Kunskapskontroll 2

Machine Learning



Alvin Lagerson

EC Utbildning

202503

# Abstract

En kort sammanfattning över ditt arbete och de viktigaste resultaten skrivet på engelska, cirka 5 meningar totalt.

**Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.**

Innehållsförteckning

[1 Inledning 1](#_Toc160654340)

[1.1 Underrubrik – Exempel 1](#_Toc160654341)

[2 Teori 2](#_Toc160654342)

[2.1 Exempel: Regressionsmodeller 2](#_Toc160654343)

[2.1.1 Exempel: Lasso 2](#_Toc160654344)

[2.1.2 Exempel: Ridge 2](#_Toc160654345)

[2.1.3 Exempel: Elastic Net 2](#_Toc160654346)

[2.2 Exempel: Neurala Nätverk 2](#_Toc160654347)

[3 Metod 3](#_Toc160654348)

[4 Resultat och Diskussion 4](#_Toc160654349)

[5 Slutsatser 5](#_Toc160654350)

[6 Teoretiska frågor 6](#_Toc160654351)

[7 Självutvärdering 7](#_Toc160654352)

[Appendix A 8](#_Toc160654353)

[Källförteckning 9](#_Toc160654354)

# Inledning

Ge en bred bakgrund till ditt arbete och varför det är relevant. Smalna successivt av och koppla bakgrunden till detta arbete. Fortsätt med syfte och frågeställning (syfte och frågeställning kan du ha i en egen underrubrik om du önskar det).

Syftet med denna rapport är \_\_\_\_\_\_\_\_\_\_\_\_, för att uppfylla syftet så kommer följande frågeställning(ar) att besvaras:

1. \_\_\_\_\_\_\_
2. \_\_\_\_\_\_\_

## Underrubrik – Exempel

Text är skriven på formatet Calibri med textstorlek 11.

# Teori

## Exempel: Regressionsmodeller

För att infoga ekvationer, gå till Infoga > Ekvation. Där kan du också välja att skriva egna ekvationer.

### Exempel: Lasso

#### Regularisering i Lasso

#### Välja Hyperparameter

### Exempel: Ridge

### Exempel: Elastic Net

## Exempel: Neurala Nätverk

# Metod

Hur har du genomfört ditt arbete? Exempelvis, hur har datan erhållits?

# Resultat och Diskussion

|  |  |
| --- | --- |
| **RMSE för olika modeller** | |
| Enkel Linjär Regression | xx |
| Lasso | xx |
| Ridge | xx |

Tabell 1: Root Mean Squared Error (RMSE) för de fyra valda modellerna.

En bild som visar text, Teckensnitt, skärmbild, linje

Automatiskt genererad beskrivning

Figur 1: Hur man lägger in tabell eller figur nummer samt beskrivning.

# Slutsatser

Här besvarar du bl.a. frågeställningarna.

# Teoretiska frågor

*Kalle delar upp sin data i ”Träning”, ”Validering” och ”Test”, vad används respektive del för?*

Träningsdata använd för att för att träna modeller, valideringsdata används för att utvärdera vilken av modellerna som är bäst. Sen tränar man om den bästa modellen på träning och valideringsdata för att sen testa den på testdata för att se hur bra modellen generaliserar på ny osedd data.

*Julia delar upp sin data i träning och test. På träningsdatan så tränar hon tre modeller; ”Linjär Regression”, ”Lasso regression” och en ”Random Forest modell”. Hur skall hon välja vilken av de tre modellerna hon skall fortsätta använda när hon inte skapat ett explicit ”validerings-dataset”?*

Hon kan göra k-fold Cross-validation. Då delas träningsdatan upp i cluster där 4 kluster används för träning och valideras mot 1 för att sen tränas på 4 andra och valideras mot 1 till dess alla kluster använts till validering. Om det är mycket data hon jobbar med så kan det ta tid, men i annat fall är det ett bra sätt att utnyttja sin data.

*Vad är ”regressionsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden?*

Ett regressionsproblem är när den beroende variabeln (y) kan anta många värden. T.ex. om y är inkomst eller ett bostadspris som kan anta kontinuerliga värden (oändligt många olika värden). Några exempel på modeller är Linear regression och Lasso som kan användas för att t.ex. förutsäga bostadspriser. Men du kan även använda SVM, Support Vector Machine, för att förutsäga bostadspriser.

*Hur kan du tolka RMSE och vad används det till?*

Root mean squared error är ett mått på medelavståndet mellan uppskattade värden och verkliga värden. Anledningen till att man kvadrerar och tar roten ur är att det inte har någon betydelse om avståndet är negativt eller positivt, det är storleken på felet som är intressant. Du kan använda RMSE för att jämföra olika modeller med varandra för att hitta den bästa modellen. Du behöver lära känna den data du jobbar med för att förstå vad som är ”bra” eller ”dåligt” eftersom talet du får fram är helt beroende av vad du tittar på. Lägre är bättre, men att få 20 i RMSE kan i sitt sammanhang vara riktigt uselt medan 20 000 i ett annat sammanhang kan vara fantastiskt, siffran i sig betyder inget utan sitt sammanhang.

*Vad är ”klassificieringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?*

Den beroende variabeln (y) kan anta två olika värden. T.ex. churn, 1 eller 0. Kan ha en eller flera andra oberoende variabler. Ett exempel på modell är logistisk regression som skulle kunna användas t.ex. för att avgöra om en patient har en viss sjukdom. Decision tree eller Random Forest är fler exempel på modeller du kan använda vid klassificeringsproblem. Decision tree kan t.ex. användas för att detektera skräppost och Random forest skulle kunna användas för att utvärdera kunder med hög kreditrisk. SVM, Support Vector Machines kan även användas vid klassificeringsproblem som i ett spamfilter. En Confusion Matrix är en matris för att grafiskt visa resultatet av en modells prestation, en enkel variant är en matris som visar verklig sant/falskt som rader och förutspådd sant/falskt som kolumner. Då visas korrekt sant/ inkorrekt falskt på första raden och inkorrekt sant / korrekt falskt på andra raden.

*Vad är K-means modellen för något? Ge ett exempel på vad det kan tillämpas på.*

Det är en oövervakad inlärningsmodell. Det betyder att det inte finns någon beroende variabel, något y-värde. Algoritmen används för att gruppera datapunkter i olika kluster (k-antal kluster). Det fungerar som så att man väljer antal kluster, de placeras ut i datarummet, varje datapunkt tilldelas närmaste centralpunkt, centroid, och sen beräknar man medelvärdet av datapunkterna i varje kluster och flyttar centroiden dit. Sen upprepar man de två sista stegen tills centroiderna stabiliseras. Ett exempel på användning är att placera kunder i olika segment för att kunna rikta relevant marknadsföring till respektive kundgrupp eller att upptäcka avvikelser, eller anomalier, i beteenden hos kundgrupper för att till exempel upptäcka potentiella brott.

*Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding. Se mappen ”l8” på GitHub om du behöver repetition.*

Ordinal encoding används när kategorierna har en naturlig ordning (t.ex. en ordinal variabel). Varje kategori tilldelas ett unikt numeriskt värde. Ett exempel är kundnöjdhet, 1 till 5 från missnöjd till mycket nöjd, men en tvåa inte nödvändigtvis är 50% av en fyra. En fördel är att det är ordning mellan kategorierna men en nackdel är att skillnaden mellan värdena inte är jämn.

One-hot encoding skapar en binär (0 eller 1) kolumn för varje kategori. Används för nominella variabler (ingen inbördes ordning). Ett exempel kan vara frukt där man ställer upp frukttyperna med svaren 1 eller 0, där ett äpple är 1 under kategorin äpple och då 0 på apelsin och 0 på mango etc. En fördel är att det inte är någon risk att modellen antar en ordning som inte finns. En nackdel är att det lätt kan bli många kolumner om det finns många unika värden (s.k. "curse of dimensionality").

Dummy encoding är en variant av one-hot encoding där en kategori utelämnas (referenskategori), för att undvika multikollinearitet, där variablerna kan förutsägas av varandra. Som i fruktexemplet kan man då ta bort kategorin äpple eftersom om det är 0 på apelsin och 0 på mango etc så betyder det att det är ett äpple. En fördel är att det blir färre kolumner än one-hot encoding, vilket minskar redundans. En nackdel är att det kräver att vi tolkar modellen med referenskategorin i åtanke.

*Göran påstår att datan antingen är ”ordinal” eller ”nominal”. Julia säger att detta måste tolkas. Hon ger ett exempel med att färger såsom {röd, grön, blå} generellt sett inte har någon inbördes ordning (nominal) men om du har en röd skjorta så är du vackrast på festen (ordinal) – vem har rätt?*

Data kan inte vara både ordinal och nominal samtidigt. Att en person med röd skjorta skulle vara vackrast på en fest är helt och hållet subjektivt och Julias personliga preferens. Om man räknar med personliga preferenser så kan i princip all data räknas som ordinal eftersom Julia även kanske tycker äpple är godare än apelsin. Detta skulle leda till att ingen modell kan användas för generalisering någonsin eftersom de subjektiva värderingarna av en mängd olika värden enbart passar Julia. Ska data kunna rangordnas och vara ordinal måste det vara objektivt sett en inbördes ordning. Ett dataset kan däremot innehålla både ordinal och nominal data.

*Kolla följande video om Streamlit: https://www.youtube.com/watch?v=ggDa-RzPP7A&list=PLgzaMbMPEHEx9Als3F3sKKXexWnyEKH45&index=12 Och besvara följande fråga: - Vad är Streamlit för något och vad kan det användas till?*

Streamlit är ett open source Python-bibliotek som används för att snabbt och enkelt bygga interaktiva webbapplikationer för dataanalys och maskininlärning. Streamlit kan bland annat användas för att visa data och låta användaren interagera med data direkt i webbläsaren utan att behöva bygga applikationen i till exempel HTML eller att helt enkelt skapa ett enkelt gränssnitt för att testa ML-modeller.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Under arbetets gång har jag stött på några problem. Bland annat fick min random forest en accuracy på knappt 8%, jag provade då att öka antal träd men det gjorde väldigt liten skillnad. Jag läste om att man kunde använda XGBoost istället och testade köra den istället men fick i princip samma accuracy. Jag testade också att göra ett enkelt neuralt nätverk, men jag kunde inte få Tensorflow att fungera på min dator. Jag fick igång det på Google Colab och fick då en accuracy på 8% och insåg då att jag måste ha fel i koden. I Colabs såg jag i loggen att accuracyn låg mellan 65% och 96% i varje epoch men att den till slut skrev ut 7% och insåg då att jag utvärderade X-valideringsdata mot y-testdata.

5000 samples i träning, 1000 i validering

En bild som visar text, Teckensnitt, skärmbild, Grafik

AI-genererat innehåll kan vara felaktigt.

50000 samples, 10000 i validering

En bild som visar text, Teckensnitt, skärmbild, Grafik

AI-genererat innehåll kan vara felaktigt.

5000 samples, 1000 i validering neuralt nätverk

En bild som visar text, skärmbild, Teckensnitt, linje

AI-genererat innehåll kan vara felaktigt.

50000 samples, 10000 i validering neuralt nätverk

En bild som visar text, skärmbild, Teckensnitt, linje

AI-genererat innehåll kan vara felaktigt.

1. Vilket betyg du anser att du skall ha och varför.
2. Något du vill lyfta fram till Antonio?

# Appendix A

# Källförteckning