從零開始的關聯式學習

Pandas 與 Mlxtend

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

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

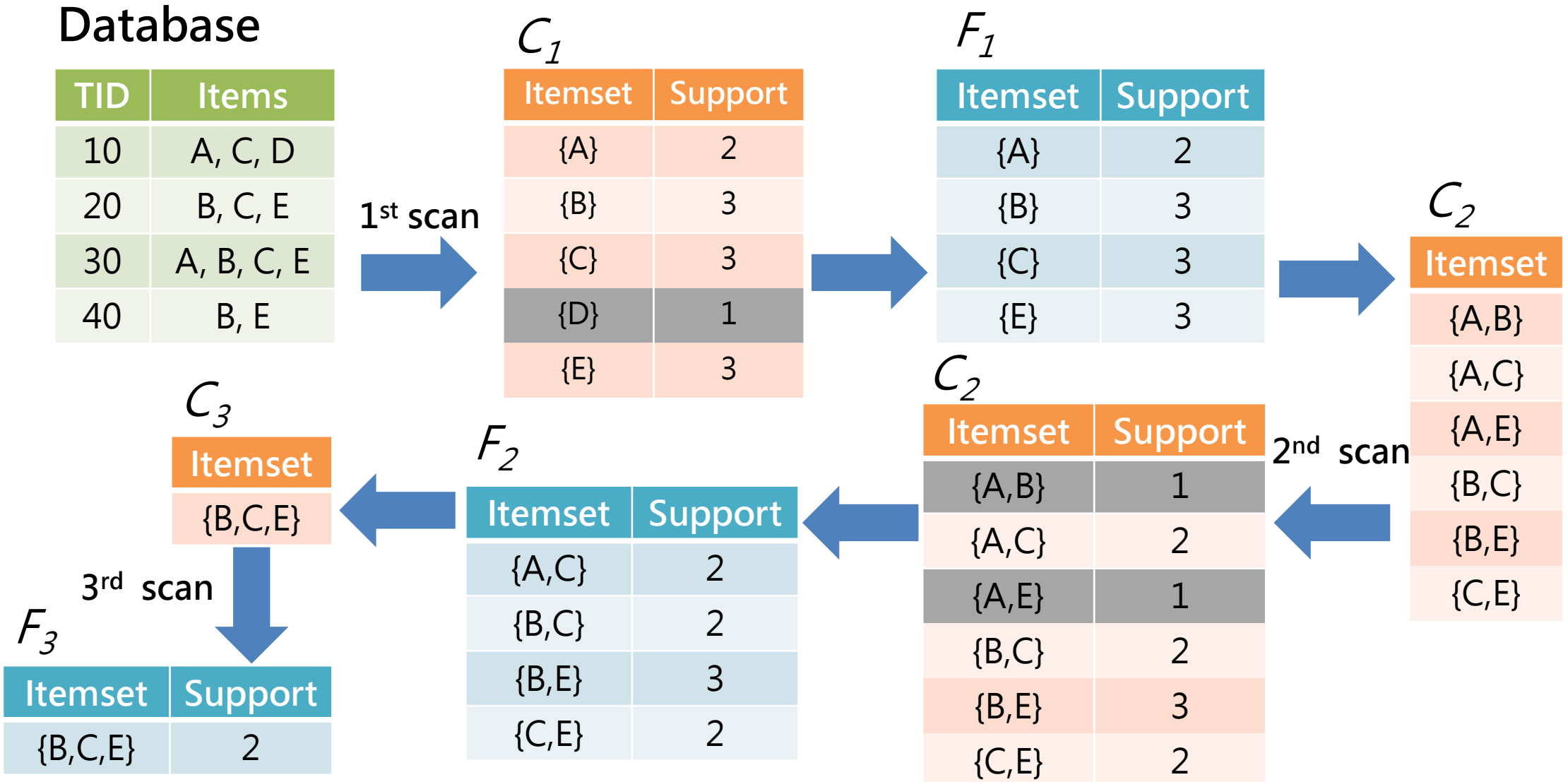
- 商品的項目集合(itemset)，$I = \{ I_1, I_2, \ldots, I_m \}$ 。#Item
- 交易資料庫(Database)，$D = \{t_1, t_2, \ldots, t_n\}$ 。 #Transaction
- 關聯規則(Association Rule)， $X \rightarrow Y$

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



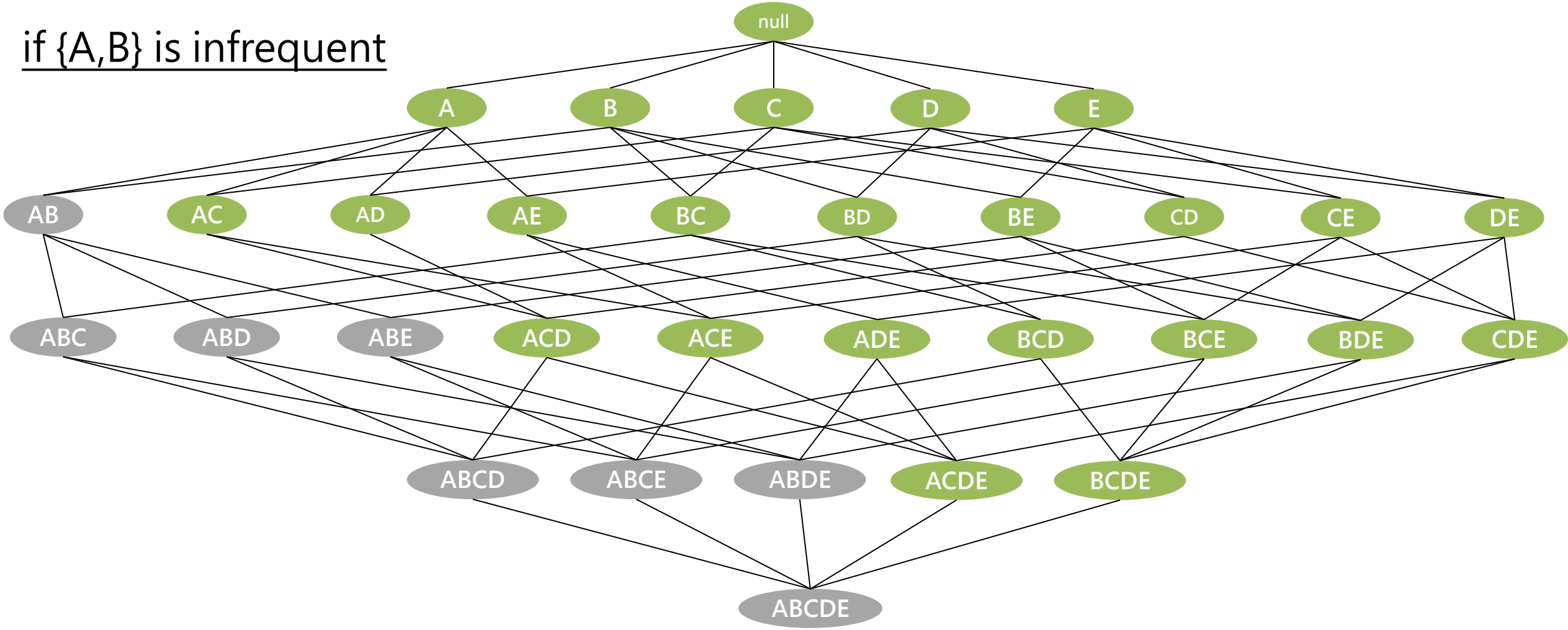
方法

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

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

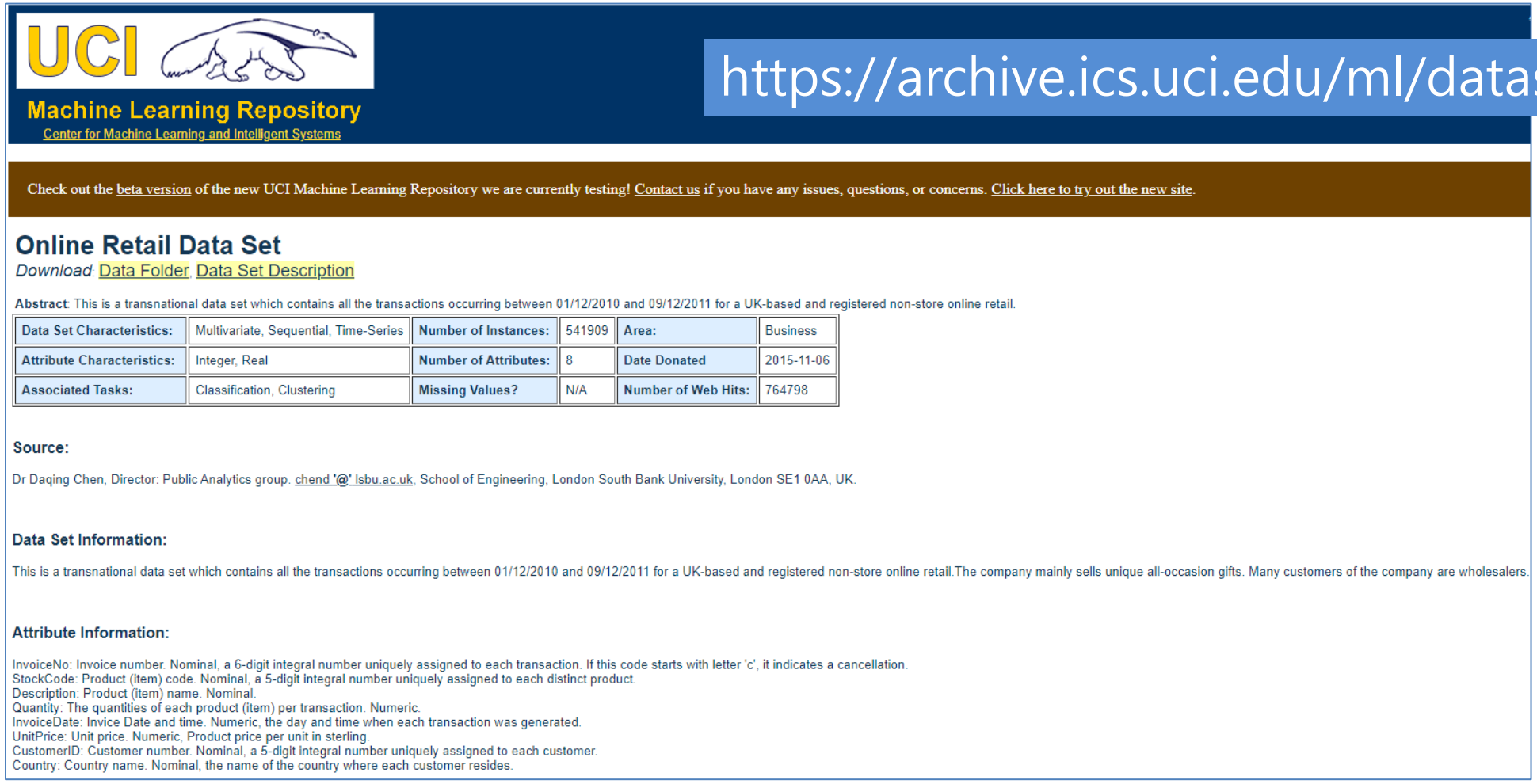
http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

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

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

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

UCI - Online Retail Data Set



UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

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

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