CSE-543/343 - Machine Learning
Course Project Guidelines

Monsoon 2017

Project Group Size: 2-4 students

Timelines:
Project Start Date: Aug. 15, 2017
0^{th} **deadline**: Aug. 18, 2017 - Group finalization
1^{st} **deadline**: Aug. 24, 2017 - Proposal - A4 size poster
2^{nd} **deadline**: 6PM, Oct. 11, 2017, - Interim Report (3 pg. limit)
final submission: Nov. 18, 2017 - Final Report (4 pg. limit)
Presentations: Nov. 30^{th}

Grading Break-up:

1. Proposal: 2.5 points
2. Intermediate review: 8.5 points
3. Final review: 14 points (report + presentation + working demo)
4. Total project: 25 points

Project Topic and Data Selection
You can choose a learning task of your choice, and one (or more) corresponding dataset(s) for evaluating your learning task. For datasets:

- **Public datasets**: Wikipedia page for List of Datasets for ML Research, UCI Machine Learning Repository and DL4J for datasets from various domains.
- **Kaggle**: You can pick current Kaggle challenges as well for your project.
- **Create**: You can also chose to collect your own dataset, but factor in the time taken to collect, label (if needed), clean and process the data.

Important Note

- Please keep in mind, the dataset you choose (or create) should be sufficiently big in size and complexity.

- Data collection cannot be the main contribution of your project.

Yet Another Important Note
Learning Techniques: Please make sure you choose your learning techniques as per the *availability of necessary computing hardware*. For example, if you plan to use Deep Learning in your project, make sure you have GPU access in order to train your model. Because of the large class size, it is impossible to provide GPUs/HPC access to all groups.
And Yet Another One
Grading: It will not just be dependent on your implementation's performance accuracy, but based on how you analyzed your models and the errors they make. How well you understand the performance? What are the insights that you obtained about the workings of your model? And how did you get these insights? A fair fraction of the grade will go to diagnostic techniques applied. We will share a rubric for the grading scheme soon.

Project Proposal Format:
The project proposal will be in the form of an A4 size poster, with the following information

1. Motivation and precise problem statement - the learning task, the dataset and a strong reason for solving this problem.
2. Data Acquisition effort (if any) - writing crawlers, indexing and initial data analysis OR the choice of a public dataset.
3. Preprocessing techniques to be explored (if any) - feature extraction/representation, reduction of dataset to suit computing requirements, etc.
4. The learning techniques you would be using to compare results (1 baseline + <team-size> $\times 1$ advanced)
5. Strategy for model selection (linear, nonlinear, kernel based) and tuning hyperparameters (e.g. cross-validation).
6. Training approach(es) to be explored (gradient descent based, newton based, stochastic gradient descent)
7. Ensemble approaches (if any) (e.g., bagging, boosting, voting)
8. Evaluation metrics
9. Deliverables of individual team members, described as clearly as possible.

The proposal will *obviously* not be perfect, however, we do expect the item numbers $1, 2, 4, 8$ & 9, i.e., problem statement, the dataset, learning techniques (linear, logistic, LASSO, kernel, Support Vector regression etc.) and the individual deliverables to be **immutable**, or at least **very well thought out**.
General Guidelines:

1. The project component is 25% of the credit. Thus the complexity of the project(s) should be roughly commensurate to the credit weightage. For example, if there are four members in the group, the effort put into a project, should be commensurate with that put in a regular 4-credit course.

2. You are strongly urged to use Python as your programming language.

3. Make extensive use of existing libraries and toolboxes. But putting together the system should be your original work. We also expect that the libraries, at least the specific learning tools that you use are not simply inserted as a black box. Your analysis should indicate that you have explored them thoroughly.

4. Your strategy for initial data analytics, your learning tool and analysis of the leraner's performance/error should be fixed by the interim report for the project.

5. Do not plagiarize. We will be running plagiarism check on all the submitted code. Strictest action against offenders will be taken.

Interim report guidelines
For the interim report, please prepare a short report of **at most three pages** (Strict limit: exceeding the page limit will amount to **grade reduction** by 25% for each additional column). Use the CVPR template from this link. You may use additional pages for figures, tables and references. You are required to submit **a single pdf file**, preferably generated using LaTeX. Any other format of the report will not be accepted. Please make sure that you **write coherently**. Do not submit a report that you have not read yourself. Your report should have the following structure:

Section 1. **Introduction** - This should contain the problem statement and the motivation.

Section 2. **Related work** - Short description of relevant related works that you have read. Here you should identify the **best results** obtained so far (state-of-the-art) on the dataset you are using.

Section 3. **Dataset and Evaluation** - Describe the dataset you are using, no. of samples in training, validation and test set. If you are extracting features, please describe the ones you have already explored in a subsection. In another subsection, you should specify what evaluation metrics are you going to use.

Section 4. **Analysis & Progress** - In this section, you should report your progress so far. List the challenges you are facing, the design choices (choice of learning method, model selection strategy, hyperparameter setting, etc.) you have made or will make to overcome these challenges. Please provide

supporting evidence (graphs, plots, visualization) to show that the data is separable/not separable, whether the training is correctly done or not, why and how the hyperparameter was selected, is the model over/under fitting the data, etc. Since every data domain will have different characteristics, **be creative with your analysis**. Your analysis should give you insights into debugging your learning system to improve performance.

Section 5. **Results** - Report any results you have obtained so far, along with a short paragraph explaining your interpretation of the results and any insights you have obtained from your analysis. Comment on the gap between your models' performance and the state-of-the-art you identified in Section 2.

Section 6. **Future Work** - Clearly state the plan ahead for the remainder of the semester. Your plan should include the following:

a) which learning techniques you are going to use (defined for each team member)?

b) any modifications in dataset choice

c) any addition/deletions/modifications in the evaluation metrics that you listed in your proposal

d) what kind of analyses are you going to perform?

e) clearly define the individual team member roles for the final evaluation.

Important Note: Please keep in mind the following points:

1. After this review, you will **NOT** be permitted to change project, regroup, etc. All future evaluations will be done based on the project topic you present in this review.

2. Only groups that have been affected by late drop **AND** have to regroup will be evaluated after the break. All other groups will be evaluated in the next week.

3. Remember that for the intermediate review, majority of the credit will be assigned for Section 4 (Analysis & Progress) and Section 6 (Future Work) above. However, this does not mean that you skip the other Sections!

Interim Review Meeting:
You will need to prepare a presentation with at most 7 slides (additional backup slides may be used) to present your work for the intermediate review. These slides should be converted to pdf and submitted through backpack before the deadline. You can not use your extension days for the review meetings.

Final report guidelines
For the final report, please prepare a short report of **at most four pages** (Sections 1 - 5) containing the following:

Section 1. **Introduction** - This should contain the problem statement and the motivation.

Section 2. **Related work** - Short description of relevant related works that you have read. Here you should identify the **best results** obtained so far (state-of-the-art) on the dataset you are using.

Section 3. **Dataset and Evaluation** - Describe the dataset you are using, no. of samples in training, validation and test set. If you are extracting features, please describe the ones you have already explored in a subsection. In another subsection, you should specify what evaluation metrics are you going to use.

Section 4. **Methodology** - In this section, you should report your methodology. For each method you used, provide supporting evidence (graphs, plots, visualization) to show that the training has been done correctly, the model is not under/over fitting, to show that the data is separable/not separable, whether the training is correctly done or not, why and how the hyperparameter was selected, is the model over/under fitting the data, etc. Since every data domain will have different characteristics, **be creative with your analysis**. Your analysis should have given you insights into debugging your learning system to improve performance.

Section 5. **Results & Analysis** - Report any results you have obtained so far, along with a short paragraph explaining your interpretation of the results and any insights you have obtained from your analysis. Comment on the gap between your models' performance and the state-of-the-art you identified in Section 2. Discuss limitations of your approaches, report failure cases and suggestions for improvement (if any).

Section 6. **Contributions** - Clearly list individual contributions made toward this project. Your plan should include the following:

a) **Deliverables**: For each team member, list all deliverables promised in the proposal, and point out the ones that *were* delivered.

b) **References & Citations**: Cite all the code (and other material like paper, tutorial, blog, etc.) that you have used. Using someone's work without giving them credit is unethical and will be penalized during the evaluation.

c) **Individual Contributions**: Please provide two parts for each team member.

i) **Brief description** of the contribution made by the team member.

ii) **List of files** comprising the functions/modules/scripts mainly contributed by the team member.

Please make sure that you write coherently. You may use **additional pages** for Section 6, figures, tables and references. Use the CVPR template from this link. You are required to submit a single pdf file, preferably generated using LaTeX. Any other format of the report will not be accepted.

1 Submission Guideline

- Create a single file **group-id.zip** and upload it to the following Google drive link.
  Final Project Submission Link
- **group-id.zip** will contain three items, two folders(**code** and **models**) and one report(**group-id-report.pdf**).
- **code** will contain all code files/scripts.
- **models** will contain all trained models.
- **group-id-report.pdf** will contain final project report.