Projekt i 732G12 Data Mining

Josef Wilzén

5 oktober 2022

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 ett datamaterial som ni ska analysera. Se sektion 3 för detaljer. När ni väl valt datamaterial ska ni komma på en frågeställning som kan besvaras genom att analysera det valda datamaterialet. Exempel på frågeställningar:

- Vilka egenskaper påverkar huruvida en komponent är trasig?
- Vilken metod predikterar temperaturen bäst med avseende på MSE och MAE?
- Vilka sidor besöker en användare innan den landar på sidan ”Resultat”?
- Finns det grupper av varor som oftast köps samtidigt?

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 ni 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. Rapportmallen finns här och här (se även kurshemsidan), och innehåller instruktioner om hur ni ska skriva er rapport. 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. Det rekomenderas att ni använder Rmarkdown för rapporten. Döp filen på formen ”gruppX_liuid1_liuid2.pdf” och ladda upp på Lisam i den anvisade samarbetsytan innan deadline. Deadlines finns sammaställda i ett dokumment [här].

Datainlämning

Ni ska göra en mindre inlämning (på Lisam) innan ni lämnar in den färdiga rapporten. Där ska ni:

- Beskriva vilket datamaterial som ni har valt
  - vilka variabler, antal variabler, antal obs mm
  - kortfattad explorativ analys: kortfattad beskrivande statistik av data och/eller några plottar av data
- Ange preliminär frågeställning (ok att ändra senare vid behov)

Inlämingen ska vara en pdf-fil som är 1-3 sidor lång. Syftet är att ni ska välja data och komma igång med inledande datahanteringen, och börja fundera över frågeställningen. Det är ok att återanvända hela eller delar av denna inlämning till den slutgiltiga rapporten om man vill.

Presentation

Under seminariet kommer varje grupp förfoga över 25 minuter där både presentation och opponering inkluderas. Ni ska under presentationens första 15 minuter presentera och 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.

- Vid den muntliga opponeringen så ska de större konceptuella frågorna och kommentarerna tas upp. Börja med de viktigaste.
- Mindre kommentarer och saker som rör formalia tas bara upp skriftligt.

Varje grupp ska sammanställa sina kommentarer i ett dokument som sedan ska skickas till rapportgruppen och lärare. Detta dokument ska innehålla både de små och stora kommentarerna.

3 Datamaterial

Er uppgift är att i grupper om två välja något datamaterial att analysera. Endast en grupp tillåtas per datamaterial. Först till kvarn gäller för dessa val!

Skriv upp ert val på projektlistan som kommer att delas i Teams under kanalen ”#DM_project”. Ni ska citera källan på datamaterialet i de fall då det krävs i er rapport.

Välja datamaterial

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

- Inget simulerat datamaterial 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, tumregel: antingen antal obs ≥ 500 eller antal variabler ≥ 10. Fråga om ni är osäkra.
- 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
- List of datasets for machine-learning research
- De databaser som finns tillgänliga via pxweb, se också här och här.

Några förslag på datamaterial

- SkillCraft1
  - https://archive.ics.uci.edu/ml/datasets/SkillCraft1+Master+Table+Dataset
  - Förslag: Klassificering
- Parkinsons
  - https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring
  - Förslag: Regression
- 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
  - Förslag: Klustering, Klassificering
- 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/Pla
  - Förslag: Klustering, Associations och sekvensanalys
- Wine Quality Data Set
  - https://archive.ics.uci.edu/ml/datasets/Wine+Quality
  - Förslag: Regression, Klassificering
- Adult Data Set
  - https://archive.ics.uci.edu/ml/datasets/Adult
  - Förslag: Klassificering
- 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
  - Förslag: Klassificering