關聯規則

電子商務

If 購物車裡有A商品，能否猜中下一個會買什麼?

用過去資料推估🡪id,item

🡺Association rule

Apnorai

FP tree(改善版)🡪把要的資料直接塞記憶體，就不用一直進資料庫

Apnorai算法

交易資料集(不討論數量、前後順序)-->|D|

|  |  |
| --- | --- |
| TID | items |
| 1 | 2,5,7 |
| 2 | 1,3,4,6 |
| 3 | 2,6,7 |
| 4 | 2,4,5 |
| 5 | 3,6 |
| 6 | 2,4,6 |
| 7 | 1,4,5 |
| 8 | 1,3,5 |
| 9 | 2,4,5 |
| 10 | 1,3,5 |

Itemset(項目集)

A={ } 如果A裡面有一個(1 itemset)…以此類推

A={2},B={1,3},C={2,4,5}

A的組合在裡面出現5次🡪5支持數量(support count) 演算法符號 sigma(A)=5

S(A)=5/10 🡪50% sigma(A)/|D|

B 3(support count) sigma(B)=3 S(B)=3/10 🡪30% 組合占比

C 1(support count) sigma(C)=1

能否從交易資料庫中看出，哪些商品經常被客戶一起買走

「經常」🡪幾次? 由分析師決定「門檻值」🡪最小支持度(minsup)+最小信賴度minconf

假設資料量大，門檻值設太低會跑到死

🡪符合實際情況分析(太高資料少，太低資料過多) 最佳門檻值未定論

門檻值可用小數點、數量、百分比

C1(1 itemset)

|  |  |
| --- | --- |
| 1 itemset | supcount |
| {1} | 4 |
| {2} | 5 |
| {3} | 4 |
| {4} | 5 |
| {5} | 6 |
| {6} | 4 |
| {7} | 2 |

Minsup=3 所以7不符合

Frequent Itemset L1

|  |  |
| --- | --- |
| 1 itemset | supcount |
| {1} | 4 |
| {2} | 5 |
| {3} | 5 |
| {4} | 4 |
| {5} | 6 |
| {6} | 4 |

C2 itemset

|  |  |
| --- | --- |
| 2 itemset | supcount |
| C62種組合 | 比對資料庫的數量 |
| {1,2} | 0 |
| {1,3} | 3 |
| {1,4} | 2 |

.

.

.

一直持續下去比對，然後不到3的砍掉變成L2

L2

|  |  |
| --- | --- |
| 2 itemset | supcount |
| {1,3} | 3 |
| {1,5} | 3 |
| {2,5} | 3 |
| {3,5} | 3 |

C3組合，首項相同者可以組成一種，不同者不能組 EX:1,3、1,5🡪1,3,5 (演算法)

C3

|  |  |
| --- | --- |
| 3 itemset | supcount |
| {1,3,5} | 2 |

得出L3被排出(空集合)

關聯規則(求出高頻項目集frequent itemset) 不考慮信賴度

R1:X🡪Y[支持度(support),信賴度(confident)]

支持度要大於最小支持度(0~1)

信賴度要大於最小信賴度(0~1)

關聯規則一定要兩個一組🡪L1沒辦法用 根據上面題目，只剩L2

支持度= 每個組合都3次，除以總數(10筆)得出30%

信賴度= 某條規則的支持度/左手項的支持度(X) L1的次數

🡺已知X被買，那Y也被買的機率

{1}🡪{3}[30%,75%]

{3}🡪{1}[30%,60%]

{1}🡪{5}[30%,75%] 信賴度算法:[30%] / [4/10]🡪75% Lift:[75%]/[6/10]🡪1.25

{5}🡪{1}[30%,50%]

{2}🡪{5}[30%,60%]

{5}🡪{2}[30%,50%]

{3}🡪{5}[30%,60%]

{5}🡪{3}[30%,50%]

設最小信賴度是70% 得出上方兩個結果可用

🡺買1後也買3的結果占資料庫的30%

🡺買1之後也買3的機率是75%

要發掘高利潤的、有趣的結果。麵包配牛奶這個組合不僅無趣也習以為常

若是應用在電子商務，就會在1的頁面放3的廣告

若是應用在賣場，就會重新安排貨架把1,3放附近

IF {1,3,5}及格的L3

{1}🡪{3,5}

{3}🡪{1,5}

{5}🡪{1,3}

+反過來的組合🡺共六種

提升度(lift)

提升度 = 某條規則的支持度/S (X)\*S(Y) 看項目次數/總數

得出的結果可看出該規則的重要度>1正相關 =1無關 <1負相關

0318 class

**分類技術補充**

#可以查看一下目前xlist變數資料內容，list型態在R中使用兩層中括號

xlist

xlist[[1]] #"飯糰","豆漿","尿布"

xlist[[2]] #"飯糰","尿布","啤酒","麥片"

xlist[[3]] #"豆漿","尿布","啤酒","綠茶"

xlist[[4]] #"飯糰","豆漿","尿布","啤酒"

xlist[[5]] #"飯糰","豆漿","尿布","綠茶"

xlist[[1]][1] #"飯糰"

xlist[[1]][2] #"豆漿"

xlist[[1]][3] #"尿布"

xlist[[1]][4] #NA表示沒資料，超出範圍所以沒資料

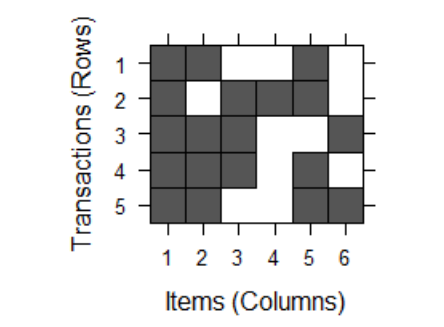
R語言的list還可以存matrix,dataframe,column(向量)

Aporai步驟

1.建立list

2.幫list命名(TID)

3.轉成transection型的資料🡪電腦把資料變成0,1矩陣(有=1(黑),無=0)



電腦計算時，會將係數矩陣的0砍掉，變成算行、列、有無值(0,1)

|  |  |  |
| --- | --- | --- |
| 行 | 列 | 值 |
| 1 | 1 | 1 |

a density of 0.6333333

分布狀況很重要🡺判斷門檻值

most frequent items:

尿布 豆漿 飯糰 啤酒 綠茶 (Other)

5 4 4 3 2 1

賣最好的從左邊數過去

writing ... [11 rule(s)] done [0.00s].

跑出11條規則

設定時x跟y不可是空集合，所以minlen要設定成2

rules\_1 = apriori(table14\_1, parameter = list(support = 0.6,

confidence = 0.6,

minlen=2,

target = "rules"))

要是都不設定，他會有預設值，但minlen=1 會導致某項出現空集合無法解釋規則

aporai演算概念:infrequent itemset🡪直接砍掉，例如D一開始就不符合門檻，後面有D的任何組合就不用算

資料整理方法

excel 資料 左上角從文字檔🡪根據區分符號選取就可以匯入

電子商務的推薦清單可以用association達成

把跟某個商品相關高的都放上去

圖片上課 5-2

MBA購物籃分析🡪關聯分析

Aporai是最早提出的，但要一直進資料庫搜索很慢

FP tree 是昇級版

Aporai 5-9

L1🡪C2 的時候有一個「結合」和「刪除」的動作(aporai特色)

若是把每種可能都列出來很耗時間

提出者透過觀察發現把根本不大於最小支持度的可能截斷

EX:A集合不符合minsup，那後面A集合的可能性全部都不納入(反單調函數)

Tip:用商品名稱跑程式，字串差可能導致錯誤或誤差，最好用id跑

演算法5-14

不一定要交易資料才能用，可以用基本資料看哪些會員的特性一起出現

但是欄位要變(跟老師問詳細)

Association rule 與 Classification(資料分類技術)

Ar 是在看資料有沒有關聯

Y=f(x1,x2,…,xn)

Y目標變數 f(x)自變數

Classfi 做出來的是模型🡺進行預測用

所以要知道目標變數是什麼🡺EX:是否購買

性別是否影響是否購買?收入是否影響是否還款?....

屬性的界定要思考清楚EX:收入高低是否分高中低

但通常都會並成兩種🡪binary data

重點在「了解目標變數」🡪然後建模

建模🡪採用技術(演算法):策略、方法🡪結論報告

監督式vs非監督式

KNN

決策樹

這兩種演算法，可以做調整參數後的正確率趨勢，可做序列分析來看上升下降

貝氏分類法

Entrophy

I(1,2)=0.917

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 性別 | p | n | 小記 | IA(D) | IA(D) | Gain |
| 男 | 1 | 1 | 2 | 2/3\*I(1,1) | 0.1 | 0.516 |
| 女 | 0 | 1 | 1 | 1/3\*I(0,1) | 0.301 |  |
|  |  |  | 3 |  | 0.401 |  |
|  |  |  |  |  |  |  |
| 婚姻 | p | n | 小記 | IA(D) | IA(D) | Gain |
| 結婚 | 0 | 1 | 1 | 1/3\*I(0,1) | 0.301 | 0.516 |
| 單身 | 1 | 1 | 2 | 2/3\*I(1,1) | 0.1 |  |
|  |  |  |  |  | 0.401 |  |

兩個都可以選，因為gain都一樣

Gini

I(1,2)=0.444

貝氏定理(請見參考圖)

自變數必須互相獨立(但現實很難達到

所以這個方法不適合用在正式研究(但可以用自己的演算法拿來比較

自己提出的算法要拿去跟其他演算法來比較，以證明自己的比較好或差

可用決策樹、貝氏、SVM(支援向量機)來比較，不能比前兩個差

CNN,RNN,LSTM,BERT, Xgboost…都是分類預測的方法

資料辨識=資料分類🡪拿來做預測

混淆矩陣🡪判斷模型好壞

正確率很容易跑掉(資料不平衡 unbalance)🡪正確分類/資料總數(看斜對角)

Balance data:正確與不正確各一半

所以要用精準度P(看直欄)、召回率R(看橫列，也等於單項的正確率)來看(參考圖1,2)

或跑ROC、F1 measure(調和平均數)🡪2(P)(R)/P+R

**總結(資料分類預測的步驟)**

1. (清洗資料)

整理資料(了解資料是否為unbalance)

找出目標變數y、自變數f(x1,x2…,xn)(你想解決或預測什麼問題?)

分割成training, testing(隨機分配or交叉驗證，約75:25)

1. (建模演算)

挑選演算法+建模(KNN,決策樹,SVM,貝氏,BPN, ANN, Deep learning…)

將分割的data丟入模型運算

1. (評估)

得出的結果拿去製作混淆矩陣

求出正確率A、準確率P、招回率R(目標偏重左上or右下正確率不同)

算調和平均數F1

可做出調參數時的正確率變化(EX:KNN=1~5的正確率變化)

將公司營運數據🡪照片🡪矩陣🡺用CNN跑

訓練資料的混淆矩陣意義不大🡪資料都看過了，猜對機率高

是用測試用的資料去評估(建混淆矩陣)🡪沒看過的資料正確率就取決於模型強度

所以是最後一步

交叉驗證(K折交叉驗證)

隨機抽樣是否公平?(歐洲人VS非洲人)🡪因此誕生的方式

假設有100個資料，前90個當訓練，後10個當驗證。下一次是80~90號的資料當驗證，以此類推🡪k代表要做幾次(幾折)

先把資料隨機編號，然後k=1跑一號的資料進模型得出結果。以此類推

然後把每個結果正確率的平均當作最後的結果。(可以用趨勢圖看每個K)

為了避免依賴某一特定的訓練和測試資料產生偏差。

處理資料:例如五等量表(4,5較好的，1,2較差的，那3呢?)

可以好壞各分一半分配

重點:空白不等於0分

空白質的補植要有理由:內插法or EM

ANN類神經網絡

訓練時間夠久、夠多就會準

但是為什麼這麼準，難以說明

Ex:判斷照片是狗還是貓

AI有分為:分類器、神經網路、統計學習方法

細胞核:接收訊息🡪判斷是否達到門檻值🡪決定是否傳遞給下一個細胞核

接收訊息:每條線有不同權重🡪加總輸入

輸出:看是否要加權🡪輸出給下一個細胞核

每個神經元(節點)的工作程序

X1~Xn\*各自的權重(W)🡪彙總+b(修正巷)🡪門檻判斷🡪輸出

神經元(perceptron)又稱為感知機🡪多重感知機(神經網絡)

+,-,\*,/ 🡺二元運算子(兩個數才能做)

And,or,not,Xor 🡺二元邏輯運算子

XOR🡪兩個運算元顛倒時才為真

感知機(資料分類🡪預測)

And、or問題處理，可區分兩筆資料(1,0)🡪可以畫很多條不同斜率，每條線就是一個分類器

但XOR要兩條線裁切的完整🡪線性不可切割，演算法會無法收斂

OR的問題

電燈泡，兩條線路、開關。只要一個開關接上就會亮，都沒接上才不亮。

用點斜式算出斜線🡪分類完成

X1+x2-1/2=0

X1,x2的係數都是1🡪權重

-1/2是修正項

這是 single layer perceptron

目標:如何讓實際值與預測值的誤差最小化

Min[sigma xiwij+b-T]

層次

Input🡪hidden🡪output

 x0是bias

|  |  |  |  |
| --- | --- | --- | --- |
| **AND** | | | |
| X0 | X1 | X2 | T |
| 1 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 1 |

T輸出值，X輸入值

1.預測值>實際值🡪權重要下調N單位(調整輸入值有1的權重就好)

2.預測值<實際值🡪權重要上調N單位(調整輸入值有1的權重就好)

3.預測值>1視為1 預測值<0視為0

N可以自訂，權重啟始值也可自訂

**公式：［X0X1X2］＝X0W0+X1W1+X2W2**

設權重為000把各個輸出放入計算決定是否調整權重

(請見資管作業的word檔案)

|  |  |  |  |
| --- | --- | --- | --- |
| **OR** | | | |
| X0 | X1 | X2 | T |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 |

設權重[1,0,0]

[1,0,0][1,0,0]=1+0+0=1 ok

[1,0,1][1,0,0]=1+0+0=1 ok

[1,1,0][1,0,0]=1+0+0=1 ok

[1,1,1][1,0,0]=1+0+0=1 ok

如何處理unbalance的資料(陰陽資料數量差很多)，且要避免oversampling

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive(pred) | negative(pred) |  |
| Positive(Actual) | TP | FN  Type 2 error | TP/(TP+FN)  真陽性比例  查全率(Recall) |
| negative(Actual) | FP  Type 1 error | TN | TN/(FP+TN)  真陰性比例 |
|  | TP/TP+FP  查準率(Precision) | TN/TN+FN | Accuracy  TP+TN/all |

查準率(Precision):

你的演算法抓回來的資料，真的是你有興趣的實際有多少比例。(涵蓋了negative的資料)。🡪看雜訊會不會很多

查全率(Precision):

你有興趣的資料，經過演算法後有多少是被正確分類的。(只有positive的資料)

這兩個有點成反比，一個高另一個通常會低一點。🡪想要的有沒有都被抓到

調和平均數(f1 measure)2PR/P+R

處理unbalance的資料，可以更好的算出平均數

使用演算法一定要知道自己要找什麼。有興趣的就是positive

X:自變數 Y:目標變數

X.train = training\_data[,1:4]

Y.train = training\_data[,5]

X.test = testing\_data[,1:4]

Y.test = testing\_data[,5]

KNN在R裡面要切成四塊 train就是訓練資料及的自變數+目標變數(答案)

Test也是

R跑出來的決策樹

EX:花瓣長度<2.45往左走(是)🡪條件成立就往左走，不是就往右

貝式分類法🡪漸增性資料是用🡪他會直接幫你更新得出結果適合快速分

但無法解釋🡪不像決策樹那樣了解來龍去脈

考關聯規則、決策樹(EGM都要)、感知機、倒傳遞、KNN、混淆矩陣

自己準備紙

文字探勘

Text🡪written resources or document(一句話也算)

包含手寫和打字

80%資料都是非結構資料(亂)

文字雲算一種

把商機高的資料挖出來🡪資料分析師

Oracle data base

集群分析

非結構資料會用到

資料量大適合用K mean

中文文字探勘

List.files 可以直接讀某資料夾裡面的檔案

用notepad++可以看出文字檔資料總共有幾行

或是直接跑進來看標記得號碼

中文斷詞系統(CKIP

文章太多的時候最後的表格會越來越大，所以要把表格縮減，才知道重要的keyword是哪些。要降維(PCA

前面都在找文章重要程度的文字

但有些人覺得會扭曲🡪直接找topic代表文章意涵

主題、文字、文件 三大元素

套件名稱:topicmodel

輸入資料室使用DTM次數矩陣

函數名稱:LDA

計算方式:VEM、Gibbs

功能:注重數值資料

套件名稱:lda

函數名稱: lda.collapsed.gibbs.sampler

計算方式:Gibbs

功能:注重視覺化呈現

這兩個的資料格式不一樣，但有個函數可以把topicmodel的資料轉到lda

dtm2ldaformat

文字探勘

至少要一千篇才準

中文🡪seg🡪矩陣(統計文字 tfidf🡪sentiment🡪建模(ID3 決策樹 KNN 貝氏 BPN

🡪預測🡪混淆矩陣+AC Recall Precision

要記得特徵篩選(tfidf家總值，排序，把重要的10~20留下

文章的topic 要用tf矩陣做

Dtm🡪topicmodel(設定K🡪視覺圖、解釋

把字分群 clustering

Recommendation 推薦 (集群分析也可以做推薦)

1.content-based

2.collaborative Filtering(CF) 相似度、權重🡪user base or itembase

又稱為口碑推薦法

3.model base

4.Hybird(混合

常用在社群或電商平台

推薦🡪減少人員的搜尋時間

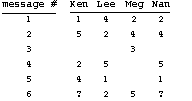
現在存在太多information，但只要記錄習慣的pattern就能把相似的東西推薦給使用者(user base)

口碑🡪相似度

顯性資料:對各產品打分數

隱性資料:個資

User item matrix



誰跟Ken比較像?(算直的) itembase方法哪兩篇文章像(算橫的)

KLMN 平均數都是3

R ken lee 相關係數🡪

(1-3)(4-3)+(5-3)(2-3)+(2-3)(5-3)+(4-3)(1-3)/

MSE、RMSE都是在看你準不準

K6pred🡪K平均+{sigma J6-J平均\*rkj(相關係數)/sigma|rkj|}

要猜6號文章的分數=3+{2-3(-0.8)+(5-3)1/|rkl|+|rkm|(分子:加上有打過分數的人的分數\*相關係數)分母:預測格與其他有打分的相關係數

2-3\*4-3+4-3\*2-3+5-3\*5-3+1-3\*1-3

作業:猜出兩個問號，用itembase跟userbase的方法

MSE

讓模型去猜全部的分數，比對正確分數，蒐集誤差，就能得出平均誤差

RMSE

每個欄位都是向量(每個人打分數的輪廓

新文章或會員遇到的問題

coldstart冷啟動🡪不知道推薦給誰 (推薦口碑最好的文章

分數可以灌水(貶低pull或架高nake都可以 🡪shilling attack

現在演算法一定都要預測和推薦清單

概念統整

範例

金曲獎🡪這首歌為何會得獎?(分析)

音樂的意義探討、了解現況

敘述性統計🡪了解資料輪廓(分析、整理)

用模型分析入圍歌曲與獲獎歌曲的差異

統計每首歌字數、段詞、不重複詞數

可畫出散佈圖看相關🡪字最多、字最少

分出不同語言的歌曲(國、台、客…混雜不同語言的)

比較入圍和得獎作品(不同變項影響下哪個多)

可以把得獎和入圍最多次的極端值拉出來(看誰厲害

也可看重複詞最多的

重點:對某些人有幫助的資訊才重要

分析情緒

TF、TFIDF文字雲🡪看出用最多的字是啥

Topicmodel 主題關鍵字🡪要怎麼賦予跑出來的主題意義是看研究者

得出的得獎主題範例:飄飄蕩蕩…戀愛、愛情迷惘、告別戀情

IDA六大主題文字雲

個文件分配主題(得獎的每首歌個描述了哪些主題)

(同一套方法再拿來跑入圍看看結果如何)

入圍歌曲的文字雲

根據年分去分每個年代的風格(不同世代的風格分析)

專案面對的是客戶。技術是埋在後面的。

第一章的緒論通常是最精采的，包裝的精采就能吸睛繼續往下看或是受到青睞

企劃書

為什麼要做這個project🡪解決某種問題(內在本質或現實)

例如:很多人會排隊打卡🡪你如何解決?

表面上是大家上班打可排隊沒效率(實際上是因為大家都快上班才來)

我會想用公司APP定位到場打卡，資料庫紀錄大家打卡的時間來morningcall

這個問題值不值得你去解決(有沒有替代的東西早就存在?)

解決問題的程度有多高

考試題目

看圖片 x3056t1

Attack故意要讓item6進入清單或退出清單(約3人)

Alice會不會喜歡item6?

誰跟alice打分很像?

TFIDF觀念、集群分析觀念、推薦方法觀念

建模和collaborative filtering example 差別在哪

資料探勘怎麼用?

大數據分析、AI也可做大數據分析

重點在用不同角度去解釋資料

1.DATA哪裡來? 資訊系統(ERP)，基層蒐集或產生🡪生產、行銷、研發、會計

收在mysql(關聯式資料庫🡪一大堆的二維表格

例如業務收進的訂單明細就會存進去資料庫變成各種欄位資料

用SQL合併就能分析🡪自己撈資料是最快的方法

假設今天有五萬張訂單狀況🡪要給他匯總性的資料(決策報表)

例如產品在國內賣的狀況🡪EX:各季賣了多少…

每個部門要的報表不同🡪BI資料倉儲(更新的資料隨時算好並根據部門跑出資

料表)

ERP在蒐集資料🡪 BI在展現資料

RDB(關聯式資料庫) star schema 星狀綱要(立體的資料)

ER model(資料要正規化) DW資料庫

DB 資料庫

BI要能讓老闆一點就出他需要的報表，而且要能追蹤有問題的部門

ERP要讓流程圖中的每個箭頭都順利執行

BI可以用各種KPI指標用來監視ERP的每個流程都順利

商業數據分析(big data)去看其他資料是否對商業有幫助

文字探勘、關聯規則….數據重新整理、利用AI、ML也是

若想知道銷售情形🡪跟人(客戶價值RFM 客戶最近下單時間，下了多少單，貢獻多少錢)但RFM不會自己紀錄要去哪裡找?(交易明細在資料庫可以自己撈如果沒有權限那就把欄位撈出來)或商品有關

假設你把客戶ID跟RFM都撈出來後要做什麼?(集群分析、分類)

假設有ABCDE五群 把它放在二維圖上 X軸是下單量 Y是單價

所以越左下(買少單價低)的就是不好的客戶

右下的(買多單價低)可以考慮讓他買貴一點

總之都想辦法讓客戶往右上跑(買多、單價高)

資料科學家要做的就是預測和釐清現況

建立模型預測(用新欄位假設是品牌忠誠度，然後就可以用決策樹之類的去歸類預測)

產品除了賣之外，行銷也很重要，行銷的檔期內商品買得好不好(行銷方案好不好)。一樣可以做RFM看各產品的貢獻