Introduction

The final project aims to provide you with an opportunity to apply the machine learning algorithms (not restricted to the ones covered in class) to interesting real-world learning problems. In a supervised learning framework, given $X := \{x_1, ..., x_m\}$ with corresponding labels $Y := \{y_1, ..., y_m\}$, where $y_i \in \{\pm 1\}$ for $i = 1, \dots, m$. We seek to infer a function $g : \mathcal{X} \to \{\pm 1\}$ to predict accurately whether a new observation will belong to class $+1$ or -1. In general, g - could be any learning machine defined on any learning from data problem (supervised, semi-supervised, unsupervised or active, reinforcement and many other types), since the core-foundational knowledge we studied in this course via VC theory of generalization and regularization techniques applies to any g.

This project is a chance for you and your team to optimize a robust machine learning model (i.e., come up with a new angle on an old problem). Successful implementation of benchmark datasets has the potential to become full-fledged research papers. Every one of you can go way above and beyond the state-of-the-art methods.

Important Dates

- Project proposal due date: **November 22, 2021**
- Final Report and Video Presentation date: **December 4, 2021**

Method of Delivery

Assignment deliverables should be submitted via Moodle to the course instructor before the due date.

Deliverables

You can use the IEEE template on overleaf:
https://www.overleaf.com/latex/templates/ieee-journal-paper-template/jbbbdkztwxrd

- **Project Proposal**
  A single page project proposal that includes the following information (submitted to Moodle)
  - Project Title.
  - The project idea (or problem definition).
  - Data set.
  - Software package that you will use.
  - Team members - up to three in each team, and their expected contributions.
  - Review and include relevant literature (1-3 research papers).

1. **Report describing in detail the work of a team with the following sections (use the Latex TEMPLATE provided in the moodle; - length 4-6 pages long)**
   - Abstract
   - Introduction
   - Methods
   - Results
   - Conclusion
   - References
   - Contribution (what and how each member contributed to the project) .
2. **Pre-recorded video presentation - (20-30 minutes length)**
3. **Source Codes (Jupyter Notebooks) + Link to Data sets** and a file named *README*, and include in it a short description of all codes files you are submitting.

Level of Collaboration Allowed

- This is a team project. You can form a team of three students at maximum for this assignment. Discussions on course materials and implementation of the project are encouraged.
- The external resources can be consulted but not copied from.
- It is expected that you discuss, work, and learn together. You may use ScikitLearn and Pytorch (or other) tools to implement your final projects.

Grading Criteria

- 40% - Implementation (well documented source code)
- 10% - Benchmark - Accuracy (i.e. comparison with the best performance in the literature)
- 20% - overall work and report quality
- 15% - discussion (for example of success/failure; limitations, etc.)
- 15% - video presentation

Machine learning tools

- Since the implementation of standard algorithms from scratch may take up a significant amount of time, in this project, you are encouraged to use available machine learning tools/libraries and concentrate on model selection problem by applying an algorithm of your interest for a real-world problem.
- However, there is no restriction if you can manage your time and want to implement a novel algorithm from scratch. Keep in mind that the deadline is the final one without any further extension.

There are many machine learning tools that provides optimized implementation of many learning algorithms including the ones listed below, which you can to achieve your final project goals:

1. **Scikit-Learn**: machine learning in Python (scikit-learn.org).
2. **Pytorch**: deep learning library in Python
3. **TensorFlow**: TensorFlow is an open source software library for numerical computation using data flow graphs for building deep learning models
- **Tutorial**: https://www.youtube.com/watch?v=dYhrCUFN0eM.

4. **Google Colaboratory** https://colab.research.google.com

1 Specific Tasks

- **Identify** data from domain of your interest, e.g., natural language processing, computer vision problems, text information retrieval, brain data analysis etc.
- **Perform model selection** to estimate the model with optimal hyper/ parameters that solves the problem of your chosen domain. We can group under **model selection** a number of problems, including:
    1. selecting the best features,
    2. selecting the best preprocessing (data normalization, mathematical transformations of feature space)
    3. selecting the best learning machine (e.g., neural network, deep learning architecture, linear model, kernel method, classification or regression algorithms, etc.),
    4. selecting the best set of hyperparameters (number of layers and hidden units in a (deep) neural network, kernel type, and regularization parameter in kernel methods, etc.).
    5. Implementing a cross validation (CV) strategy (standard 10-fold CV or nested CV, leave-one out CV and/or others)
- Different learning machines exist in literature such as linear models, neural networks, deep learning, convolutional networks and/or kernel methods, etc., **you can adapt and apply** an algorithm of your interest and compare with other at least three ways try to outperform them in terms of generalization performance. Explain what you have done to combat overfitting due to deterministic and stochastic noise.
    - For instance, you decided to work on a handwritten digit recognition task, and you chose as your main algorithms a deep neural network (DNN) model.
    - Then, you are expected to optimize the DNN via your model selection strategy and compare its performance with the other standard algorithms (at least three), i.e., linear soft-margin SVMs, Kernel soft-margin SVMs, Regularized Logistic Regression, Regularized Linear Regression.

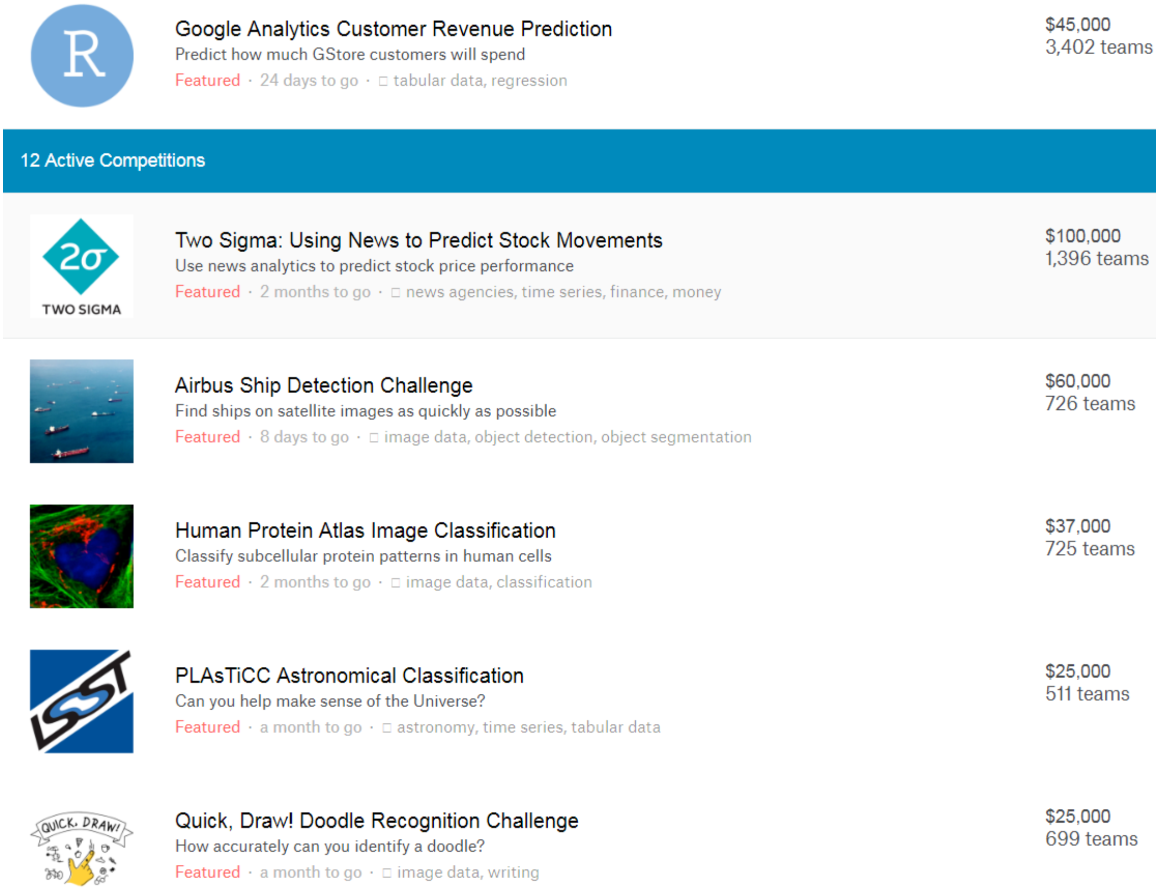
Try to follow the following steps in your project:

1. Consider a dataset D (from any domain of your interest)
2. **Apply an algorithm of your choice on** D
3. Estimate its generalization error (E_{test})
4. **If**: generalization error smaller than what exists in the literature for the same dataset:
    - **End of the process: Outcome → Grade A:)**
5. **Else**:
    - Go back to step 2 with another algorithm or change the model selection strategy.

Project Databases and Ideas

Determining a domain of interest is one of the important task in this project. You have to spend time with your team members to explore internet databases dedicated for machine learning problems. For instance,

- For computer vision related datasets refer to : http://www.cvpapers.com/datasets.html
- Another popular and diverse data sets can be accessed in UC Irvine Machine Learning Repository : http://archive.ics.uci.edu/ml/
- Similarly, check Amazon- dataset repository https://aws.amazon.com/datasets/
- 19 Free Public Data Sets For Your First Data Science Project https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- Review machine learning competitions and dataset at **www.kaggle.com**
  - ↠ You can try to participate at any current competition posted at kaggle.com and use deep learning algorithms. For instance take a look at the competitions:
    - ∗ https://www.kaggle.com/competitions



Google Analytics Customer Revenue Prediction
Predict how much GStore customers will spend
Featured · 24 days to go · □ tabular data, regression
$45,000
3,402 teams

12 Active Competitions

Two Sigma: Using News to Predict Stock Movements
Use news analytics to predict stock price performance
Featured · 2 months to go · □ news agencies, time series, finance, money
$100,000
1,396 teams

Airbus Ship Detection Challenge
Find ships on satellite images as quickly as possible
Featured · 8 days to go · □ image data, object detection, object segmentation
$60,000
726 teams

Human Protein Atlas Image Classification
Classify subcellular protein patterns in human cells
Featured · 2 months to go · □ image data, classification
$37,000
725 teams

PLAsTiCC Astronomical Classification
Can you help make sense of the Universe?
Featured · a month to go · □ astronomy, time series, tabular data
$25,000
511 teams

Quick, Draw! Doodle Recognition Challenge
How accurately can you identify a doodle?
Featured · a month to go · □ image data, writing
$25,000
699 teams

  - Decoding the Human Brain
    - ∗ https://www.kaggle.com/c/decoding-the-human-brain

    - https://www.kaggle.com/c/grasp-and-lift-eeg-detection
  - Develop a Gesture Recognizer for Microsoft Kinect
    - https://www.kaggle.com/c/GestureChallenge
  - The CIFAR-10 dataset

    https://en.wikipedia.org/wiki/CIFAR-10

If you have any questions regarding this project/data, please contact me.