CSE-343/ECE363/563 - Machine Learning
Course Project Guidelines

Winter 2022

Project Group Size: 3 students (in exceptional cases 2 students will be permitted)

Deadlines:
Project Start Date: Jan. 28, 2022
1^{st} **deadline**: Feb. 08, 2022 - Proposal - 1-page A4 size
2^{nd} **deadline**: 11:59AM, Apr. 4, 2022, - Interim Report (2 pg. limit)
Interim Presentations: Apr. 4-5, 2022
Final report submission: 11:59PM, May 15, 2022 - Final Report (4 pg. limit)
Final Presentation Slides: 7:00AM, May 17, 2022 (instructions on Classroom)
Final Presentations: May 17, 2022

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.
- **Data Challenges**: You can pick current or past data challenges from Kaggle, and workshops from various different conferences. See an sample list at the end of this document.

- **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. **The focus should be on applying ML techniques, evaluating and analyzing them**[1].

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. **You can (and should) leverage Google Colab as much as you can; it provides a reasonable GPU free of cost**.

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. You are highly encouraged to use techniques in the *Machine Learning Yearning* book by Andrew Ng for analyzing your model's performance, and improving upon the baseline methods that you try.

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.

[1] Best avoided if this is your first hands-on ML project.

4. Your strategy for initial data analytics, your learning tool and analysis of the learner'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.

Project proposal format:

The project proposal will be in the form of a single page A4 size document, 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(s).

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 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 and Error Analysis approaches.

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**.

Interim report and presentation guidelines

For the interim report, please prepare a short report of **at most two 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. 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 projects.

Final report and presentation 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.

Final Project Presentation:
You will need to prepare a presentation with at most 8 slides (additional backup slides may be used) to present your work for the project. These slides should be converted to pdf and submitted through classroom before the deadline. You can not use your extension days for projects. Your presentation should be organized in the following manner:

- Problem Statement & Dataset (1 slide)
- Progress summary until intermediate submission (1 slide)
- Progress after intermediate submission - approaches and results (2-3 slides)
- Analysis and Ablation (1-2 slides)
- Individual Contribution (1 slide)

Example Projects
You may pick projects from data challenges from various sources. Some examples are given below. Reach out to the TAs if you are unable to find the resources.

- **Kaggle**
    1. Translate chemical images to text
    2. Determine if two products are the same by their images
    3. Find individual human cell differences in microscope images
    4. Predict which Tweets are about real disasters and which ones are not
- **CVPR workshops**
    1. Computer vision for fashion retrieval
    2. Skin Image Analysis
    3. Computer Vision in Sports performance statistics
    4. Fine-Grained Visual Categorization
    5. Large Scale Computer Vision for Remote Sensing Imagery
    6. Image Restoration and Enhancement
    7. Visual Question Answering and Dialog
- **ACL/EMNLP**
    1. Information Extraction
    2. Machine Translation and Multilinguality
    3. Sentiment Analysis
    4. Spoken Language Translation
    5. Summarization
    6. Question Generation
    7. Dialogue Generation
    8. Coreference resolution
    9. Question Answering
    10. Identifying real or fake news
    11. Fact Extraction
- **MICCAI**
    1. Thyroid Nodule Segmentation and Classification in Ultrasound Images
    2. Image analysis of anatomical structures and lesions
    3. Medical image reconstruction

4. Cellular image analysis
5. Medical image retrieval
6. Computer-aided detection/diagnosis
7. Modeling and predicting disease development or evolution from a limited number of observations
8. Forecasting disease/cancer progression over time
9. Predicting missing data
10. Predicting clinical outcome from medical data

- **INTERSPEECH**

1. Speaker and Language Identification
2. Analysis of Speech and Audio Signals
3. Speech Synthesis and Spoken Language Generation
4. Speech Recognition
5. Spoken Language Processing (Dialog, Summarization, Understanding, Translation and Information Retrieval)
6. Analysis of Paralinguistics in Speech and Language
7. Speech Perception, Production and Acquisition

- **ECML**

1. Parallel, Distributed, and Federated Learning
2. News Recommendation and Analytics
3. Machine Learning and Data Mining for Sports Analytics
4. Machine Learning for Cybersecurity
5. Machine Learning for Earth Observation

Important Note about picking Datasets
Depending on the size of the datasets, you may not be able to train full-blown state-of-the-art models for these problems. One way to resolve this would be to scope the problem / dataset / model down to a level that you can handle with the compute available. This trade-off is very important from a practical standpoint as you may have compute resource / power constraints at the time of training and / or deployment.

1. Reinforcement Learning for Autonomous Driving (using simulators) (SeeOpenAI Gym) or Udacity's Self-Driving Car Simulator.
2. Reinforcement Learning for Atari Games (SeeOpenAI Gym)
3. Adversarial attacks on machine learning (design attacks that systematically apply transformations to data that trigger a failure of an otherwise working model). See the Kaggle Challenge on Targeted Adversarial Attacks

4. Non-human primate face/gender recognition or age prediction, e.g., see the Chimpanzee Faces in the Wild

5. Traffic light detection and recognition system LISA or LARA

6. NLP based Q&A (See DL4J: Question answering)

7. Sentiment analysis DL4J: Sentiment Analysis

8. Recommendation and Ranking DL4J: Recommendation & Ranking

9. Text Classification, News Summarization, Topic prediction, etc. Text Data.

10. DL4J has a number of datasets that you can use for a variety of learning tasks. **Note**: If the dataset is too large, feel free to use a smaller subset in your project.

11. Graph based projects: Analytics on Networks (Communication/Social/Biological). See datasets at the Stanford Large Network Dataset Collection.

12. Deep learning for reconstruction of compressively sensed videos:
This topic includes applying deep learning to recover videos for which a fewer measurement data is sensed or collected at the receiver instead of sensing the entire video. Thus, effectively this implies that only few samples (instead of complete frames) can be transmitted saving the time and bandwidth both. And recovering full video data at the receiver. Some sample reference papers are:

    - Iliadis, Michael, Leonidas Spinoulas, and Aggelos K. Katsaggelos. "Deep fully-connected networks for video compressive sensing." arXiv preprint arXiv:1603.04930 (2016).
    - Xu, Kai, and Fengbo Ren. "CSVideoNet: A Recurrent Convolutional Neural Network for Compressive Sensing Video Reconstruction." arXiv preprint arXiv:1612.05203 (2016).

13. Deep learning for accelerated MRI reconstruction:
This topic includes applying deep learning method to reconstruct Magnetic resonance images from a fewer data collected from the MR scanner in the k-space. In MRI, particularly, in Dynamic MRI where the subject is scanned a number of times, it is crucial that minimum possible scanning time is consumed without compromising the quality of images captured for diagnosis. In other words, subjects should require to spend minimum possible time inside the scanner. This project is aimed to meet this requirement via ML. Some sample reference papers are:

    - Yang, Yan, et al. "ADMM-Net: A Deep Learning Approach for Compressive Sensing MRI." arXiv preprint arXiv:1705.06869 (2017).

- Schlemper, Jo, et al. "A Deep Cascade of Convolutional Neural Networks for MR Image Reconstruction." International Conference on Information Processing in Medical Imaging. Springer, Cham, 2017.

14. Machine Learning for disease diagnosis using fMRI data:
    This topic includes applying deep learning method for disease diagnosis using functional MRI data. For example, one may be interested in diagnosing Alzheimer, Autism, or Parkinsons. Some sample reference papers are:

    - Ktena, Sofia Ira, et al. "Distance Metric Learning using Graph Convolutional Networks: Application to Functional Brain Networks." arXiv preprint arXiv:1703.02161 (2017).
    - Parisot, Sarah, et al. "Spectral Graph Convolutions on Population Graphs for Disease Prediction." arXiv preprint arXiv:1703.03020 (2017).

15. Predicting brain image from raw imaging data
    Reference paper: Cole, James H., et al. "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker." NeuroImage (2017).

16. Decoding of visual activity patterns from fMRI
    Reference: Zafar, Raheel, et al. "Decoding of visual activity patterns from fMRI responses using multivariate pattern analyses and convolutional neural network." Journal of Integrative Neuroscience 16.3 (2017): 275-289.

17. Brain MRI segmentation
    Reference papers:

    - Moeskops, Pim, et al. "Automatic segmentation of MR brain images with a convolutional neural network." IEEE transactions on medical imaging 35.5 (2016): 1252-1261.
    - Zhang, Wenlu, et al. "Deep convolutional neural networks for multimodality isointense infant brain image segmentation." NeuroImage 108 (2015): 214-224.

18. Lesion segmentation for disease diagnosis
    Reference papers:

    - Kamnitsas, Konstantinos, et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation." Medical image analysis 36 (2017): 61-78.
    - Brosch, Tom, et al. "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation." IEEE transactions on medical imaging 35.5 (2016): 1229-1239.

19. Cell classification
    Reference paper: Sirinukunwattana, Korsuk, et al. "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images." IEEE transactions on medical imaging 35.5 (2016): 1196-1206.

20. Peptide classification
    https://www.nature.com/articles/srep22843
    https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-016-1103-6
    https://www.nature.com/articles/srep12512

21. Protein localization in subcellular structures
    https://academic.oup.com/bioinformatics/article-abstract/33/16/2464/3603546/SubCons-a-new-ensemble-method-for-improved-human?redirectedFrom=fulltext
    https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1699-4

22. Cancer stage detection
    https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3604-y
    http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161501
    https://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-8-S6-S2

23. More projects in medical domain including segmentation, disease diagnosis, cancer imaging, reconstruction can be looked at https://grand-challenge.org/
    All_Challenges/ for the challenges in medical domain.