Kunskapskontroll 2 - Maskininlärning

Uppgiften görs och lämnas in individuellt men ni uppmuntras till att samarbeta och använda alla hjälpmedel så länge som ni förstår vad ni gjort och det ni lämnar in är ert egna arbeta, precis som i arbetslivet. För de teoretiska frågorna nedan får ni inte använda ChatGPT.

Från och med nu i kursen så har det historiskt sett varit en avsevärd mindre andel som får VG då examinationerna är mer krävande, för att bli godkänd så skall du visa att du använder koncepten vi lärt oss korrekt. Att uppnå godkänt är i sig inte ”enkelt” och kräver ansträngning. Satsar du på VG så förbered dig på att det blir mer arbete och kräver djupare förståelse samt problemlösning (se exakta betygskriterier i studiehandledningen). Två tips är att börja så snart som möjligt för att snabbt upptäcka vad som kommer vara utmanande samt kommunicera med varandra för att få tips.

I denna examination så kommer ni även träna i att skriva en riktig rapport, något vi kommer göra i kurser framöver samt i ert slutgiltiga examensarbete och som gör ert arbete lättare och mer systematiskt.

På Omniway skall ni Lämna in en GitHub länk som innehåller:

1. Rapporten.
2. Koden.

Vid frågor / funderingar, prata med Antonio på lektionerna eller skicka mejl via Omniway. Skall bli väldigt spännande att följa och läsa era arbeten. Kör hårt och ha kul!

/Antonio

G-delen

G-delen består av att ni skall (1) besvara teoretiska frågor (de skrivs i rapporten från steg 3), (2) modellera MNIST datan och (3) skriva en rapport. **För rapporten; läs dokumentet ”rapport_guide” för att se hur en rapport skall skrivas och använd mallen ”rapport_mall” när du skriver din rapport som kommer inkludera en självutvärdering.** För inspiration hur en mer omfattande rapport (slutgiltigt examensarbete) kan se ut, se t.ex. tidigare studentuppsatser som finns uppladdade på GitHub.

1. Teoretiska frågor

Besvara nedanstående teoretiska frågor koncist.

1. Kalle delar upp sin data i ”Träning”, ”Validering” och ”Test”, vad används respektive del för?
2. 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”?
3. Vad är ”regressionsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden?
4. Hur kan du tolka RMSE och vad används det till:

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2}$$

5. Vad är ”klassificieringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?
6. Vad är K-means modellen för något? Ge ett exempel på vad det kan tillämpas på.
7. Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding. Se mappen ”l8” på GitHub om du behöver repetition.
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?
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?

2. Modellera MNIST

Använd maskininlärning för att modellera MNIST datan. Du skall utvärdera minst två olika modeller i ditt arbete och göra ett komplett ML-flöde, från början där du laddar in data till slut där du utvärderar den bäst valda modellen på din test data. Hur du laddar ned MNIST datan kan du se här.

```
import numpy as np
from sklearn.datasets import fetch_openml
```

```
mnist = fetch_openml('mnist_784', version = 1, cache = True, as_frame = False)
print(mnist.DESCR)
```

```
X = mnist["data"]
y = mnist["target"].astype(np.uint8)
```

3. Rapport

Skriv kort och koncist. Läs dokumentet ”rapport_guide” för att se hur en rapport skall skrivas och använd mallen ”rapport_mall” när du skriver din rapport

VG delen

De som satsar på VG behöver genomföra arbetet i G delen på ett bra sätt (se betygskriterier i studiehandledningen för exakta detaljer) samt nedanstående:

* Skapa en Streamlit applikation där man med hjälp av en modell tränad på MNIST (kan vara det som görs i G delen) används för att prediktera ”ny data” som matas in via Streamlit. Den ”nya datan” kan t.ex. komma från att användaren laddar upp en egen hand-ritad bild, att man ritar en siffra via Streamlit eller att man tar en bild via kameran på datorn. Du väljer själv.