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 analysis of one or more datasets 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 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, and this will be one of the criteria for grades.

- 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:
http://cs229.stanford.edu/projects2015.html
http://cs229.stanford.edu/projects2014.html
http://cs229.stanford.edu/projects2013.html

Your report will be checked for plagiarism. Do not lift sentences from other sources.

- Team size requirements: Project can be done in teams of 2 to 5 students.

- Project selections should be unique, which means that two teams cannot work on the exact same problem. It's acceptable to work on the same dataset, but techniques must be different.

- *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/boVcqO0j2wZmJpSm1

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.

Below are some suggested topics.
Note: Two teams can not 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 cup challenge

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

You can take part in a previous year's cup also.

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

https://www.drivendata.org/

5. Build a recommender system on DBLP's conference/publications dataset:

http://dblp.uni-trier.de/xml/

Ideas:

- DBLP stores details of publications of authors in journals and conferences.
- You will build a recommender system on author-conference, author-journal, author-title keyword, author-author. That is, recommend how likely is an author to publish at a top conference, or prestigious journal.

6. [Bioinformatics] Create a Machine Learning approach for next generation sequence comparison and analysis.

References:
http://bioinformatics.oxfordjournals.org/content/early/2013/10/01/bioinformatics.btt528.full
http://www.osti.gov/scitech/servlets/purl/1050659

7. 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 of backtesting using R packages such as backtest, PerformanceAnalytics, quantmod, etc

8. Building a strong recommender system for the IMDB movie dataset:

- IMDB movie dataset
http://www.imdb.com/interfaces

Deliverables and Deadlines

Deadline	Project Phase	Deliverable
Sunday October 23 Midnight	Project Selection Team Formation	Submit your details on Google Forms https://goo.gl/forms/boVcqO0j2wZmJpSm1
Sunday November 6 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 • Coding language / technique to be used • Preliminary Results (if available)
Sunday November 27 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 **