Project Instructions
Advanced Machine Learning
DATA 442/642

**1. Project description**

The main goals of the project is to prepare students to apply machine learning algorithms to real-world tasks, or to leave them well-qualified to start machine learning research. The project will have two parts. The poster presentation as well as the project proposal. You may only discuss technical details of your projects with your own group members (and me, of course).
The due date for the project proposal is **Tuesday, October 26th, 2021, at 11:59 pm**. Please submit this proposal on Canvas by that time. Your proposal should be **well-written**, **well-organized**, and **reproducible** following the current standards of machine learning reproducibility in research[1]. This will ensure evidence of the correctness of your results and will enable other researchers to make use of your methods and results. The class projects will be presented as a poster presentation. You should prepare a poster, and be prepared to give a very short explanation (10 minutes), about your work. You will also have an opportunity to see what everyone else did for their projects. You have to present your project presentation by **Friday, December 4th, 2021**. You will also need to submit your poster as a PDF the day before the presentation.

**2. Project topics**

If you are looking for project ideas, please talk to me early enough, and I will be happy to brainstorm and suggest some project ideas. There are three type of projects that you can pick and these include:

- **Application project**. Pick an application that interests you, and explore how best to apply learning algorithms to solve it.

- **Algorithmic project**. Pick a problem or family of problems, and try to develop a novel variant of an existing algorithm, to solve it.

- **Theoretical project**. Prove some interesting/non-trivial properties of a new or an existing learning algorithm.

**3. Evaluation**

Projects will be evaluated based on:

- **The technical quality of the work**. Does the technical material make sense? Are the things tried reasonable? Are the proposed algorithms or applications clever and interesting?

- **Significance**. Did the authors choose an interesting or a "real" problem to work on, or only a small ?toy" problem? Is this work likely to be useful and/or have impact?

- **The novelty of the work**. Is this project applying a common technique to a well-studied problem, or is the problem or method relatively unexplored?

[1] Pineau, Joelle, et al. "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)." arXiv preprint arXiv:2003.12206 (2020)

**4. Project proposals**

The project proposal is mainly intended to make sure you decide on a project topic and get feedback from me early. Your proposal should be a PDF document. Please make sure that you give the title of the project, the project category, the full names of all of your team members, and a 300-500 word description of what you planning to work on. Your project proposal should include the following information:

- **Motivation**: What problem are you tackling? Is this an application or a theoretical result? What makes this problem interesting and important?

- **Method**: What machine learning techniques are you planning to apply or improve upon? Note that you have to only provide high level technical details of the techniques.

- **Intended experiments**: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

If you are planning to use a real world dataset, make sure that you are including all the references and links. In addition, present at least one example of prior research on the topic and include all the information of the papers that you are planning to use for your project. Use a standard format for your references (such as APA or MLA).

**5. Poster presentations**

Posters will be graded on the poster quality and clarity, the technical content of the poster, as well as the knowledge demonstrated by the team when discussing their work. For a poster example please check (`https://sigport.org/sites/default/files/docs/MLSP_2019.pdf`). For poster templates check (`https://www.overleaf.com/learn/latex/Posters`).

**6. Project ideas (This list will be updated during the semester)**

Below are suggested project ideas. The following ideas are not considered "unique projects". Each project idea can be easily extended to multiple projects and can either form an application, algorithmic, or theoretical project.

**Deep fakes detection**

Develop machine learning techniques for the detection of manipulated videos. Openface 2.0 is a great software that can be used for feature extraction from image sequences.

- Yang X, Li Y, Lyu S. Exposing deep fakes using inconsistent head poses. InICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019 May 12 (pp. 8261-8265). IEEE.

- Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H. Protecting World Leaders Against Deep Fakes. InCVPR Workshops 2019 Jun (pp. 38-45).

- Baltrusaitis, Tadas, et al. "Openface 2.0: Facial behavior analysis toolkit." 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018.

**Molecular property prediction and knowledge discovery for material design**

Develop aggregation information approaches for molecular property in-painting and graph-based methods enhanced with belief propagation techniques for molecular labeling.

- Elton DC, Boukouvalas Z, Butrico MS, Fuge MD, Chung PW. Applying machine learning techniques to predict the properties of energetic materials. Scientific reports. 2018 Jun 13;8(1):1-2.
- Boukouvalas Z, Elton DC, Chung PW, Fuge MD. Independent vector analysis for data fusion prior to molecular property prediction with machine learning. arXiv preprint arXiv:1811.00628. 2018 Nov 1.
- `http://quantum-machine.org/datasets/`
- Contact the instructor for energetics dataset.

**Multivariate and explainable data fusion for misinformation detection during high impact events**

Develop and discover machine learning methods that enable efficient, generalizable, and explainable detection of misinformation across social media modalities.

- Nørregaard, Jeppe, Benjamin D. Horne, and Sibel Adal?. "NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 13. No. 01. 2019.
- `http://hazyresearch.stanford.edu/flyingsquid`
- Islam MS, Sarkar T, Khan SH, Mostofa Kamal AH, Hasan SM, Kabir A, Yeasmin D, Islam MA, Amin Chowdhury KI, Anwar KS, Chughtai AA. COVID-19?Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. The American Journal of Tropical Medicine and Hygiene. 2020 Aug 10:tpmd200812.
- Boukouvalas Z, Mallinson C, Crothers E, Japkowicz N, Piplai A, Sudip M, Joshi A, Adal? T. Independent Component Analysis for Trustworthy Cyberspace during High Impact Events: An Application to Covid-19. arXiv preprint arXiv:2006.01284. 2020 Jun 1.