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)?*

Man skulle kunna tänka på det som ett medelvärde för hur många fel modellen gör (eller egentligen är det ett medelvärde av de kvadrerade felen).

Loss function, 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.

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,

Syfte och frågeställning

Databeskrivning/EDA

Metod och modeller

Resultat och analys

Slutsats och förslag på potentiell vidareutveckling