Project Instructions

Zois Boukouvalas

2021-09-01

Description

For the project, you will explore a real dataset from a real study and present your results as well as write up a mid-semester proposal. You will work in teams with at most two members.

The entire proposal may be no more than three pages (double spaced, 12 point font, 1 inch margins). The references do not count toward your page limits. The project proposal should be written in R Markdown and knitted to PDF.

You may only discuss technical details of your projects with your own group members (and me, of course).

The due date for the research proposal is **Friday, October 20th, 2021, at 11:59 pm**. Please submit this proposal on Canvas by that time.

Your proposal should be **well written**, **reproducible**, and **well-organized**.

Your project proposal should include all of the following elements:

1. **Title Page**: Include the names of all team members.
2. **Introduction**: Provide as much information about the research problem and the data. Try to answer the following questions. Why is the research problem of significant interest? What are the main research questions? What are the challenges? You should include a *brief* literature review. For the lit review, you should find peer-review articles related to the topic. Use a publication search engine such as Google Scholar.
3. **Initial Hypotheses**: Before you look at the data, provide a list of *detailed* hypotheses. E.g. you might believe that a particular set of features can be used for the detection of fraud. *Use scientific arguments to support your belief.*
4. **Data-drivin Hypotheses**: What new hypotheses are you planning to develop as you explore the data? Provide some description of things that you find interesting when you were looking at the problem and the data.
5. **Proposed work and discussion**: Desribe what you are proposing to do and provide a brief discussion of how your results may be placed in the context of the literature.
6. **References**: A list of references cited in your report. Use a standard format for references (such as APA or MLA).

Presentation (Poster presentations, Friday, December 4th, 2020)

- Your group should also prepare a 10-12 minute poster presentation. The content of the presentation should be the same as the project report (background, hypotheses, data explorations, discussion).
- No `R` code should be shown during the presentation.

Project ideas

Below are some possible topics. The following topics are **not** considered "unique projects". Each topic can be easily extended to multiple projects.

1. *Deep Fakes detection*

   - Useful references to start:
     - Agarwal, Shruti, et al. "Protecting world leaders against deep fakes." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.
     - 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.
   - Dataset: Contact the instructor for a possible dataset.

2. *Molecular property prediction using machine learning*

   - Useful reference to start:
     - Elton, Daniel C., et al. "Applying machine learning techniques to predict the properties of energetic materials." Scientific reports 8.1 (2018): 1-12.
   - Datasets:
     - http://quantum-machine.org/datasets/
     - Contact the instructor for energetics dataset.

3. *Industrial air pollution and its effect on mortality*

   - Useful references to start:
     - Leogrande, Simona, et al. "Industrial air pollution and mortality in the Taranto area, Southern Italy: A difference-in-differences approach." Environment international 132 (2019): 105030.
     - Schwartz, Joel, et al. "The effect of dose and timing of dose on the association between airborne particles and survival." Environmental Health Perspectives 116.1 (2008): 64-69.
   - Dataset: Contact the instructor for dataset.

4. *Discovering associations between true and false information using machine learning and NLP*

   - Useful references to start:
     - Nørregaard, Jeppe, Benjamin D. Horne, and Sibel Adalı. "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.
     - Sharma, Karishma, et al. "Combating fake news: A survey on identification and mitigation techniques." ACM Transactions on Intelligent Systems and Technology (TIST) 10.3 (2019): 1-42.
   - Dataset: NELA-GT-2018

5. *Ethics in machine learning*

   - Useful references to start:
     - Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." arXiv preprint arXiv:1908.09635 (2019).
     - Yapo, Adrienne, and Joseph Weiss. "Ethical implications of bias in machine learning." (2018).
   - Datasets: For interesting datasets use "A survey on bias and fairness in machine learning."

Important Dates

- **Friday, October 20th, 2021, at 11:59 pm**: Submit proposal on Canvas
- **Friday, December 8th, 2021**: Poster presentation.