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**: Oct. 6, 2017, - Interim Report I (CVPR format - 3 pg. limit)
final submission: Nov. 10, 2016 - Final Report (CVPR format - 4 pg. limit)
Presentations: Nov. 18 (tentative)

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.

Example Projects

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 multi-modality 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.