Datorlaboration 6

Josef Wilzén

21 september 2020

Allmänt

Datorlaborationerna kräver att ni har R och Rstudio installerat.

- Kodmanual: länk
- Dataset till vissa uppgifter finns här.
- **ISL**: An Introduction to Statistical Learning,
  - Allmän info: länk
  - Boken: länk
  - R-kod till labbar: länk
  - Dataset: länk

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 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.

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: Intressemått

Använd här samma material som i övning 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.

(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. Sortera den resulterande tabellen utefter alla intressemått och visa de tio regler med högst värden. Vilka av dessa regler anses vara intressanta?

4. Utforska hur listorna med regler skiljer sig från varandra och försök dra några slutsatser om intressemåttens huvudsakliga egenskaper. Hur skiljer sig asymmetriska och sym- metriska intressemått? Tips: Kolla i denna artikel Selecting the right objective measure for association analysis

Del 3

(Kommer snart)