Imperial College London

Machine Learning for Neuroscience

Ethical considerations and responsible machine learning

Payam Barnaghi
Department of Brain Sciences
Imperial College London
January 2023

1

Machine learning for healthcare and medicine

Imperial College London

- The ultimate goal of developing machine learning models and systems in healthcare is deploying them in real-world setting and using them for patient care.

2

2

ML deployments in real-world settings

Imperial College London

- The systems are often are used for decision-support by keeping human-in-the-loop.
- However, trustworthiness, reliability and robustness of the systems/models must be considered and investigated prior to the deployments.
- The users' perceptions of the system and appropriate training should be also considered.

3

3

Offline and online models

Imperial College London

- Some models are designed to process offline and pre-collected datasets and can provide exploratory analysis and insights to the contributing factors to a diseases, cluster and groups of features or patients based on different fractures.
- Online models are trained based on some existing data and deployed in operational settings. Sometimes their parameters are fixed and sometimes they are either periodically re-trained or learn as new data emerges (i.e. online or continual learning).

4

4

Continual learning

Imperial College London

- "Modern machine learning excels at training powerful models from fixed datasets and stationary environments, often exceeding human-level ability."
- "Yet, these models fail to emulate the process of human learning, which is efficient, robust, and able to learn incrementally, from sequential experience in a non-stationary world."

Source: R. Hadsell et al., Embracing Change: Continual Learning in Deep Neural Networks, Trends in Cognitive Sciences, December 2020, Vol. 24, No. 12.

5

5

Why do we need continual learning?

Imperial College London

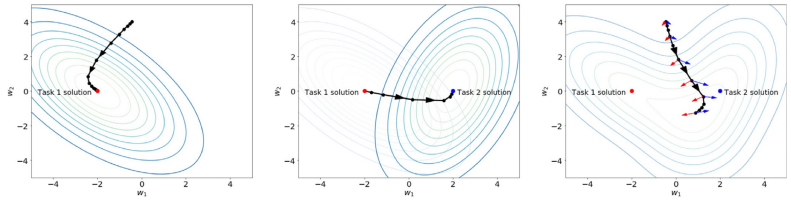
- In continual learning, the model repeatedly receives new data, and the training data is not complete at a fixed given time.
- If we re-train the entire model whenever there are new instances, it would be very inefficient, and we have to store the trained samples.
- The key challenge continual learning scenarios in changing environments is how to incrementally and continually learn new tasks without forgetting the previous or creating highly complex models that may require accessing the entire training data.

6

6

Continual learning

Imperial College London



Source: R. Hadsell et al., Embracing Change: Continual Learning in Deep Neural Networks, Trends in Cognitive Sciences, December 2020, Vol. 24, No. 12.

7

7

Forgetting problem in machine learning models

Imperial College London

- Most of the common deep learning models are not capable of adapting to different tasks without forgetting what they have learned in the past.
- Updating and altering tasks of an already learned model leads to the loss of the previously learned knowledge as the network is not able to maintain the important weights for various distributions.
- The attempt to sequentially or continuously learn and adapt to various distributions will eventually result in a model collapse. This phenomenon is referred as catastrophic forgetting or interference (*McCloskey et al., 1989*) (*Goodfellow et al., 2013*).

8

8

Developing clinically and/or care applicable solutions

Imperial College London

- Before designing any model the dataset should be thoroughly investigated and the collection setting/condition, noise, bias, imbalance and suitability of the dataset for the planned analysis be carefully considered.
- We should investigate how and when the data is collected and how and when it is going to be used and for what purpose.

9

9

Outcomes and end-points

Imperial College London

- The clinical and care end-points should be clearly defined.
- For example if the model is going to be used for prediction of an adverse health condition in a hospital setting, what timeframe would be clinically useful for the model to make the predictions? i.e., predicting a specific condition 5 mins before it happens may not be as useful in a real-world setting.

10

10

Multi-source data

Imperial College London

- Harmonising the data and investigating different sources of noise, potential error and inconsistencies are important.
- If different devices are used to collect the data, you need to consider solutions to reduce the effect of calibration and measurements errors and variations.
- You need to investigate the protocols and procedures that have been used in each site to collect the data to make sure the data is consistent.
- For more information please refer to Alexander Capstick's and Francesca Palermo's work on the LAP model:
  - https://github.com/alexcapstick/LossAdaptedPlasticity
  - https://arxiv.org/abs/2212.02895

11

11

Ethical implications - bias

Imperial College London

- Several works have identified ways in which non-health-related ML can exacerbate existing social inequalities by reflecting and amplifying existing race, sex and other biases.
- Health care is not immune to bias.
- The health data on which algorithms are trained are likely to be influenced by many facets of social inequality, including bias toward those who contribute the most data.

Source: Wiens, J., Saria, S., Sendak, M. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 25, 1337–1340 (2019).

12

12

Algorithmic bias

Imperial College London

- Algorithms that predict an individual's risk for a condition or suitability of a specific treatment may be biased toward those who are able to access and afford the procedure.
- This could happen by feeding data from the people have had that treatment in the past which won't include people who couldn't afford it in the first place or didn't have access to it.
- Some of this bias can be corrected for during model training when the data is divided to training, validation and test sets.
- In general, awareness is necessary to investigate when potential biases could be present in the data and what can be done to mitigate their effect.

Source: Wiens, J., Saria, S., Sendak, M. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 25, 1337–1340 (2019).

13

13

Robust evaluation of the models

Imperial College London

- When a model is designed for an environment, validation within that environment requires careful thought to ensure that no unintended label leakage has occurred between the datasets used for model tuning and independent testing.
- For example, the 'radiologist-level' performance recently achieved across several tasks using chest X-rays. The data used in the analysis consisted of multiple frontal-view X-ray images per patient. It was important to split data at the patient level, as opposed to random splitting, so that no images from the same patient appeared in both the training and testing sets.

Source: Wiens, J., Saria, S., Sendak, M. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 25, 1337–1340 (2019).

14

14

Importance of qualitative analysis

Imperial College London

- Beyond quantitative measures of performance, qualitative approaches can expose concerns associated with bias and confounding that the quantitative measures might have missed.
- For example, clinical experts can investigate explanations provided at individual test points to determine whether the model is plausible and relevant.

Source: Wiens, J., Saria, S., Sendak, M. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 25, 1337–1340 (2019).

15

15

Reporting and context of the application

Imperial College London

- Proposed reporting guidelines developed by the community provide good ideas to outline the importance of clear descriptions of the source of the data, participants, outcomes and predictors, and in some cases require the model itself (e.g., regression coefficients) to be presented.
- This last requirement creates a potential for unintended consequences and even harm, if the model is then applied inappropriately.
- For example, a recent study in building models to predict healthcare-associated infections found that variables associated with risk at one hospital were protective in another.
- So it is important to report the context(s) in which the model applies and was validated

Source: Wiens, J., Saria, S., Sendak, M. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 25, 1337–1340 (2019).

16

16

Deployment and maintenance responsibility

Imperial College London

- Effectively applying a predictive model in an ethical, legal and morally responsible manner within a real-world healthcare setting can be substantially more difficult than developing a model in a curated experimental environment.
- Before integrating in patient care, it is critical to test the system in 'silent' mode, in which predictions are made in real time and exposed to a group of clinical experts but not acted upon.
- This prospective validation allows clinicians to identify and review errors in real-world settings.

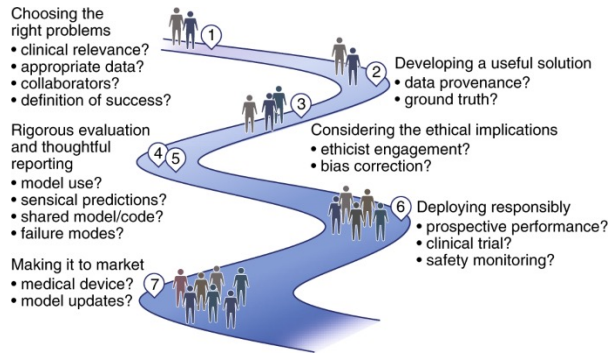
Source: Wiens, J., Saria, S., Sendak, M. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 25, 1337–1340 (2019).

17

17

A roadmap for deploying effective ML systems

Imperial College London



Source: Wiens, J., Saria, S., Sendak, M. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 25, 1337–1340 (2019).

18

18

Explainability of the models

Imperial College London

- Providing explanation to the decision making process in ML models is important to make them ore interpretable.
- For example providing feature importance/contributions in the predictions helps to provide more explanation to the decision making/prediction process of the model.
- In recent years using methods such as SHAP graph, highlights in imaging or heatmaps are used to add more explainability to the models.
- However, evaluating the explainability of the models should not be limited just to presenting them. Further investigation and clinical/care expert knowledge and analysis should be sought for the provided explanations.

19

19

Explaining what and to whom

Imperial College London

- For example in some models designed for medical imaging a heatmap or a region highlight is used to show the area that the model has been look for to obtain the results/predictions.
- "However, the important question for users trying to understand an individual decision is not where the model was looking but instead whether it was reasonable that the model was looking in this region" [1].
- A good article on this topic:

[1] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021 Nov;3(11):e745-e750. doi: 10.1016/S2589-7500(21)00208-9. PMID: 34711379.

20

20

Privacy issues

Imperial College London

- Methods such as imputation and predictive analysis could reveal information about participants that they have declared to provide.
- For example, a data imputation methods could provide an estimated for a variable (e.g. alcohol consumption) that they had declined to provide.
- Or narrowing down the analysis to small sub-group in a dataset in not handled carefully could risk making the participants identifiable.

21

21

Fairness and imbalance

Imperial College London

- Analysing the outcomes and actions driven by the designed methods could help to investigate if data or sampling issues have led to biased decisions.
- Predictions and decisions should be also investigated for potential disparity across different demographics and groups.
- The training data should be investigated for unfair bias or imbalance.

22

22

Further reading

Imperial College London

- A good article on this topic:
- Wiens, J., Saria, S., Sendak, M. et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 25, 1337–1340 (2019). https://doi.org/10.1038/s41591-019-0548-6

23

23

If you have any questions

Imperial College London

- Please feel free to come and see me (9th Floor, Sir Michael Uren Research Hub, White City Campus) or email (p.barnaghi@imperial.ac.uk).

24

24