Projekt i 732G12 Data Mining

Josef Wilzén

29 september 2020

1 Lärandemål

Det huvudsakliga målet med denna inlämningsuppgift är att använda den teoretiska och praktiska kunskap som övats upp under tidigare del av kursen. Ni förväntas även få en praktisk övning i hur man kan analysera verkliga datamaterial samt de problem som kan uppstå med dessa. Det ingår även en övning i muntlig och skriftlig redovisning av analysresultatet.

2 Instruktioner

Er uppgift är att i par välja någon av nedanstående material att analysera. Från början kommer det endast tillåtas en grupp per material och metodfamilj, men ifall det tar slut material kommer några få jobba med samma. Det är även tillåtet att välja ett annat material som finns tillgänligt på internet, men då måste godkänade fås. Först till kvarn gäller för dessa val!

När ni väl valt material ska ni komma på en frågeställning som lämpas att besvaras med hjälp utav metodfamiljen som ni valt ut. Det kan vara alltifrån, Vilka egenskaper påverkar huruvida en komponent är trasig?, Vilka sidor besöker en användare innan den landar på Resultat?, Finns det grupper av varor som oftast köps samtidigt? osv.

Under arbetets gång kommer ni säkert stöta på problem som till exempel att datamaterialet inte har det format som ni använt tidigare under kursen eller att en viss tilltänkt metod inte alls fungerar på just det specifika datamaterialet. En del av denna inlämningsuppgift är att ni ska självständigt lösa dessa problem men kan självklart fråga om hjälp under de schemalagda handledningspassen som finns tillhands. Lösningar som ni kommer på, måste tydligt presenteras i rapporten som ni skriver för att uppfylla kravet om reproducerbarhet som råder för akademiska rapporter.

När ni väl kommit fram till ett svar på er frågeställning ska allting sammanställas till en rapport som ska formas enligt rapportmallen. Huvudfokus ska ligga på databeskrivningen och dess bearbetning samt rapportens metodkapitel. Alla analyser och slutsatser ska vara motiverade med lämpliga grafer och tabeller.

Rapporterna ska skrivas med någon följande programvaror:

- Rmarkdown med knitr.
- LaTeX: typsättningssystem som är speciellt lämpligt för vetenskapliga texter och matematisk notation. Valfri programvara för LaTex går bra.
  - Lyx: grafiskt program som generar en LaTeX-rapport i bakgrunden, som kan komplieras till en pdf. Kan användas med knitr.

Rapporten ska lämnas in som pdf-fil, lämna även in er källfil, vilket kan vara Rmd-fil eller lyx/latex-fil beroende på vad ni väljer. Det rekomenderas att ni använder Rmarkdown för rapporterna.

Presentation

Under seminariet kommer varje grupp förfoga över 20 minuter där både presentation och opponering inkluderas. Ni ska under presentationens första 10 minuter sammanfatta den rapport som ni gjort och sedan lämnas 10 minuter för opponering från opponentgruppen.

Opponering

Varje grupp ska opponera på en annan rapport enligt det schema som kommer att presenteras. Det förväntas att fokus ligger på det statistiska, det vill säga hur metoderna presenteras, används och tolkas. Varje grupp ska sammanställa sina kommentarer i ett dokument som sedan ska skickas till rapportgruppen och lärare.

3 Datamaterial

Nedan presenteras förslag på datamaterial. Notera att ni kan välja ett annat material, se rubriken ”Eget datamaterial” nedan. Notera att det är först till kvarn som gäller. Skriv upp ert val på projektlistan som kommer att delas i Teams under kanalen ”#DM_project”. Datamaterialen är indelade efter föreslagen metodklass, men ni kan använda annan metod, men det ska motiveras i rapporten. ’

Klustering

- Växter
  - En databas över 22 632 växter som finns i USA och Canada. Informationen för varje växt som finns är vilken stat eller provins/territorium som den växer i. Källa: https://archive.ics.uci.edu/ml/datasets/Plants
  - Klustering
- USAs befolkning
  - En databas över 2 458 285 slumpmässigt utvalda individer från 1990 års folkräkning i USA. Källa: https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29
  - Klustering
- Köpcentrum (Gifts)
  - Transaktionsdatabas över en UK-baserad affär. Källa https://archive.ics.uci.edu/ml/datasets/Online+Retail
  - Klustering

Klassificering

- Bilsensorer
  - 11 sensorer i en bil har samlats in data för att benämna ifall en komponent är trasig eller ej. Tyvärr finns ingen information kring vad de olika variablerna betyder exakt. Källa: https://archive.ics.uci.edu/ml/datasets/Dataset+for+Sensorless+Drive+Diagnosis
- Bokstavsigenkänning
  - Datamaterial som behandlar en 20 000 bokstäver med variabler som beskriver hur dessa ser ut. Källa: https://archive.ics.uci.edu/ml/datasets/Letter+Recognition

Associations och sekvensanalys

- FIFA
  - Ett datamaterial taget från FIFA World Cup 98s hemsida med 20 450 unika sekvenser av klickningar på 2 990 distinkt olika hemsidor. Tyvärr finns ingen information om vilka sidorna egentligen är. Källa: http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php
- Köpcentrum (Retail)
  - Data från en anonym belgisk affär LÄS IGENOM OCH ANVÄND RAPPORTMALLEN! med 88 162 transaktioner och ett antal unika varor eller varugrupper Källa: http://fimi.ua.ac.be/data/ (Tom Brijs, notera källan i länken)
- Växter
  - Samma som ovan
- Köpcentrum (Gifts)
  - Samma som ovan

Eget datamaterial

Ni är fria att välja ett eget datamaterial. Då gäller följande regler:

- Inget simulerat dataset eller ”toy data”. Det ska vara ett riktig data, som kan användas för en riktig frågeställning.
- Inte för ”enkelt”: inte för några observationer eller variabler
- När ni hittat ett datamaterial: Fråga Josef om det är ok att använda det. Ge en kort beskrivning av det och vilken metodklass ni tänker er.

Förslag på ställen att hitta data:

- Machine Learning Repository
- Kaggle datasets
- Datasets for Data Mining, Data Science, and Machine Learning