4.4 Synthetic Data

ZhaoZhi Qian
zq224@cam.ac.uk

Agenda

- Motivation: synthetic dat
  - Used case for ML training on sensitive medical data
- Generation of synthetic data
  - Maintaining data fidelity
  - Ensuring privacy preservation
- Evaluation of synthetic data

Healthcare data

- An essential resource
- Availability of healthcare data resources:
  - Catalyze a complete transformation in healthcare in ML
- Made available due to digitalising data
- History
  - Open access datasets = significant progress
  - WordNet → progress in NLP
  - ImageNet -> imaging
- Complex for healthcare data
  - Ethical considerations
    - Privacy concerns
  - Multiple sources and modalities:
    - Complex
    - Diverse populations
    - Different uses
  - Some initiatives:
    - MIMIC dataset
      - Still focusing on ICU patients
      - Much more can be done for accessibility
  - Sharing data
    - Companies/organization trying to lock up access to data
      - to productize their models
      - Privacy
      - Strict regulations
  - Subsequently, lack of high quality data

De-identified data vs Synthetic data

- De-identified/anaonymized data
  - Real data with all personal identifiers removed/ data fields scrambled
  - Gender + ZIP code + DoB can identify a person
- Synthetic data
  - Created from scratch
  - Cannot be synced back to any individual
  - However, require ML/ statistical modeling

How can synthetic data help?

- Share a synthetic (proximal) version of data that resembles real data but contains no real samples for any specific individual
-

- Use cases for synthetic data:
    - Enabling data sharing for developing analytics
    - Facilitating reproducibility of clinical studies and analyses
    - Augmenting small-sample data sets:
        - Data for rare diseases
        - Data for underrepresented patient subgroups (to guard against model bias)
    - Increasing robustness and adaptability of ML models (transferring across hospitals)
- Synthetic clinical data in action: biomedical imaging
    - Nature Biomedical Engineering 2021

Desiderata for synthetic data generation

- Generative modelling
    - Coupled with discriminative modelling
        - Does not condition on impute features
    - Application for clinical data is different
        - Complex and diverse data structures
        - Domain knowledge

https://www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/