Projects for UCL Computer Science M.Sc. Programmes and 4th Year Undergraduate Students (2024/25)

Vasileios Lampos
Department of Computer Science
University College London
v.lampos@ucl.ac.uk

N.B. I am also open to discuss different research directions that fit within my group's research interests and can be supported by data sets that are already at our disposal or do not require an additional ethics approval. Please do not hesitate to get in touch via email.

Project 1 – Time series forecasting

Time series models play an important role in many domains, from public or personalised health to physics, but also climate science and finance. Most traditional machine learning forecasting models have various limitations such as stationarity assumptions, need for extensive feature engineering, and ineffective handling of longer input sequences. Recent research efforts have focused on the deployment of modern neural network architectures (esp. Transformer) to improve time series forecasting models. However, these models have limited capacity in providing statistically accurate uncertainty bounds for their predictions, and oftentimes their development introduces irrational modelling discounts (e.g. removing focus from inter- and intra-variable dependencies) that seem to improve forecasting performance in a set of established benchmarks although in practical terms they might not (caveats explained in the Appendix of Shu and Lampos (2024)). Our research seeks to improve forecasting performance under scenarios with tangible outcomes (meteorology, epidemiology, finance), generate foundational (task-agnostic) models that can learn from a diverse set of predictors, and rigorously compare to almost forgotten (but still quite competitive) traditional forecasting approaches. We are also interested in investigating improved ways for deriving uncertainty estimates going beyond Bayesian neural networks (Morris et al., 2023), and starting from conformal prediction approaches, with the aim to retain (or improve) mean forecasting accuracy while maintaining a similar level of computational complexity. Please feel free to specify your own project within this context.

Project 2 – Machine-generated text detection

With the emergence of large language models, online text (and text content in general) is increasingly becoming machine-generated. In this project, you can attempt to build your own machine-generated text detector and assess its capacity under different scenarios (e.g. different base language models, different domains of text and so on). Please have a look at the paper by Li et al. (2024) to understand the task better.

Project 3 – COVID-19 incidence and health burden modelling at finer geographies using the Google symptoms data set

In this task, we will use a data set of symptom-related web search activity that has been released by Google.[1] In addition to covering more than 400 symptom categories, this data set offers very fine geographical granularity. The focus of the modelling will be on COVID-19. Can we use this data to understand how COVID-19 spread across a country (e.g. the UK or the US), especially during the first wave(s) of the pandemic?

N.B. This project requires a more established understanding of epidemiology. Please have a look at this related paper by Lampos et al. (2021). It is not designed to be a cutting edge machine learning project. It can also be used for the development of interesting visualisation solutions (dynamic dashboards for epidemiological monitoring). Hence, if your M.Sc. has a strong machine learning focus, I would advise to choose between projects 1 or 2.

[1] Google symptoms search trends, pair-code.github.io/covid19_symptom_dataset/

Data and computational resources

Data for the aforementioned projects is either publicly available or will be provided by us. You will be able to use our group's computational resources (including some basic GPU support), if necessary.

Ethics

These projects will use aggregate information. Data for individual users is not available to us. Given this, the projects described here have obtained an ethics exemption by UCL Computer Science.

References

Vasileios Lampos, Maimuna S. Majumder, Elad Yom-Tov, Michael Edelstein, Simon Moura, Yohhei Hamada, Molebogeng X. Rangaka, Rachel A. McKendry, and Ingemar J. Cox. 2021. Tracking COVID-19 using online search. *npj Digital Medicine*, 4(17).

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53.

Michael Morris, Peter Hayes, Ingemar J. Cox, and Vasileios Lampos. 2023. Neural network models for influenza forecasting with associated uncertainty using Web search activity trends. *PLOS Computational Biology*, 19(8).

Yuxuan Shu and Vasileios Lampos. 2024. DEFORMTIME: Capturing Variable Dependencies with Deformable Attention for Time Series Forecasting. *arXiv preprint (2406.07438)*.