

Datorlaboration 7

Josef Wilzén

28 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: Webbsideanalys

Datamaterialet "clickstream.csv" innehåller en log-fil från en e-handel webbsida 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.

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