Kerstin Palö, 2025-01-24

Kunskapskontroll 2 – Del 1, Teoretiska frågor

Teoretiska frågor

Besvara nedanstående teoretiska frågor koncist.

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

   Svar:

   Träningsdata

   Används för att träna modellen, det vill säga för att justera modellens interna parametrar (vikter och biasar).

   Modellen lär sig mönster och samband från den här datan.

   Vanligtvis utgör träningsdatan 60–80% av hela datasetet.

   Valideringsdata

   Används för att finjustera modellen och välja hyperparametrar (som inlärningshastighet, antal lager i ett neuralt nätverk, eller maxdjup i ett beslutsträd).

   Hjälper till att upptäcka om modellen överanpassar (overfitting) till träningsdatan.

   Modellen använder inte valideringsdatan under träningen – det är som ett "mellanprov".

   Vanligtvis utgör valideringsdatan 10–20% av hela datasetet.

   Testdata

   Används för att utvärdera modellens prestanda efter att den är färdigtränad och optimerad.

   Testdatan används endast en gång, i slutet, för att ge en oberoende bedömning av modellens generaliseringsförmåga.

   Vanligtvis utgör testdatan 10–20% av hela datasetet.

Kerstin Palö, 2025-01-24

2. Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding.

Svar:

Ordinal encoding:

- Kategoriska värden ersätts med siffror som återspeglar en **naturlig rangordning**.
- Används för **ordnade kategorier** där det finns en inbördes hierarki, men där avstånden mellan värden inte är jämna.

Exempel:
Kategorin **Utbildningsnivå**:

- Grundskola → 0
- Gymnasium → 1
- Högskola → 2
- Doktorand → 3

One-hot encoding:

- Varje kategori representeras som en binär vektor där endast en position är "1", medan övriga är "0".
- Används för **nominella variabler** (utan inbördes ordning).

  Exempel:
  Kategorin **Färg**:

  - Röd → [1, 0, 0]
  - Grön → [0, 1, 0]
  - Blå → [0, 0, 1]

Dummy variable:

- En variant av one-hot encoding där en kategori **utelämnas** för att undvika **multikollinearitet** (s.k. "dummy trap").
- Den utelämnade kategorin fungerar som **referenskategori** i analysen.

Exempel:
Kategorin **Färg**:

- Röd → [1, 0]
- Grön → [0, 1]
- Blå → [0, 0] (referenskategorin).

Kerstin Palö, 2025-01-24

3. 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?

> Svar: Julia har rätt. Det beror på den kunskap man har om de olika kategorierna om de kan sägas vara ordinal eller nominal. Betydelsen i sammanhanget.
>
> Man skall fråga sig själv: Om jag byter plats på kategorierna, ändrar det tolkningen av datan?

4. Läs följande länk: https://stackoverflow.com/questions/56107259/how-to-save-a-trained model-by-scikit-learn

(speciellt svaret från användaren som heter ”sentence”) som beskriver ”joblib” och ”pickle”.

Det är alltså ett sätt att spara modeller och innebär att man kan träna en modell och sedan återanvända den för att göra prediktioner utan att behöva träna om modellen. Detta kommer ni ha nytta av om ni satsar på VG delen.

Svara på frågan: Vad används joblib och pickle till?Frågorna är kopplade till slides 1-3 från kursvecka

> Svar:
>
> Joblib och pickle är Python-bibliotek som används för att spara och läsa in Python-objekt till/från en fil för senare användning, så att man slipper generera en ny modell varje gång.