classification	Categorical	226		1987
Audiology (Standardized)	Multivariate	Classification	Categorical	226	69	1992
Auto MPG	Multivariate	Regression	Categorical, Real	398	8	1993
Automobile	Multivariate	Regression	Categorical, Integer, Real	205	26	1987
Badges	Univariate, Text	Classification		294	1	1994

Kaggle DataSets

https://www.kaggle.com/datasets



Public

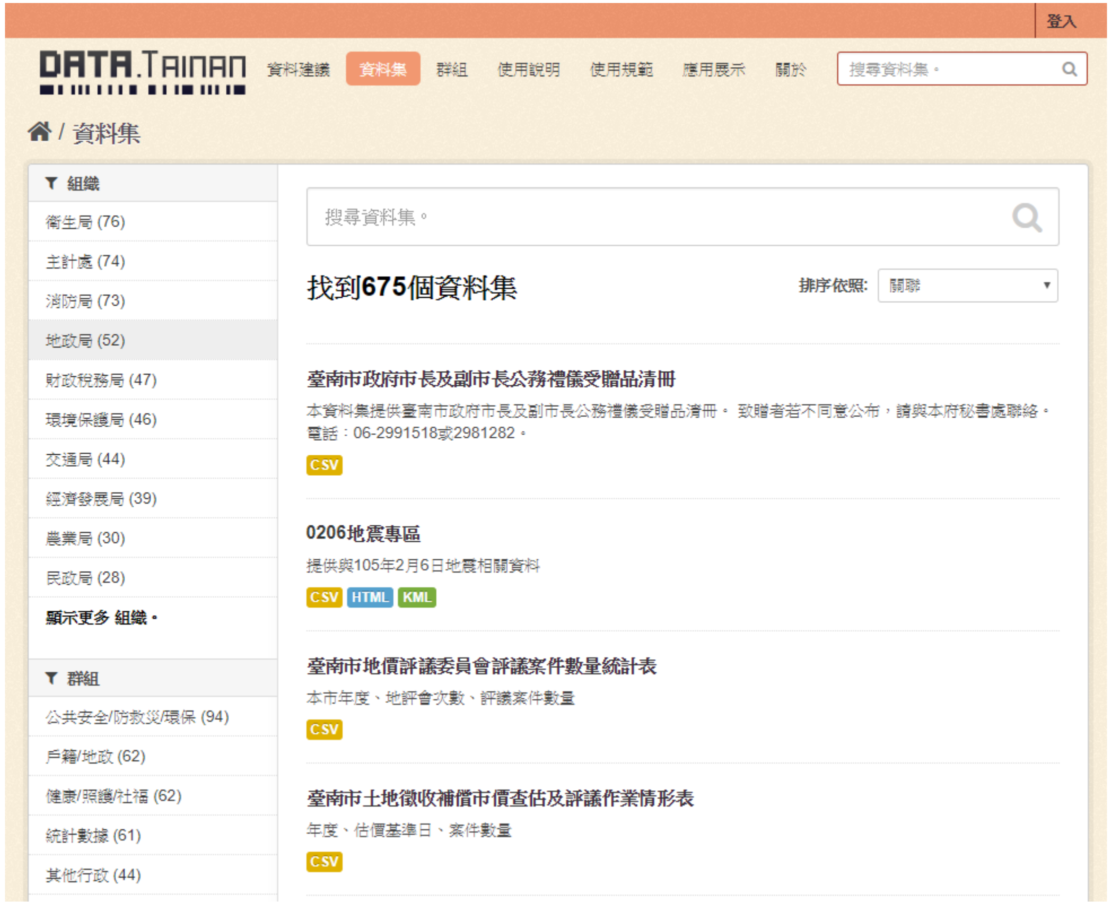
Sort by Hotness

12,800 Datasets

373	**Data Science for Good: Kiva Crowdfunding** Use Kernels to assess welfare of Kiva borrowers for $30k in prizes Kiva updated 10 days ago	geography finance lending + 2 more...	CSV 42 MB CC0	106 27 45k
295	**Trending YouTube Video Statistics** Daily statistics for trending YouTube videos Mitchell J updated 12 hours ago	languages popular culture statistics + 2 more...	CSV 64 MB CC0	27 10 51k
254	**Huge Stock Market Dataset** Historical daily prices and volumes of all U.S. stocks and ETFs Boris Marjanovic updated 4 months ago	business finance economics artificial intelligen...	Other 245 MB CC0	12 7 28k
154	**Bitcoin Blockchain** Complete live historical Bitcoin blockchain data Google BigQuery updated a month ago	finance money internet bigquery	BigQuery 821 GB CC0	430 6 31k
70	**Historical Air Quality** Air Quality Data Collected at Outdoor Monitors Across the US US Environmental Protection Agency updated 3 months ago	pollution bigquery	BigQuery 323 GB CC0	16 1 17k
63	**HackerRank Developer Survey 2018** Survey of 25,000 professionals and students on the state of developer skills HackerRank updated 12 days ago	women demographics programming + 2 more...	CSV 5 MB CC4	17 1 6k

臺南市開放資料

http://data.tainan.gov.tw/dataset



/ 資料集

組織
- 衛生局 (76)
- 主計處 (74)
- 消防局 (73)
- 地政局 (52)
- 財政稅務局 (47)
- 環境保護局 (46)
- 交通局 (44)
- 經濟發展局 (39)
- 農業局 (30)
- 民政局 (28)
- **顯示更多 組織。**

群組
- 公共安全/防救災/環保 (94)
- 戶籍/地政 (62)
- 健康/照護/社福 (62)
- 統計數據 (61)
- 其他行政 (44)

搜尋資料集。

找到**675**個資料集

排序依照: 關聯

臺南市政府市長及副市長公務禮儀受贈品清冊

本資料集提供臺南市政府市長及副市長公務禮儀受贈品清冊。 致贈者若不同意公布，請與本府秘書處聯絡。電話：06-2991518或2981282。

CSV

0206地震專區

提供與105年2月6日地震相關資料

CSV HTML KML

臺南市地價評議委員會評議案件數量統計表

本市年度、地評會次數、評議案件數量

CSV

臺南市土地徵收補償市價查估及評議作業情形表

年度、估價基準日、案件數量

CSV

關聯規則學習

Association Rule Learning

概念

- 在大型資料庫中發現項目間關聯的方法。
  - {牛奶, 麵包}→{可樂}：代表某人同時買了牛奶和麵包，就可能會買可樂。
- 該方法常使用於電子商務上，通常可為**促銷**、**產品推薦**等行銷活動的決策依據。

定義

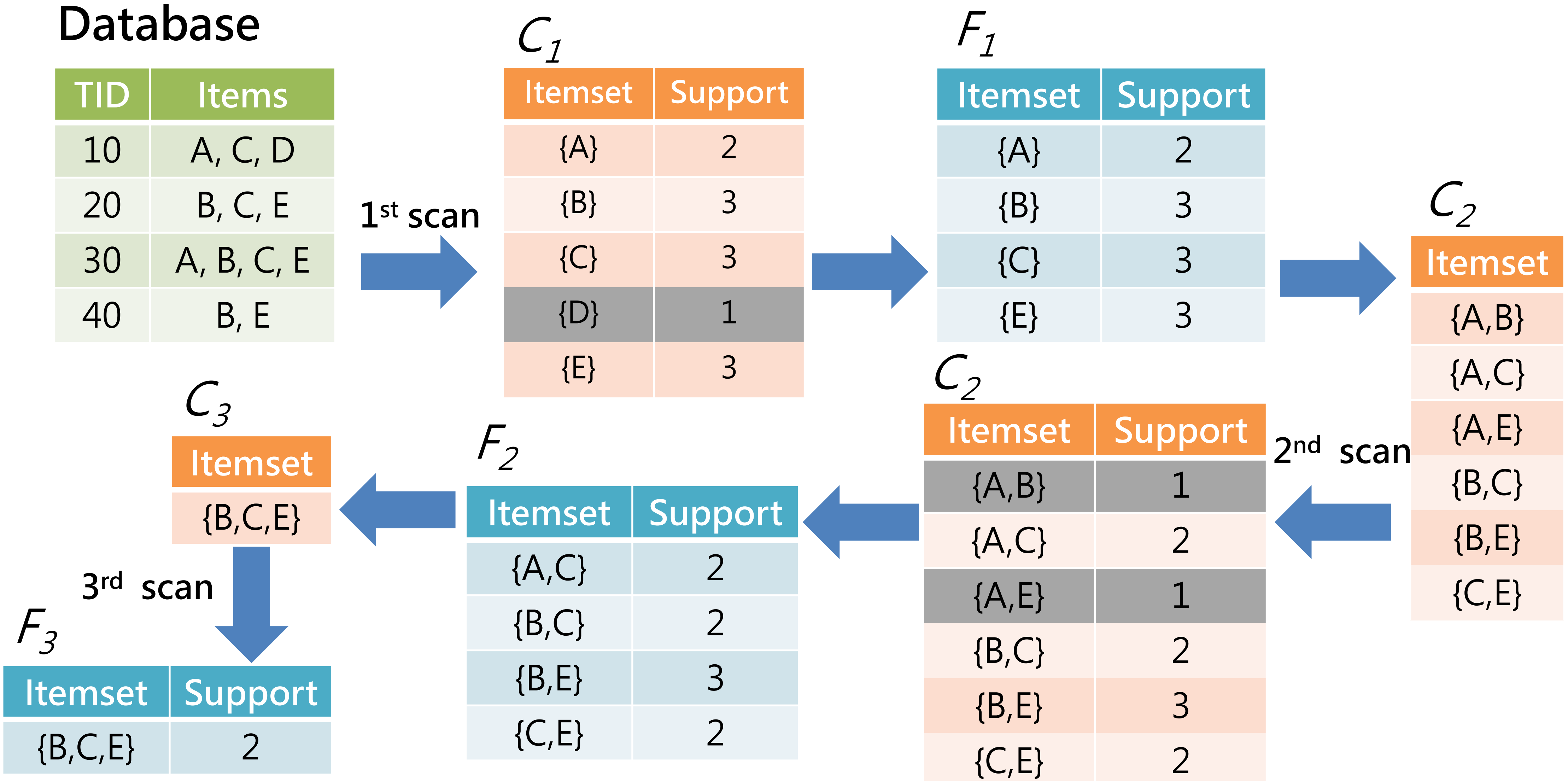
- 商品的項目集合(itemset)，$I = \{ I_1, I_2, \ldots, I_m \}$。#Item
- 交易資料庫(Database)，$D = \{t_1, t_2, \ldots, t_n\}$。 #Transaction
- 關聯規則(Association Rule)， $X \rightarrow Y$

Apriori

概念

- 逐層搜索的迭代方法。
- k -itemset 用於探索（ k + 1 ） - itemset。
  1. 找出 frequent 1-itemset，F_1。F_1用來找 frequent 2-itemset，F_2。而 F_2 用來找到 F_3。直到不能找到 k-itemset。
  2. 每找一個 F_k 需要掃描一次資料庫。為提高頻繁項集逐層產生的效率，Apriori 性質則可減少搜索。
- Apriori 性質：frequent itemset 的所有非空子集都必須是頻繁的。
  - 若某個 k-itemset 的 candidate 的 subsets 不在 (k-1)-itemset 時，這個 candidate 就可以直接刪除。

當最小支持度為 2 時的情況



Database

TID	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan

C_1

Itemset	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

F_1

Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A,B}
{A,C}
{A,E}
{B,C}
{B,E}
{C,E}

2nd scan

C_2

Itemset	Support
{A,B}	1
{A,C}	2
{A,E}	1
{B,C}	2
{B,E}	3
{C,E}	2

F_2

Itemset	Support
{A,C}	2
{B,C}	2
{B,E}	3
{C,E}	2

C_3

Itemset
{B,C,E}

3rd scan

F_3

Itemset	Support
{B,C,E}	2

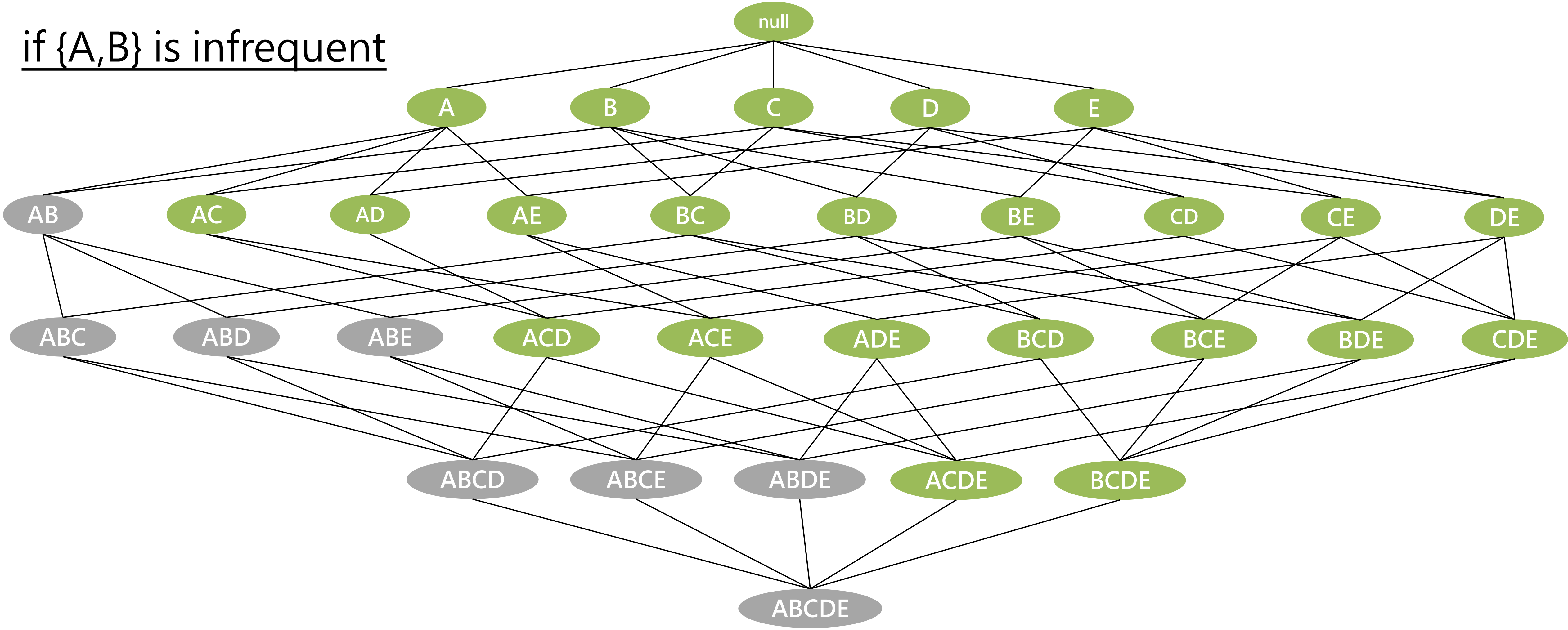
方法

1. $C_3 = F_2$的組合
   - F_2 = {{A, C}, {B, C}, {B, E}, {C, E}}

     C_3 = {{A, B, C}, {A, C, E}, {B, C, E}}
2. 使用 Apriori 性質剪枝：某個 frequent itemset 的所有 subsets 必須是頻繁的，對 candidate itemset C_3，我們可以刪除其非頻繁的 subsets ：
   - {A, B, C} 的 2-itemset 是 {A, B}, {A, C}, {B, C}，其中 {A, B} 不是 F_2的元素，所以刪除;
   - {A, C, E} 的 2-itemset 是 {A, C}, {A, E}, {C, E}，其中 {A, E} 不是 F_2的元素，所以刪除;
   - {B, C, E} 的 2-itemset 是 {B, C}, {B, E}, {C, E}，所有 2-itemset 都是 F_2的元素，因此保留。
3. 剪枝後得到 C_3 ={{B, C, E}}

剪枝

if {A,B} is infrequent



案例

TID	網球拍	網球	運動鞋	羽毛球
1	1	1	1	0
2	1	1	0	0
3	1	0	0	0
4	1	0	1	0
5	0	1	1	1
6	1	1	0	0

- 顧客購買記錄的資料庫 *D*，包含 6 個 Transactions
- 項目集 *I* = {網球拍, 網球, 運動鞋, 羽毛球}

觀察關聯規則，網球拍 → 網球。

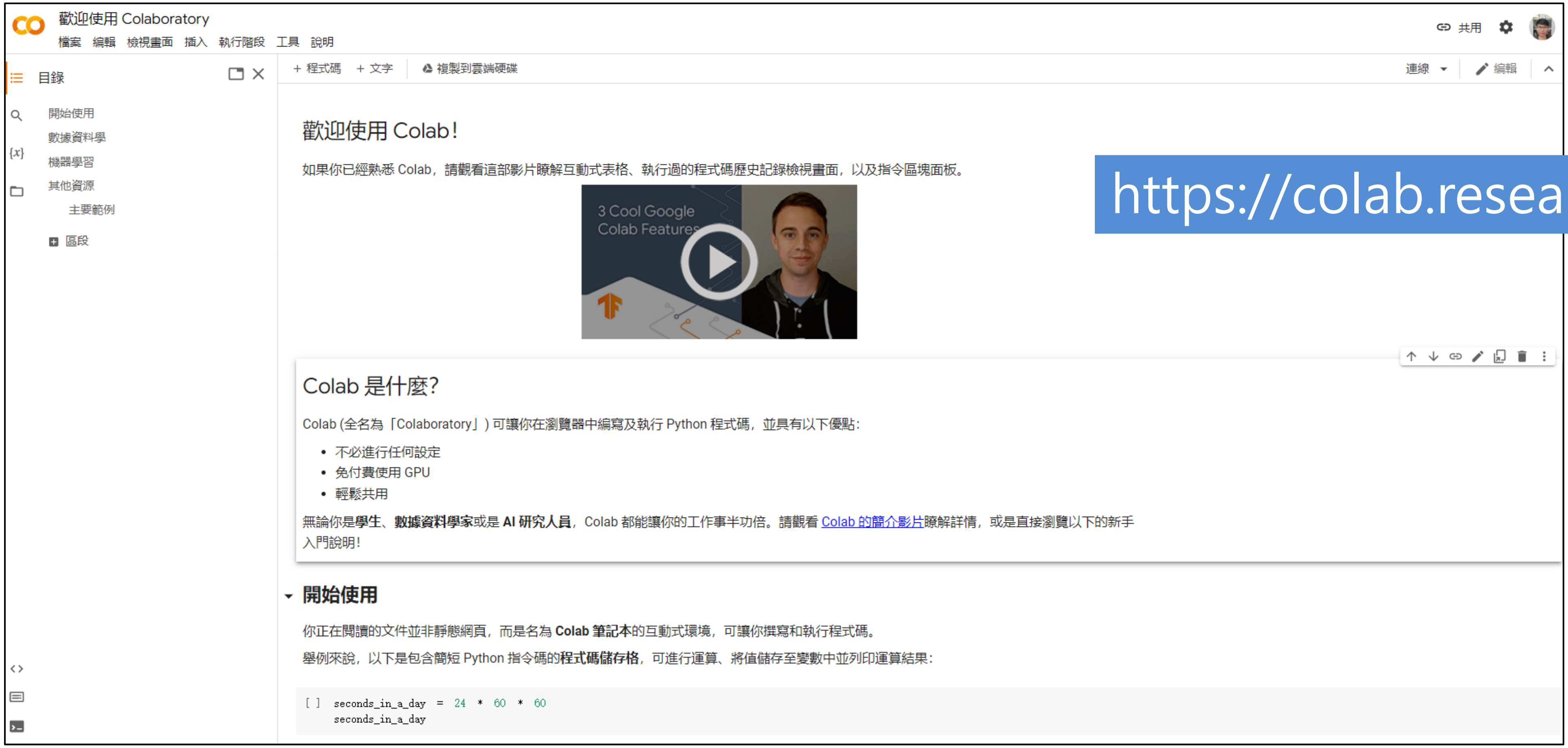
1. Transaction 1, 2, 3, 4, 6 包含網球拍。
2. Transaction 1, 2, 6 同時包含網球拍和網球。
3. 支持度 = 3/6 = 0.5，信心度 = 3/5 = 0.6。

- 若最小支持度為 0.5，最小信心度為 0.6。
- 關聯規則 “網球拍→網球” 是存在強關聯的。

- 1-itemset (4): {網球拍}, {網球}, {運動鞋}, ~~{羽毛球}~~
- 2-itemset (7): {網球拍, 網球}, {網球拍, 運動鞋}, ~~{網球拍, 羽毛球}~~, {網球, 運動鞋}, ~~{網球, 羽毛球},~~ ~~{運動鞋, 羽毛球}~~
- 3-itemset (4): {網球拍, 網球, 運動鞋}, ~~{網球拍, 網球, 羽毛球}~~, ~~{網球拍, 運動鞋, 羽毛球}~~ ~~{網球, 運動鞋, 羽毛球}~~

實作開始

Google Colab


歡迎使用 Colaboratory
檔案 編輯 檢視畫面 插入 執行階段 工具 說明
共用
+ 程式碼 + 文字 複製到雲端硬碟
連線
編輯
目錄
開始使用
數據資料學
機器學習
其他資源
主要範例
區段
3 Cool Google Colab Features

歡迎使用 Colab!

如果你已經熟悉 Colab，請觀看這部影片瞭解互動式表格、執行過的程式碼歷史記錄檢視畫面，以及指令區塊面板。

Colab 是什麼?

Colab (全名為「Colaboratory」) 可讓你在瀏覽器中編寫及執行 Python 程式碼，並具有以下優點:

- 不必進行任何設定
- 免付費使用 GPU
- 輕鬆共用

無論你是**學生**、**數據資料學家**或是 **AI 研究人員**，Colab 都能讓你的工作事半功倍。請觀看 Colab 的簡介影片瞭解詳情，或是直接瀏覽以下的新手入門說明!

開始使用

你正在閱讀的文件並非靜態網頁，而是名為 **Colab 筆記本**的互動式環境，可讓你撰寫和執行程式碼。

舉例來說，以下是包含簡短 Python 指令碼的**程式碼儲存格**，可進行運算、將值儲存至變數中並列印運算結果:

```
[ ] seconds_in_a_day = 24 * 60 * 60
    seconds_in_a_day
```

https://colab.research.google.com/

Mlxtend

http://rasbt.github.io/mlxtend/api_subpackages/mlxtend.frequent_patterns/

min_support : float (default: 0.5)

✓ A float between 0 and 1 for minumum support of the itemsets returned.

transactions_where_item(s)_occur / total_transactions

$$\text{support}(A \rightarrow C) = \text{support}(A \cup C), \quad \text{range: } [0, 1]$$

$$\text{confidence}(A \rightarrow C) = \frac{\text{support}(A \rightarrow C)}{\text{support}(A)}, \quad \text{range: } [0, 1]$$

Mlxtend

http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

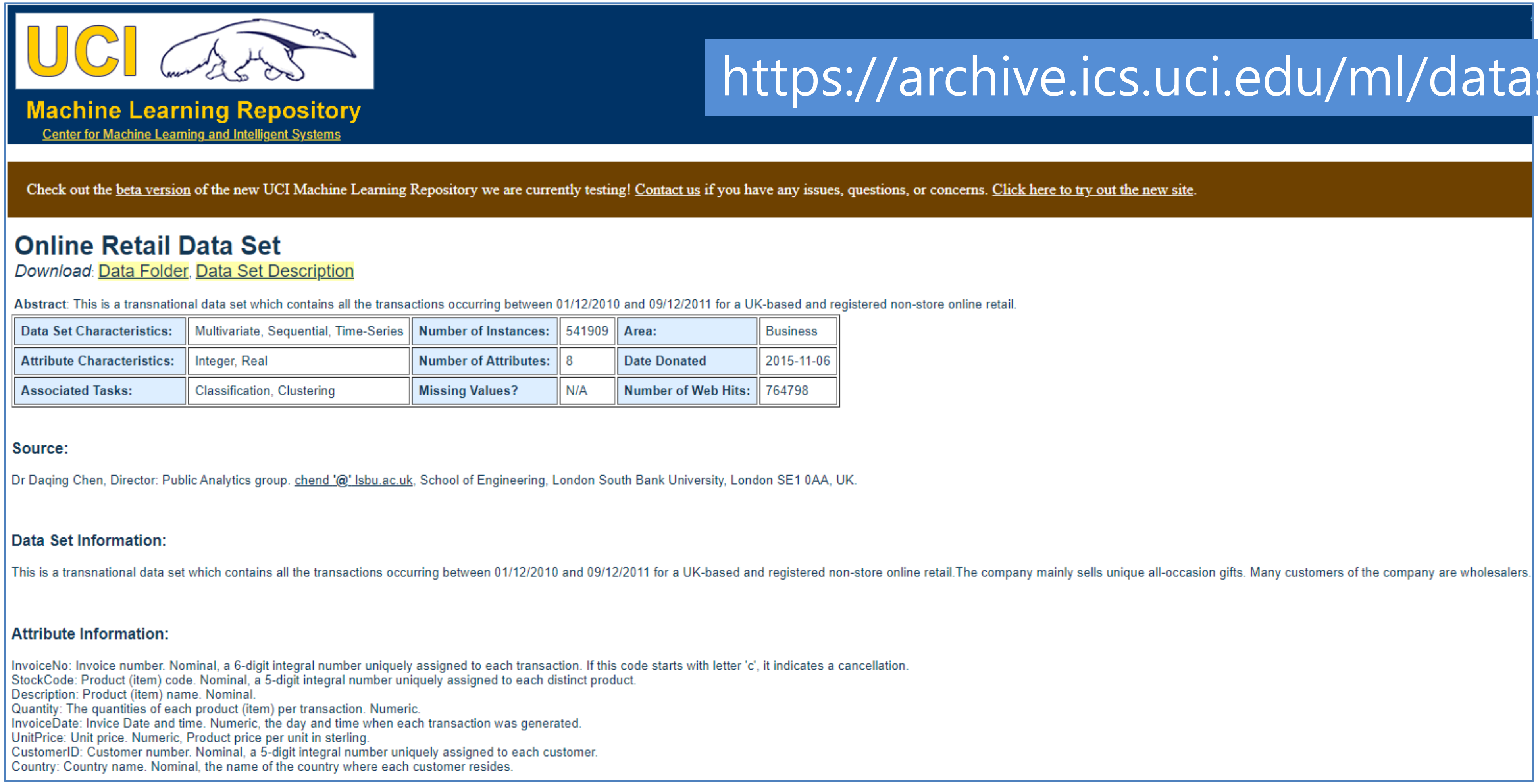
$$\text{lift}(A \rightarrow C) = \frac{\text{confidence}(A \rightarrow C)}{\text{support}(C)}, \quad \text{range: } [0, \infty] \quad = \quad \frac{Support}{Supp(X) \times Supp(Y)}$$

$$\text{levarage}(A \rightarrow C) = \text{support}(A \rightarrow C) - \text{support}(A) \times \text{support}(C), \quad \text{range: } [-1, 1]$$

$$\text{conviction}(A \rightarrow C) = \frac{1 - \text{support}(C)}{1 - \text{confidence}(A \rightarrow C)}, \quad \text{range: } [0, \infty] \quad = \quad P(A)P(B')/P(A \cap B')$$

UCI - Online Retail Data Set

https://archive.ics.uci.edu/ml/datasets/online+retail



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if you have any issues, questions, or concerns. Click here to try out the new site.

Online Retail Data Set

Download: Data Folder, Data Set Description

Abstract: This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

Data Set Characteristics:	Multivariate, Sequential, Time-Series	**Number of Instances:**	541909	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	2015-11-06
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	764798

Source:

Dr Daqing Chen, Director: Public Analytics group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

Data Set Information:

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Attribute Information:

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
Description: Product (item) name. Nominal.
Quantity: The quantities of each product (item) per transaction. Numeric.
InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.
UnitPrice: Unit price. Numeric, Product price per unit in sterling.
CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
Country: Country name. Nominal, the name of the country where each customer resides.

懶人包

[link]

Thank you