Kunskapskontroll Sandra Persson

Teoretiska Frågor

*Hur kan vi definiera ”Maskininlärning”?*

Att bygga system som kan lära sig av data utan att bli explicit programmerade. The art and science. Bli bättre.

*Vad är supervised learning?*

Då vi har data med ”labels”, alltså data där vi vet vad den beroende variabeln är. Uppgiften är då att förutspå vilken label den nya datan kommer att ha. Datan vi ger till algoritmen har ”labels”, den innehåller alltså ett facit. Vi vet vad svaret är. Vill bygga en modell som kan ge ny data en korrekt label.

*Vad är skillnaden mellan Regressionsproblem och Klassificieringsproblem?*

Definitionen enligt boken är att regressionsproblem är ett problem där vi förutspår data som är kontinuerlig. Jag tycker personligen att den definitionen är förvirrande och föredrar att tänka på det som att ”Ett tal som inte behöver vara ett heltal”. Ratio, equal interval. Data där intervallerna mellan varje möjlig datapunkt är jämn.

*Kan du ge ett exempel på vad regressionsmodeller respektive klassificeringsmodeller kan*

*användas till?*

En regressionsmodell skulle kunna användas till att förutspå

Hur många kunder, vad en kund kommer att spendera, temperatur

En klassificeringsmodell skulle kunna användas för att tex förutspå vilken kategori en person tänker på, djur, byggnad eller transportmedel. Höger, vänster, upp, ner.

*Vad är Root Mean Squared Error (RMSE)?*

RMSE används som ett mått för att kunna utvärdera hur bra en regressionmodell är (en av flera olika typer av mått), alltså hur bra den har lyckats att förutspå ny data. För att utvärdera modellen. För att få en intuitiv förståelse för RMSE skulle man kunna tänka på det som ett mått som visar medelvärdet för hur mycket fel modellen gör (eller egentligen standardavvikelsen för felen). Euclidean norm. RMSE. Roten ur gör att det blir lättare för oss människor att tolka värdet.

Loss function (fast bara MSE är loss function för OLS), evaluation, ett mått på hur bra modellen är. Använd inom regression för att hitta den bästa ”linjen”, alltså den modell som bäst predicerar datan. Hitta de bästa koefficienterna.

Medelvärdet av de kvadrerade felen, alltså skillnaden mellan det faktiska värdet och det predicerade värdet. Roten ur.

*Vad är en ”confusion matrix”?*

En matris som visar en frekvenstabell över predicerade värden jämfört med de faktiska värdena. För en klassificerare. Används för att utvärdera modellen. Kan få ut flera intressanta mått från matrisen, såsom accuracy, precsion, recall och False Positive Rate.

*Om man delar upp datan i träning, validering och test – hur används respektive del?*

Träningsdelen används för att träna modeller, fit metoden.

Valideringsdatan används för att välja modell eller hyperparametrar.

Och testdatan används i slutet för att testa vår valda modell på ny data.

*Vad är en parameter för något? Ge ett exempel.*

*Vad är en hyperparameter för något? Ge ett exempel.*

*Vad är GridSearchCV i Scikit-learn? Ge ett exempel på vad det kan användas till.*

2. Rapport

En ”grov mall” för hur rapporten skall vara strukturerad kan ni se nedan.

Introduktion som innehåller underrubrikerna: - Bakgrund

- Syfte och Frågeställning

Databeskrivning / EDA (Exploratory Data Analysis)

Metod och Modeller (Teori)

Projekt Resultat och Analys

- Resultat är deskriptiva i sin natur, t.ex. att man presenterar RMSE för sina olika modeller. Ofta kan tabeller vara användbara vid redogörelse av resultaten.

1

5. Slutsats och förslag på potentiell vidareutveckling.  
**Rapporten skall vara ca 2-3 sidor**. Skriv koncist och fundera på vad du vill lyfta fram.

**Efter att du är klar skall du skriva en kort redogörelse i slutet av rapporten (detta ingår inte i de 2-3 sidorna):**

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

**Vilket betyg du anser att du skall ha och varför.**

**Tips du hade ”gett till dig själv” i början av kursen nu när du slutfört den.**

**Introduktion**

**Bakgrund**

Dataset housing, från Handson ML, från början ett dataset från StatLib. California Housing Prices dataset. 1990.

Från kartan i boken ser det ut som att plats för huset verkar ha betydelse, men frågan är om latitud och longitud är bra prediktorer, eller om närhet till havet och population kanske är bättre då det ser ut som att de dyraste husen finns nära havet i städer/tätbefolkade orter.

”For each block group in California. Block groups are the smallest geographical unit for which the US Census Bureau publishes sample data. A block group typically has a population of 600 to 3000 people. Distrikt.”

**Syfte och frågeställning**

Syftet med analysen är att bygga en modell som kan predicera medianhuspriser utifrån de andra variablerna i datasetet.

**Databeskrivning/EDA**

Innan jag delar upp datan i tränings och testdata gör jag en kortare EDA för att undersöka vilka slags variabler datan har, kontrollera om det finns saknad data, att värdena ser rimliga ut och hur distributionen ser ut för varje variabel.

10 variabler, 20 640 observationer.

9 numeriska och 1 kategorisk.

Longitude: interval scale

Latitude: interval scale

housing\_median\_age: ratio, discrete

total\_rooms: ratio, discrete

total\_bedrooms: ratio, discrete

population: ratio, discrete

households: ratio, discrete

median\_income: ratio

median\_house\_value: median house value for households within a block, in US dollars.

ocean\_proximity: kategorisk variabel med nominalskala. Fem kategorier, värt att notera är att bara fem stycken distrikt har värdet ”island”.

Variabeln total bedrooms har saknade värden som jag behöver titta på innan jag matar datan till mina modeller.

Medin house value: när vi tittar på frekvens finns flest antal, närmare 1.000 hos max-värdet. Alltså mode är samma som max-värdet.

Korrelationer: latitud och longitud har en negativ korrelation. Vi ser även att population, antal rum, antal sovrum, och hushåll alla visar en positiv korrelation. Och även att median house value visar en positiv korrelation med median income. Och att median house value inte visar ett linjärt samband med någon annan variabel än medianinkomst.

**Metod och modeller**

**Preprocessing**

Total bedrooms har saknade värden. Jag använder simple imputer för att ersätta de saknade värdena och väljer att använda medianen eftersom vi sett att distributionen ser skewed ut.

**Resultat och analys**

**Slutsats och förslag på potentiell vidareutveckling**

Efter att du är klar skall du skriva en kort redogörelse i slutet av rapporten (detta ingår inte i de 2-3 sidorna):

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

Den största utmaningen var nog att jag kände mig betydligt mindre motiverad än jag brukar. Jag tycker att det är kul att svara på teoretiska frågor och att upptäcka och analysera ett nytt dataset. Men den här gången hade vi redan gått igenom alla frågor på lektionen och dessutom diskuterat frågorna på tidigare lektioner. Datasetet har vi också tittat på under tidigare lektioner och dessutom är det ett dataset som återkommer i vår kursbok.

**Vilket betyg du anser att du skall ha och varför.**

**Tips du hade ”gett till dig själv” i början av kursen nu när du slutfört den.**