Project Instructions

IDS576: Advanced Prediction Models
Theja Tulabandhula

Aim

The aim of the project is to simulate the real-world process of conceptualizing a data analytics project and bringing unique insights using statistical modeling. More specifically, the project component of this course allows you to explore a dataset and asks you to report your experience of building a statistical model to achieve some (business) goal

Group

You should form groups of 4 students for this project component. Reach out to your classmates early.

Project Outcomes

There are two due-dates for project deliverables: one intermediate and one final. See the course logistics page for the exact dates. The deliverables are:

1. **Project Report:** In at most 8 pages[1], you should explain your creative contributions in the project (modeling, optimization, inference, analysis, insights etc). For example, describe how you have built a tailored statistical model(s) for your dataset. You should also have inferences and discussion on what went wrong, went right and what can be improved (be technical here). The report can optionally be combined with code as a Jupyter notebook, and should be uploaded on Blackboard.

2. **Code and data:** The code (e.g., Jupyter notebook, if not combined with the report above) and a small sample of the data should be provided along with the report.

Presentation

You should also aim for a 10-15 minute presentation at the end of the semester (see date on the course logistics page) explaining, via your Jupyter notebook or slides, the whole project.

Ideas

Here are some ideas:

[1] 12 point, single column. You can have an appendix for supplementary material that may or may not be checked.

- Joint visual and textual sentiment analysis by embedding images and words into the same vector space.
- Reinforcement learning based control of UAVs/vehicles (say for door-to-door delivery) via simulators.
- Build seq2seq models on reddit data to generate fake and compelling content that can spread mis-information.
- Identify issues or special use cases of GPT2.
- Other options: look at recently published research articles at applied conference venues such as KDD, ICWSM, CIKM, SDM, WSDM, ICDM, CVPR, AAAI, NeurIPS and ICML, and discuss which paper you want to reproduce the results of (take note of the availability of the data!).
- Datasets:
  - Deep vision examples
  - Stanford CS221, CS229, CS224w and CS221n
  - Deep learning gallery
  - Datasets in vision.
  - Datasets in text.
- Do not train deep networks from scratch if it can be avoided. Always go for transfer learning initially. If easy feature creation and a shallow model suffices, then don't jump into using a deep network immediately.
- Team members should try to parallelize model training and other compute intensive operations, and speed up their experimental investigations. For example, this can be done on Google Colab by running multiple experiments simultaneously, one by each team member.

Grading Rubric

- Projects will be graded based on the creativity shown in handling the data and the insights drawn. The reports should be very clearly written and presented, and will be evaluated based on the *correctness*, *content*, *creativity* and *clarity*:
  - Correctness will be assessed based on the evaluation metrics used for the results, valid experimental setup and experimentation, technical correctness and the assumptions laid out.

  - Content will be assessed based on the novel contributions made in the project and project depth (e.g., why this data, why this problem, what did you do, visualization and interesting conclusions, insights, discussion of methodology). You should try to demonstrate your understanding of the relevant topics and their use in your innovative non-trivial project.
  - Creativity will be assessed based on how no-obvious your solution or contribution is and how different choices were made in the execution of the project.
  - Clarity will be assessed based on the language used, the structure of the report, the references cited, the capability of explaining in a clear and professional manner, and the clarity demonstrated in your discussions etc.

- All external material/sources (code/idea/theory/insights) used should be cited without failure. Use of pre-trained models, databases, web servers, frontend frameworks, visualization tools etc for your report/presentation-demo is allowed and encouraged, although use of proprietary software (such as Matlab, Mathematica etc.) is discouraged. This project cannot be used as part of any other course or requirement.