PROJECT DESCRIPTION

The project is an essential part of this class. It will allow you to demonstrate your Machine Learning (ML) skills and create something that you are proud of. It can also be a valuable addition to your projects portfolio that you can demonstrate to prospective employers.

Project Requirements

For the project, you have to perform a thorough processing and analysis of a dataset using ML techniques. Some of the requirements of the project are:

- The datasets should be chosen from a standard repository, such as Kaggle competitions, KDD cup competitions. If you are not sure, please consult the instructor or the TA.

- You should apply multiple techniques and algorithms to the same dataset, and also compare their performance. In the end, you should identify which is your strongest technique and use that as your competition entry.

- You should use "strong" or "powerful" learners. Examples could be:
  - Deep Learning techniques
  - Ensemble Learning techniques, for example boosting or random forests
  - SVM with non-linear kernels
  - Recent ML libraries such as

    Spark MLLib: http://spark.apache.org/docs/latest/mllib-guide.html
    Flink: https://flink.apache.org/news/2015/06/24/announcing-apache-flink-0.9.0-release.html
    Storm: http://storm.apache.org/
    GO language: http://www.datasciencecentral.com/profiles/blogs/machine-learning-libraries-in-go-language-3

- Your results should be strong enough in terms of accuracy and other evaluation metrics e.g. ROC curve, area under ROC curve, and this will be one of the criteria for grades. Note that just using accuracy as the evaluation criteria is not sufficient.

- You should create a well formatted project **report** that should cover the following sections:
  - Introduction and problem description,
  - Related work
  - Dataset description (including features, attributes, etc)

    - Pre-processing techniques
    - Your proposed solution, and methods [This section should have enough details – both theoretical, and practical]
    - Experimental results and analysis [Details are expected]
    - Conclusion
    - Contribution of team members
    - References

An excellent example of what to include in such a report can be found here: http://www.cs.utexas.edu/~mooney/cs391L/paper-template.html

Some examples of excellent reports can be found at: (Note: You cannot choose these project topics)
http://cs229.stanford.edu/projects2015.html
http://cs229.stanford.edu/projects2014.html
http://cs229.stanford.edu/projects2013.html

All contents of your report must be original. You cannot copy sentences, paragraphs, figures, or anything else from outside sources. As a graduate student, you are expected to work with maturity and diligence.
Again, your report will be checked for plagiarism. Any violation will carry strong penalties, including reporting the incident to university authorities.

- Team size requirements: Project can be done in teams of 1 to 4 students. More than 4 students cannot be in a team under any circumstances. You can only form team within the same class and section. You are not allowed to work or collaborate with students from other sections of this class.

- Project selections should be unique, which means that two teams cannot work on the exact same problem. Please do not request exceptions to this rule.

- ***Projects will be assigned on first come first serve basis*.** After selecting you project, please be sure to fill out your details here: https://goo.gl/forms/s9c4pjdKr4ZBgIL72

- The final project report is due at midnight Friday December 1. Project demos and presentations will be required in front of the TA during the first week of December, most likely between December 4 to 6. These are strict deadlines.

Project Ideas

Below are some of the project ideas. You can choose any one of them. Note that for the data science competitions, you have multiple options. You are free to choose any active competition, but you will have to follow the requirements completely. You cannot pick and choose which requirements you will satisfy.

Note: Two teams cannot work on the exact same topic. Projects will be assigned on a first come first serve basis.

1. Participate in the Yelp dataset challenge and submit a good entry:

http://www.yelp.com/dataset_challenge

2. Take part in an **active** Kaggle competition that involves significant amount of Machine Learning technologies

https://www.kaggle.com/competitions

3. Take part in the KDD 2017 cup. Details are available at:
http://www.kdd.org/kdd2017/announcements/view/announcing-kdd-cup-2017-highway-tollgates-traffic-flow-prediction

The competition website where datasets and other details are available is at:
https://tianchi.aliyun.com/competition/information.htm?spm=5176.100067.5678.2.8CnCPt&raceId=231597

4. Take part in a previous KDD cup challenge

http://www.kdd.org/kdd-cup

You can take part in any previous year's cup.

5. Take part in an **active** Driven Data competition.

https://www.drivendata.org/

6. Machine learning based analysis of stock market investing techniques

Ideas:

- Simulation of systematic trading techniques, such as backtesting
https://en.wikipedia.org/wiki/Technical_analysis#Systematic_trading
- Simulation and analysis of backtesting using R packages such as backtest, PerformanceAnalytics, quantmod, etc

7. Take part in the thinkorswim challenge:
https://www.thinkorswimchallenge.com

Note: This is a financial data challenge and requires some knowledge of finance and the stock market.

8. Take part in a competition from KDnuggets
https://www.kdnuggets.com/competitions/

9. Take part in a competition from Innocentive
https://www.innocentive.com/

10. Take part in a competition from TunedIT
http://tunedit.org/

Deliverables and Deadlines

Deadline	Project Phase	Deliverable
Wednesday Oct 25 Midnight	Project Selection Team Formation	Submit your details on Google Forms https://goo.gl/forms/s9c4pjdKr4ZBgIL72 Please check for instructor's comments and approval at: https://goo.gl/iZy9md
Friday Nov 10 Midnight	Project Status Report	Submit a report containing following on eLearning: • Dataset details, such as number of features, instances, data distribution • Techniques you plan to use • Experimental methodology (how you plan to pre-process, create training, validation, and test datasets, and other such details) • Coding language / technique to be used • Preliminary Results (if available)
Friday December 1 Midnight	Final Report	Submit final documents on eLearning: • Detailed Final Project Report • Code • README file indicating how to run your code ** Your report and code will be checked for plagiarism **