Seminar - SOK3023

Fagansvarlig: Markus J. Aase
Universitetslektor i matematikk og statistikk

February 4, 2025

Introduksjon

Dette seminaret fokuserer på forståelse og intuisjon i maskinlæring - tatt fra pensum i SOK-3023 *Maskinlæring for økonomer* på UiT. Dere skal jobbe i par, svare på spørsmål og forklare dem til hverandre. Noen spørsmål krever at dere har vært i forelesning, leser i kompendiet, eller søker opp på internett.

Læringsmål:

- Forstå viktige maskinlæringskonsepter, og kunne forklare dem.
- Forberedelse på muntlig eksamen.
- Vurdere hvilke modeller som egner seg for ulike typer datasett.

Oppgave 1: Diskusjonsspørsmål

Diskuter og forklar følgende spørsmål til hverandre. Skriv ned korte svar.

1. **Hva er forskjellen på prediksjon og inferens?** Kan du gi et økonomisk eksempel på begge?
2. Forklar **bias-variance tradeoff**. Hvilken modell har høy bias? Hvilken har høy varians?
3. Når bør du bruke **L1-regularisering (Lasso)** kontra **L2-regularisering (Ridge)**? Hva skjer med koeffisientene?
4. Hva betyr **en epoch** i treningen av et nevralt nettverk? Hva skjer om vi trener for få eller for mange epoch?
5. Hva er **en batch**?
6. Hvilke evalueringsmetoder ville du valgt for et **klassifikasjonsproblem**? Hva med et **regresjonsproblem**?

7. Kom med noen eksampler hvor maskinlæring kan være nyttig i samfunnsøkonomien?

8. Hva er forskjellen mellom **stokastisk gradient descent (SGD)**, **Adam** og **vanlig gradient descent**?

9. Hva er et beslutningstre, og hvordan funker de?

10. Hva er random forests, og boosting? Hvordan relaterer disse til uttrykket *ensemble learning*?

11. Skisser opp Sigmoid-funksjonen og ReLU-funksjonen. Hva er de, og *what is it good for*?

12. Hvorfor er **max pooling** nyttig i en CNN? Hva med **konvolusjon**?

13. Forklar hvorfor **Whisper-modellen** er relevant for maskinlæring? Og hva er greia med Nasjonalbiblioteket og Whisper?

14. Hva er *kryssvalidering*, og hva kan det brukes til? Kom med konkrete eksempler.

15. Hva betyr det at et nevralt nettverk er *fully-connected*?

16. **Vurdering av en binær klassifikasjonsmodell**

    En modell har følgende **confusion matrix**:

$$\begin{bmatrix} 980 & 10 \\ 8 & 2 \end{bmatrix}$$

    Der:

    - 980 er sanne negative (TN)
    - 10 er falske positive (FP)
    - 8 er falske negative (FN)
    - 2 er sanne positive (TP)

    (a) Beregn **accuracy**, **precision** og **recall** for modellen. Hva med F1-score?

    (b) Hvorfor kan modellen ha høy accuracy, men lav precision og recall?

    (c) Er dette en god modell? I hvilke situasjoner vil den være problematisk?

17. **Forventningsverdi og feil**

    Vis at forventningsverdien av kvadratisk feil kan skrives som:

$$E[(Y - \hat{Y})^2] = (f(X) - \hat{f}(X))^2 + \text{Var}(\epsilon)$$

    (a) Definer forventningsverdi og varians av feilleddet, altså ϵ.

    (b) Bruk forventningsoperatoren $E[\cdot]$ til å vise at uttrykket holder.

(c) Hva forteller dette oss om hvordan feil i maskinlæringsmodeller oppstår?

18. **Gradient Descent - Oppdatering av vekter** Vi har en kostnadsfunksjon $C(w)$ som avhenger av vekten w (vi antar at vi bare har én vekt her). Vi ønsker å minimere denne med gradient descent, hvor oppdateringsregelen er:

$$w^{(t+1)} = w^{(t)} - \alpha \cdot \frac{dC}{dw}$$

Anta at kostnadsfunksjonen er $C(w) = (w-3)^2$ og læringsraten er $\alpha = 0.1$. Start med $w^{(0)} = 5$.

(a) Beregn gradienten $\frac{dC}{dw}$.

(b) Utfør tre iterasjoner av gradient descent.

(c) Hvordan utvikler w^t seg, når t går fra 1, 2, 3?

Oppgave 2: Valg av maskinlæringsmetode

Dere får følgende tabell med datasett. For hvert tilfelle, diskuter og bestem hvilken maskinlæringsmodell dere ville brukt, og hvorfor.

Datasett	**Input-variabler**	**Output**
Boligpriser	Kvadratmeter, beliggenhet, antall rom, byggear	Pris (kr)
Kundesegmentering	Alder, inntekt, kjøpshistorikk	Kundekategori (A, B, C)
Svindeldeteksjon	Transaksjonsbeløp, sted, tid, type vare	Svindel? (Ja/Nei)
Aksjeprisforutsigelse	Historiske priser, rente, volatilitet	Neste dags pris
Tekstanalyse	Kundeanmeldelser (tekst)	Positiv/negativ

Table 1: Hvilken metode ville du valgt til de ulike datasettene?

Oppgave 2 - del 2

Rent praktisk, hvordan ville dere løst et av problemene over? Hvilke loss-funksjon hadde egnet seg, hva må dere gjøre med dataene?

Oppgave 2 - del 3

Åpne Google Colab, eller en form for Python-interpreter, og kjør koden under:

```
from sklearn.datasets import fetch_california_housing
data = fetch_california_housing()
data
```

Her er et eksempel på kode som laster inn datasettet "California Housing" fra `scikit-learn`. Kan du/dere lage en maskinlæringsmodell på dette datasettet? Hvilke metoder da?:) Gjerne bruk kunstig intelligens for å komme i gang.

Oppsummering

Etter dere har jobbet med dette (to-og-to **personer**), skal dere gå sammen to og to **grupper** å diskutere hva dere har kommet frem til.