Projekt i 732G12 Data Mining

Josef Wilzén

September 24, 2024

1 Lärandemål

Det huvudsakliga målet med denna inlämningsuppgift är att använda den teoretiska och praktiska kunskap som övats upp under första delen 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 har valt ett datamaterial ska ni komma på en frågeställning som kan besvaras genom att analysera det valda datamaterialet. Det huvudsakliga problemet i projektet ska behandla övervakad inlärning. Det är dock tillåtet att använda oövervakad inlärning som del i att anlysera det huvudsakliga problemet. I analysen ska minst en modell vara ett neuralt nätverk.

Exempel på frågeställningar:

- Vilka egenskaper påverkar hurvida en komponent är trasig?
- Vilken modell ger bäst predikation av framtida inkomst?
- Vilken metod predikterar temperaturen bäst med avseende på MSE och MAE?

Under arbetets gång kommer ni säkert stöta på problem som till exempel att datamaterialet inte har det format som ni använt tidigare eller att en viss tilltänkt metod inte fungerar alls på det specifika datamaterialet. En del av denna inlämningsuppgift är att ni självständigt ska lösa dessa problem, men ni kan självklart ställa frågor under de schemalagda undervisningspassen. 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 frågeställningen ska allting sammanställas i 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 ligger på databeskrivningen och dess bearbetning samt rapportens metodkapitel. Alla analyser och slutsatser ska vara motiverade med lämpliga gradet och tabeller.

Rapporten ska skrivas med någon av följande programvaror:

- Rmarkdown[1] (med knitr) Rekomenderas för detta projekt!
- LaTeX Vanligt vid skrivandet av akademiska rapporter. Kan använda Overleaf för att skriva tillsammans (utan strul med installationer).
- LyX Grafiskt program som generar en LaTeX-rapport i bakgrunden. Kan användas med knitr för att köra R-kod.

Om ni använder knitr med Rmarkdown/latex/lyx, då rekomenderas det att ni har seprata R-filer där ni har era anayser, och att ni sen bara läser in lämpliga resultat i er rapport-fil. Så man inte måste köra om alla skattningar etc varje gånga som man ska kompliera sin rapport-fil.

Rapporten ska lämnas in som en pdf-fil. Döp filen på formen `gruppX_liuid1_liuid2.pdf` och ladda upp på Lisam. Er rapport ska vara snygg och välstruktuerad. Se rapportmallen för mer instruktioner.

[1] Det går även att använda Quarto om man vill.

Tidigare studier

En del datamaterial har analyserats av andra och olika analyser kan finnas tillgängliga på internet. Det är inte ok att direkt kopiera någon annans analys av samma datamaterial. Det är ok att använda vissa avgränsade delar av andras analyser om tydligt anger vad man använt och citerar källan. Exempel kan vara att man använder samma transformation av förklarande varaiabler som i en annan analys. Man kan hämta inspiration av vilka metoder som har använts eller inte använts på just ert datamaterial. Om ni läser om andra analyser av ert datamaterial så bör ni skriva kort om dessa i er bakgrundsektion i rapporten (och citera!), alltså vad som har gjorts tidigare på "området".

Det är också ok att jämföra era egna resultat med resultatet från någon annans analys (glöm inte att citera då). Exempel: Ni har valt att använda neurala nätverk, ni ser att någon annan har använt random forest på samma data. Då kan ni i diskussionen citera den andra källan och jämföra era träffsäkerhet med deras träffsäkerhet. Det är inte ok att ta någon annas resultat från modellskattningar som ert eget resultat, utan ni måste skatta alla era egna modeller själv.

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 observationer osv.
  - Vilken variabel är er responsvariabel?
  - kortfattad explorativ analys: kortfattad beskrivande statistik av data och/eller några plottar av data.
- Ange preliminär frågeställning (ok att ändra senare vid behov).

Inlämningen ska vara en pdf-fil som är 1-3 sidor lång. Syftet är att ni ska välja data och komma igång med inledande datahantering, 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 det ges 25 minuter till varje grupp. Under de första 15 minuterna ska ni presentera och sammanfatta den rapport som ni gjort, övriga 10 minuter lämnas för opponering från opponentsgruppen.

Opponering

Varje grupp ska opponera på en annan grupp. Det förväntas att fokus ligger på det statistika, det vill säga hur metoderna presenteras, används och tolkas.

- Vid den muntliga opponeringen så ska de större konceptuella frågorna oh kommentarerna tas upp.
- 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

Varje grupp ska välja ett eget datamaterial. Två grupper kan inte ha samma datamaterial och först till kvarn gäller. På Lisam under sammarbetsyta finns ett excel ark där ni kan skriva upp vilket datamaterial ni valt. Kom ihåg att citera källan på ert datamaterial i er rapport.

Välja datamaterial

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

- Inget datamaterial som ni har arbetet med under datorlaborationerna under kursen.
- Inget simulerat datamaterial eller "toy data". Det ska vara ett riktigt datamaterial som kan användas för en riktig frågeställning.
- Inte för "enkelt": Tumregel, minst 500 observationer eller minst 10 variabler. Fråga om ni är osäkra.
- När ni hittat ett datamaterial, skriv upp det i dokumentet. Fråga Johan om ni är osäkra på valet.

- Problemet ska vara inom **övervakad inlärning** (kan vara antingen regression eller klassificering).

Ni väljer själva var ni vill hämta data ifrån. Här kommer några förslag på datakällor:

- Machine Learning Repository
- Kaggle datasets
- MedMNIST
- Datasets for Data Mining, Data Science, and Machine Learning
- Awesome Public Datasets
- List of datasets for machine-learning research
- Kan också använda databaser från pxweb, se också här och här.