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**

(Ett klassiskt data set för att lära sig ML. Syftet är att jag ska lära mig. Eller sälja in hur bra det hade varit att kunna bygga en modell som predicerar huspriser per distrikt.)

Datasetet som används i analysen är “housing” från boken *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems.* Datasetet är från början från 1990 och visar median-huspriser angivet i US dollar för varje distrikt, eller ”block”, i Kalifornien.

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 rapporten är att bygga en modell som kan predicera median-huspriser 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, och för att kontrollera om det finns saknad data, att värdena ser rimliga ut, samt hur distributionen ser ut för varje variabel.

Datasetet har 10 variabler, varav 9 numeriska och en kategorisk variabel. Se nedan tabell för en sammanfattning över variablerna, samt deras medelvärde och standardavvikelse. Variabelnamnen är beskrivande men värt att minnas är att varje observation gäller för ett distrikt i Kalifornien. Antal observationer, alltså antal distrikt, är 20 640.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variabelnamn** | **Mätskala** | **Mean** | **Std** |
| longitude | Interval scale |  |  |
| latitude | Interval scale |  |  |
| housing\_median\_age | ratio |  |  |
| total\_rooms | ratio |  |  |
| total\_bedrooms | ratio |  |  |
| population | ratio |  |  |
| households | ratio |  |  |
| median\_income | ratio |  |  |
| median\_house\_value | ratio |  |  |
| ocean\_proximity | nominal |  |  |

EDA:n visar att variabeln total bedrooms har saknade värden, vilket jag kommer att behöva behandla under preprocessing. Plottar av distributionen för varje variabel visar att vår beroende variabel, median house value, har samma mode som max värde, det vanligaste värdet är alltså det högsta värdet. De visar även att flera variabler har en skewed distribution samt flera outliers.

Innan jag tittar på samband mellan variabler, delar jag upp datan i ett tränings-set och ett test-set för att undvika att ”tjuvkika” på mönster i testdatan. En plot över korrelationen mellan de numeriska variablerna i träningsdatan visar att den beroende variabeln bara visar ett linjärt samband med median-inkomst. Den visar även att flera av de oberoende variablerna visar en stark korrelation med varandra vilket innebär att en variabel i varje par antagligen är öveflödig, eftersom den inte tillför någon information. Till exempel visar antal rum och antal sovrum visar en stark positiv korrelation. Latitud och longitud visar en stark negativ korrelation (vad innebär det egentligen?), vilket helt enkelt beror på formen av Kalifornien. Vår kursbok skapar nya variabler genom att,,, vilket käns som en bra idé och jag väljer att göra samma sak.Korrelationerna ger en indikation på att inte alla features är relevanta som prediktorer för huspriset och jag väljer därför att göra en feature selection med hjälp av en random forest regressor och feature importance.

Innan jag går vidare med modellval skapar jag en pipeline för preprocessing. Saknade värden i numeriska variabler ersätter jag med medianen. Anledningen till att jag väljer median som strategi är att den variabel som har saknade värden, total bedrooms, visar en skewed distribution. De numeriska variablerna skalar jag med hjälp av robust scaler, som skalar datan genom att subtrahera median och sedan dela med IQR, vilket gör den mer robust mot outliers jämfört med standard scaler. Den kategoriska variabeln gör jag om till en numerisk med hjälp av One-hot-encoding. Två varianter av träningsdatan skapas, en som använder dummy variable encoding för att användas i linjär regression, och en vanlig one-hot-encoding för övriga modeller.

Features

10 variabler, 20 640 observationer.

9 numeriska och 1 kategorisk.

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.

Median house value: när vi tittar på frekvens finns flest antal, närmare 1.000 hos 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.**