C S 487/519 Applied Machine Learning I
Fall 2018
Project 2: Open ML project

1 Objective

In this *team* project with 2-3 students, you are required to apply ML knowledge to solve a real-world problem.

2 Requirements

Design a problem that comes from real-world applications and design solutions to solve the problem. The problem can be totally new (from your own investigation) or from other sources (e.g., a kaggle competition problem, a problem defined in a research article). The problem can also be part of a problem that you are currently working on (e.g., towards your PhD dissertation, or Master's project, or Master's thesis).

- **Motivation**: Please clearly describe the applications/motivations of the problem.
- **Problem**: Clearly define the problem.
- **Solution**: Design reasonable solutions to solve the problem by utilizing the ML knowledge that you have learned and making use of other related ML tools.
- **Data**: Obtain proper datasets to test your solution. You can use self-created datasets or publicly available datasets. If you utilize existing datasets, they had better be reasonably big in size (e.g., with more than 10K instances). If you create your own datasets (e.g., by writing script to crawl data), your dataset may not be huge in size (e.g., with several hundreds of instances).
- **Analysis**: Properly analyze the performance of your solution.

3 Submission instructions

Everything needs to be submitted through github repository. In particular,

- Create a github repository for this project (or a folder in your github repository).
- In your github repository, create a project folder `open-proj`.
- Create folders for different stages with required information (see below).
- Submit the link to your github repository folder (for different stages) through Canvas.
- ONE member from the team can be representative to submit the project.

Stages.

- **Stage 1** (10 points): Form a team. In your `stage1` folder, create a `team.txt` file with student names.
- **Stage 2** (30 points): Formulate your problem. In your `stage2` folder, put your report with the motivations and problem definitions. Note that this report can be updated in later stages.
- **Stage 3** (30 points): Add your proposed solution to your report, continue to refine your motivations and problem definition, and update your report; Start to write code to do basic analysis or to get the data. In your `stage3` folder, put the updated report, code (if there is any), small datasets (if created by you), or links to your large datasets.
- **Stage 4** (60 points): Continue to refine your report; finish half of the code (approximately); have some preliminary results from at least one dataset. In your `stage4` folder, put the updated report, code, small datasets (if created by you), or links to your large datasets.

- **Stage 5** (170 points): Finish your solution and your algorithm analysis; finish your report. Put your final product in the `stage5` folder with the following information:
  - (2 points) `team.txt`
  - (8 points) A readme file `readme.txt` with (1) how the code base is organized, (2) the commands to run your code, and (3) data set information: if you use existing datasets to test your solutions, the readme file should include the link to the existing datasets.
  - (110 points) Code base containing your code for crawling the data (if there is any), preprocessing the data, solving the problem, analyzing your solutions, plotting analysis figures, etc.
    - Your Python code should be written for Python version 3.5.2 or higher.
    - Please properly organize your Python code (e.g., create proper classes, modules).
    - If your team created your own datasets, put the datasets in the repository.
  - (50 points) A 3-5 page report, `report.pdf`, to include the above content (motivation, problem definition, solution explanation, data description, result analysis).

Format requirement of the report.

- For LaTeX users: please use `sample-sigconf.tex` from the ACM article template (`https://www.acm.org/binaries/content/assets/publications/consolidated-tex-template/acmart-master.zip`); Additional information about formatting and style files is available online at: `https://www.acm.org/publications/proceedings-template`.
- For word users: Margin at each of the top, bottom, left, and right sides is 1.0 inch; double column; The font size is 10pt; font type is Times New Roman. Single line space.