Datorlaboration 6

Josef Wilzén

29 september 2022

Allmänt

Datorlaborationerna kräver att ni har R och Rstudio installerat.

- Kodmanual: länk
- Dataset till vissa uppgifter finns här.
- **ISL**: An Introduction to Statistical Learning,
  - Boken: länk
  - R-kod till labbar: länk
  - Dataset: länk och länk
- **IDM**: Introduction to Data Mining
  - Kod till boken finns här här
  - Sample chapters

Notera att ni inte behöver göra alla delar på alla uppgifter. Det viktiga är att ni får en förståelse för de olika principerna och modellerna som avhandlats. Dessa uppgifter ska inte lämnas in, utan är till för er övning.

Datauppdelning

För att motverka överanpassning bör ni dela upp data till träning-, validering-, (och testmängd). Detta kan göras med `createDataPartition()` från `caret`-paketet. Argument till den funktionen som är av vikt här är p som hur stor andel av observationerna som ska användas till träningsmängden. Ni kan också använda `subset()` för att göra detta också, men det blir svårare att tydligt ange de observationer som ska tilldelas till valideringsmängden. Denna uppdelning ska ske slumpmässigt. Notera att om en testmängd ska skapas måste uppdelningen ske en gång till från valideringsmängden.

Del 0: kod

- Associations- och sekvensanalys i Kodmanualen
- IDM, kod för Association Analysis
- Association Mining (Market Basket Analysis)
- Visualize Market Basket analysis in R
- Exempel på användning av arules finns här

Del 1: Utvinning av frekventa enhetsmängder och regler med hög konfidens

Datamaterialet ”marbas.csv” innehåller transactioner från ett antal livsmedelbutiker i södra Italien. Filen innehåller två variabler TRANS_ID och PRODUCT, där TRANS_ID beskriver vilken transaktion som observationen hör till och PRODUCT beskriver vilken produkt som köptes. Kod till vissa av uppgifterna finns här.

1. Importera filen till R och kontrollera datastrukturen. Kom ihåg att läsa in datamaterialet först som vanligt och sedan konvertera det till transaktionsformat efteråt.
2. Skapa en associationsanalys med en supporttröskel på 5 procent, det maximala antalet enheter i en regel till 2, och konfidenströskel på 50 procent.
3. Hur många regler fås ut från algoritmen? Visa de fem regler som har högst support och de fem regler som har högst konfidens. Vilka av dessa anser ni vara intressanta från er synvinkel?

4. Vilka är de riktiga minsta värdena på vardera mått som visas i slutresultatet? Överensstämmer de med de trösklar som angavs i steg 2?
5. Skapa en associationsanalys igen men denna gång ange 4 som det maximalt tillåtna antal enheter i en regel, supportnivåtröskeln till 100 transaktioner och samma konfidenströskel som tidigare. Plocka ut de tio regler som har högst konfidens och tolka vilka utav dessa som ni anser vara intressanta.

Del 2

Gå igenom koden här

Del 3: Intressemått

Använd här samma material som i del 1 och repetera steg 2 från denna.

1. Sortera de resulterande reglerna utefter Lift och visa de tio regler med högst Lift-värde.
2. Skapa följande nya intressemått utifrån den resulterande regeltabellen. I formlerna nedan är $P(A)$ och $P(B)$ supporten för vänster- respektive högerledet av regeln $A \rightarrow B$, och $P(A, B)$ är supporten för hela regeln. Tips: `lhs()`, `rhs()`, `support()`, viss kod finns här.

   (a)
   $$IS = \frac{P(A,B)}{\sqrt{P(A) \cdot P(B)}}$$

   (b)
   $$Klosgen = \sqrt{P(A,B)} \cdot (P(B|A) - P(B))$$

   (c)
   $$Jaccard = \frac{P(A,B)}{P(A) + P(B) - P(A,B)}$$

   (d)
   $$Laplace = \frac{P(A,B)}{P(A) + 2}$$
3. Många olika intressemått går att få fram med funktionen interestMeasure(), se även här för förklaringar.
4. Sortera den resulterande tabellen utefter alla intressemått och visa de tio regler med högst värden. Vilka av dessa regler anses vara intressanta?
5. Utforska hur listorna med regler skiljer sig från varandra och försök dra några slutsatser om intressemåttens huvudsakliga egenskaper. Hur skiljer sig asymmetriska och symmetriska intressemått? Tips: Kolla i denna artikel Selecting the right objective measure for association analysis

Del 4: Webssideanalys

Datamaterialet ”clickstream.csv” innehåller en log-fil från en e-handel webssida som säljer hård- och mjuk- varaprodukter. Varje rad innehåller ett användar-ID (COOKIE), och information om `Datum`, `Time`, `Click-order`, och `Webpage`. Bilagan innehåller en lista med alla sidor som finns på hemsidan. Utgå från kodmanualen för sekvensanalys. Notera att det går att köra `cSPADE()` i Rstudio.

Sekvensanalys utan tidsbegränsningar

1. Sätt maximala antalet enheter i sekvenserna till 2 och titta på de fem flest förekommande sekvenserna. Försök förklara varför dessa verkar vara vanligt förekommande. Vad är sannolikheten att `Product`-sidan besöks minst två gånger av en användare?
2. Vilka sidor leder till att en användare besöker `Pay_Res`?

Sekvensanalys med tidsbegränsningar

Besvara följande frågor genom sekvensanalyser där ni anger korrekta värden för de tidsbegränsningar som kan styras.

1. Vilka sidor leder **direkt** till `Help`?
2. Vilken är den flest förekommande sidan som en användare besöker efter att ha startat på `Start_Session` och besökt tre andra sidor emellan? Är denna sekvens rimlig?

Bilaga

List of websites that were visited

- Home: the homepage of the website
- Login: where a user has to enter their name and other personal information
- Logpost: prompts a message that informs whether the login has been successful
- Register: to be recognized later on, the user has to obtain a userid and password
- Regpost: shows the partial results of the registration, asking for missing information
- Help: it answers questions that may arise during navigation through the website
- News: presents the most up-to-date products
- Shelf: contains of the programs that can be downloaded from the website
- Program: gives detailed information about the software programs that can be bought
- Download: allows the user to download software programs of interest
- Catalog: contains a complete list of products on sale in the website
- Product: shows detailed information on each product that can be purchased
- P-info: sets out the payment terms for purchasing products on the website
- Addcart: where the virtual basket can be filled with items to be purchased
- Cart: shows the current status of the basket
- Mdfycart: allows the user to modify the current content of the basket
- Pay_req: displays the amount to pay for the products in the basket
- Pay_res: here the visitor agrees to pay, and payment data is inserted
- Freeze: where the requested payment can be suspended, perhaps to add new products to the basket
- Agb: general terms of purchase
- start_session session start
- end_session session end