**Teoretiska frågor:**

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

Ans: Träning data : Dessa data används för att träna modellen. Modeller lär sig av data och gör förutsägelser. Det recommenderas att träna modellen med en stor andel av total data för noggrannhet och förutsägelser.Beroende på data och dess egenskapar använder vi olika modeller.

Validering data : Dessa data används för välja den bästa modellen från de olika modeller som testats under träningen.Det hjälper till att modellen inte är underpassad eller överanpassas till träningsdata.

Testdata : När modellen är valts används testdata för att kontroleera prestanda och pöang genom att göra opartiska förutsägelser. Testdata hjälper oss att förstå hur vår valda modell kommer att fungera på osynliga data.

1. 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”?

Ans:Julia kan forsätta med Korsvalidation för alla modeller. Baserad på resultan hon kan välja modell med mindre RMSE(Root Mean Square Error).

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

Ans: Regressionsproblem handlar om att arbeta med numeriska värden. Vi ger numeriska indata och förväntar oss att den modellen förutsäger numeriska utdata.

I vissa fall omvandlar vi vissa kategoriska värden till numeriska genom att använda vissa tekniker som "Label Encoding", skapa "Dummy Variables", "One Hot Encoder" och etc.

Exempel:

1. Förutsäga huspriser baserat på storlek och ortsnamn.

2. Förutsäga anställningsbeslutet baserat på erfarenhet, testresultat i intervju och kvalifikationer.

3. Förutsäga bilpriser baserat på bilens ålder, körsträcka och antal tidigare ägare.

4. Försäljningsförutsägelser genom att använda tidigare data baserade på månad, kvartal och år

1. Hur kan du tolka RMSE och vad används det till:

Ans: Mean square Error(MSE) mäter det genomsnittliga avståndet mellan de förutsagda värdena och de verkliga värdena. Eftersom de verkliga värdena kan vara över eller under de förutsagda värdena, kan skillnaderna vara både positiva och negativa. För att undvika negativa värden kvadrerar vi dessa skillnader.

För att återföra det till de ursprungliga enheterna tar vi roten ur MSE,vilket ger oss RMSE

RMSE används för att bedöma fel i modeller som förutspår kontinuerliga värden.

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

Ans: Klassificeringsproblem innebär att tilldela indata till fördefinerade kategorier och även förutsäga i vilken kategori nya punktar passar in.

Exemple:

1.Förutsäga hand skrivna siffror eller arbeta med siffror from sklearn datasets med hjälp av Random Forest Model.

2.Churn förutsäga med hjälp av SVM.

3.Spam detektering för e-post.

Confusion Matrix : Denna matrix används för att utvärdera prestanden hos en klassificeringsmodell. Detta har 4 utfall:

1.True Positive - Modellen förutspår korrect när det är sant.

2.True Negative – Modellen förutspår korrect när det är falskt.

3.False Positive – Modellen förutspår sant när det egentligen är falsk..

4.False Negative – Modellen förutspår falsk när det egentligen är sant.

Från confusion matrix vi kan defenera följande mått:

1.Presicion : Andelen positiva förutsägelser som faktiskt är korrekta.

2.Recall : Andelen av den possitiva klassen som vi förutspår korrekt.

3.F1 score : Harmoniska medel värde av presicion och recall.

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

Ans: K-means är en algoritm som används för unsupervised learning för att gruppera datapunkter i olika kluster. Det korrekta antalet kluster väljs genom att använda Elbow-metoden.

Vi väljer slumpmässigt centroids och grupperar datapunkterna nära dem, sedan justerar vi dessa centroids tills datapunkterna inte ändrar sitt kluster.

Huvudmålet är att få datapunkterna inom ett kluster att hålla sig nära centroiden samtidigt som de är långt ifrån andra kluster.

Exempel: 1.Vi kan ta irisflower dataset och försöka bilda kluster med deras egenskaper.

2.Vi kan klustra människor baserat på deras ålder och inkomst med hjälp av k- means-modellen.

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

Ans: Ordinal Encoding: Det innebär att kategoriska data som inneboende rangordning.När vi arbeta med ordinal data (exemple "Tillfredsställd", "Inte tillfredsställd", "Neutral"), kan vi använda denna kodning och tilldela värden baserat på deras ordning.

Exemple: När vi diskutera om utbildningnivår ordinal encoding hjäler oss att tilldela siffror till varje utbildningsnivå.

1. Graduation : 0
2. Masters : 1
3. phd : 2

data = [‘Graduation’,’phd’,’phd’,’Masters’,’Graduation’,’Phd’]

Output = [0 2 2 1 0 2]

One-hot-encoding: Denna metod används för kategoriska data utan någon ordning (nominaldata). One-hot encoding används för att omvandla kategoriska variabler till binära värden. För varje observation skapas en ny kolumn som motsvarar en kategori, där den aktuella kategorin får värde 1 och alla andra kategorier får värdet 0.

Efter omvandlingen, kan vi ta bort den ursprungliga kategoriska kolumnen och behålla de nya binära kolumnerna (en per kategori).

|  |  |  |  |
| --- | --- | --- | --- |
| city | Goteborg | Malmo | Örebro |
| Goteborg | 1 | 0 | 0 |
| Malmo | 0 | 1 | 0 |
| örebro | 0 | 0 | 1 |

Exemple:

Här vi kan se för varje variable vi skapat ett nytt column.

Dummy Variable:

Det är också samma som One-hot-encoding , här specialt är att vi tas bort for att undvika multikollinearitet.

|  |  |  |
| --- | --- | --- |
| city | Goteborg | Malmo |
| Goteborg | 1 | 0 |
| Malmo | 0 | 1 |
| örebro | 0 | 0 |

Här vi tas bort örebro binary värdet column, och vi tas bort city column också for att göra nästa steg i model flödet.

8. 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?

Ans: Nominal data : Datan utan nån inre rangordning eller ordning(exemple: färger, kön, stadnamn, etc.)

Ordinal data : Data med inre rangordning och specifik ordning(exemple: kundtillfredsställelse, en persons humör)

Julia är rätt, i den övanstående exemplet kommer färgen på skjortorna under nominal data eftersom färger inte har någon ordning eller relation mellan dem. När du säger att en person i röd skjorta ser mer attraktiv ut på en fest innebär det att vi sätter rött på första plats, eftersom rangordningen börjar automatiskt, vilket omvandlar det till ordinaldata eftersom rangordning är inblandad.

9. 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?

Ans: Streamlit är ett python bibiliotek används för att bygga web-applikationer med Python-kod för Machine Learning(ML) och Data Science(DS).

Vi änvands Streamlit för att skapa interaktiva web applikationer där vi kan integrera med de byggda regressions eller klassifieringmodellerna. Vi kan göra förutsägelser genom att ge ny data. Vi använder också Streamlit för att distribuera(deploy) applikationer.

# Självutvärdering

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

Ans: Det tog lite tid att förstå hela innehållet, men genom att lära mig steg för steg blev det enklare.

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

Ans:

1. Något du vill lyfta fram till Antonio?

Ans : Nej.