Nasjonalt senter for
e-helseforskning

Syntetiske data

Per Atle Bakkevoll
Seniorrådgiver/stipendiat

Hva skal vi snakke om?

- Hva er syntetiske data?
- Hvordan produseres de?
- Fordeler og utfordringer
- Bruksområder
- Evaluering av syntetiske data
- Oppsummering

Bakgrunn

- Tilgang til treningsdata er en flaskehals for KI i helse- og omsorgstjenesten
- Bedre personvernfremmende teknologier kan gjøre tilgangen enklere og raskere
- Syntetiske data er et eksempel på en slik teknologi

Hva er syntetiske data?

- Kunstig genererte data som ligner på data fra den virkelige verden
- Datamodaliteter: bilder, video, tekst, tabelldata, osv.
- De kan brukes til å trene maskinlæringsmodeller, til testing av programvare eller for å utføre statistiske analyser
- Genereringen kan skje på flere måter:
  - ML-metoder (datadrevne metoder)
  - Kunnskapsdrevne metoder
    - Beslutningsregler, statistiske og matematiske modeller
  - Hybride metoder, som kombinerer tilfeldig generering ved ML-metoder med kunnskapsdrevne metoder

Genering av syntetiske data med ML-metoder

- En generativ ML-modell trenes med ekte data
  - Nevrale nettverk-modeller er de mest populære
    - Generative Adversarial Networks (GAN)
    - Variational Autoencoders
  - Beslutningstre-baserte modeller
  - Bayesianske nettverk
- Den trente modellen lærer sannsynlighetsfordelinger og sammenhenger mellom variabler i det ekte datasettet
- Dette brukes til å generere syntetiske data som har tilnærmet samme egenskaper som det ekte datasettet

Fordeler med syntetiske data

- Mindre risiko for personvernbrudd
  - Tilfeldig genererte data, representerer ikke virkelige personer (dersom modelleringen var vellykket)
- Kan gi større og mer varierte datagrunnlag der det fins lite ekte data
  - Sjeldne sykdommer
- Kan korrigere for kjente skjevheter, urettferdigheter og mangler i det originale datasettet
    - F.eks. underrepresenterte demografiske grupper

Utfordringer ved syntetiske data

- Hvordan redusere risiko for personvernbrudd, samtidig som nytten av dataene bevares?
- Personvern
  - For å ha nytte må de syntetiske dataene være like de ekte treningsdataene, men ikke *for* like
    - Kan inkludere personvernfremmende teknologier som differential privacy eller ulike krypteringsteknikker
  - ML-metoder kan lekke personidentifiserbar informasjon fra treningsdataene de syntetiske dataene er generert fra (overtilpasning)
  - Sårbare for avdekking av medlemskap og attributter
- Datakvalitet
  - Datakvaliteten avhenger av kvaliteten til de ekte treningsdataene
    - Feil og mangler vil arves av de syntetiske dataene
  - Ikke gitt at ML-modellen lærer alle relevante sammenhenger i det ekte treningsdatasettet
    - Nye feil og mangler introduseres

Syntetiske data som treningsdata

- For å bruke syntetiske data som treningsdata må de være representative for de dataene som ML-modellen skal brukes på
- Syntetiske data er en umoden teknologi og bruken i helse- og omsorgstjenesten er beskjeden
- Det er nødvendig å bygge kunnskap og erfaring med å bruke syntetiske data til å trene ML-modeller
- Empiriske evalueringer som sammenligner syntetiske datasett med reelle datasett er nødvendig for å skape tillit

Bruksområder

- Bruke syntetiske data i tidlige faser i utvikling av ML-modeller som senere evalueres med reelle data
  - Train on Synthetic, Test on Real (TSTR)
    - En ML-modell trenes med syntetiske data og ytelsen testes med reelle data
  - Transfer learning (fra syntetiske data til reelle data)
    - En ML-modell trenes, valideres og testes med syntetiske data
    - Den forhåndstrente modellen forbedres gjennom en ny runde med trening, validering og testing med reelle data

Eksempel

- En forskergruppe planlegger å utvikle en ML-modell og har behov for pasientdata til trening
- Kan spare tid ved å få rask tilgang til et anonymt, syntetisk datasett basert på et ekte datasett mens de venter på tilgang til det ekte datasettet
- Modellutviklingen starter med det syntetiske datasettet
- Når de får tilgang til det ekte datasettet brukes dette til å forbedre ML-modellen som ble trent med det syntetiske datasettet

Validering av KI-systemer

- Begrepet «validering» har ulik betydning i KI-feltet og blant klinikere
- Teknisk validering er et av stegene i treningen av en ML-modell
  - Kan gjøres med syntetiske data
- Klinisk validering betyr at KI-systemet evalueres i en ekte klinisk setting for å bedømme effekt og pasientsikkerhet
  - Data fra ekte pasienter i en klinisk setting
- Teknisk valideringer er ikke tilstrekkelig for å validere den kliniske ytelsen eller generaliserbarheten til modellen
- Manglende klinisk validering er en viktig grunn til at KI-baserte systemer ikke er tatt i bruk i større skala

Tilgjengeliggjøre data for forskning

- Mye forskning skjer på de få tilgjengelig åpne datasettene med pasientdata som eksisterer, som MIMIC-III
  - Kan føre til skjevheter i forskningen
- Syntetiske data har lavere risiko for re-identifisering, kan de publiseres som åpne datasett?
- Vanskelig å generere syntetiske data som har høy nytte, samtidig som de har lav risiko for personvernbrudd
- Grundig evaluering av personvern og datakvalitet
- Akseptabel risiko for re-identifisering vil avhenge av hvilken type tilgang som gis

Evaluering av syntetiske data

- Sammenligne det syntetiske datasettet med et ekte datasett
- Nøyaktighet
  - Evalueres ved å utføre statistiske analyser på begge datasettene. Nøyaktigheten er høy dersom det er stor grad av statistisk likhet mellom datasettene
- Nytte
  - Nytten avhenger av det konkrete bruksområdet
  - Evalueres ved å teste hvor godt det syntetiske datasettet presterer på en bestemt oppgave og sammenligne med hvordan det ekte datasettet klarer samme oppgave
  - Nytte-evalueringer kan ikke generaliseres til andre brukstilfeller eller ML-modeller
- Personvern
  - Metrics
  - Simulere ulike personvernangrep
- Utfordring: det eksisterer ingen standardiserte måter å evaluere statistisk likhet, nytte eller personvern.

Oppsummering

- Syntetiske data kan være et nyttig verktøy, men må brukes med forsiktighet
- Ikke mulig å gi et generelt svar på om vi bør bruke syntetiske data til å trene ML-modeller
  - Vil variere for ulike brukstilfeller
  - Trenger empiriske evalueringer av datakvalitet og personvern
- Trenger mer forskning og utvikling av metoder og verktøy for generering og evaluering av syntetiske data for å kunne gi bedre svar på nytteverdien