CS461 Machine Learning
Course Project Guidelines

Overview:

Your class project is an opportunity for you to explore an interesting machine learning problem of your choice in the context of a real-world data set. The purpose of the project must be related to some aspect of the material, but may explore an avenue that was left unaddressed in class.

You are expected to perform the following tasks:

- Identify a specific problem in your chosen topic in Machine Learning.
- Summarize a few recent solutions that tackled your problem of interest.
- Identify a question related to your problem of interest and form a plan to answer the
- question by implementing a program.
- Implement the program and produce results.
- Discuss your findings and future direction

The projects can be done individually, or in teams of two students.:

1. **Proposal**: (15/11/2025) at most two-pages, 12-point font, single spacing, 1 inch margins.
2. **Midway Report**: (TBA)
3. **Final Report**: (18/12/2025)
4. **Presentation**: (TBA)

Note that all write-ups (Midway and final report) must be in the form of IEEE conference paper.

Project Proposal

In order to help guide your choice of a project, you are required to submit a brief proposal (at most one-page, 12-point font, single spacing, 1 inch margins) that describes the idea for a project. The proposal should be one page maximum. Include the following information:

- Project title
- Introduction. What is the problem tackled?– Motivation
  - This should identify the project type, the problem you plan to address, the motivation for why you find the problem important or interesting, any previous work you already know about (if applicable), and a rough tentative approach to solving the problem.

- Method: Explain the proposed solution at a high level.
- Data set. Describe the dataset, such as how many columns, rows, the source of the dataset, etc.
- Midterm milestone: A table shows dates, and what will you complete by end of the semester?

Note: If you are having trouble writing a proposal, feel free to contact the instructor (Dr. Ali Aburas).

Midway Report

This should be a 4-5 pages short report in the form of a IEEE conference paper, and it serves as a check-point. It should consist of the same sections as your **final report** (background, method, experiment, conclusion and references). The introduction and related work sections should be in their almost final form; the section on the proposed method should be almost finished; the sections on the experiments and conclusions will have whatever results you have obtained, as well as “place-holders” for additional results you plan/hope to obtain.

In the Result section (the core of the project)– Interesting challenges in your implementations– showcase the findings from your data-set– Performance analysis– Anything else interesting you want to share.

Grading scheme for the project report:

- 70% for proposed method and some experiments results so far
- 30% for the design of upcoming experiments

Poster Presentation

All project members should present during the presentation hours. The session will be open to everybody (**if applicable**).

You can create a bunch of "normal" presentation slides, print out each one on a piece of (letter-sized) paper, and put them all together on a poster board. I will provide a template if you need it.

Your Project Must Apply at least 3 of the following

1. ***Comparison of algorithms***: Throughout the course, we've been discussing various algorithms and their properties. Oftentimes, algorithms don't work like expected and algorithms may need to be adapted or modified to better fit the assumptions inherent in the data. What work needs to be done to adapt a model to an interesting set of data that you've found? How do various algorithms perform on the same set of data?
2. ***Missing information***: Various real world classification problems involve missing components in the input vectors. How can you deal with such missing information? Do you expect your method to degrade rapidly if more information is missing?
3. ***The Hyperparameter Selection***: Machine learning algorithms automatically adjust (learn) their internal parameters based on data. However, there is a subset of parameters that is not learned and that has to be manually configured. The performance of a model can depend on the choice of its hyperparameters. Four commonly used optimization strategies: Grid search, Random search, Hill climbing, and Bayesian optimization.
4. **Approaches to Solve Imbalanced Classes**. Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally. There are systematic algorithms that you can use to generate synthetic samples. A simple way to generate synthetic samples is to randomly sample the attributes from instances in the minority class. It consists of removing samples from the majority class (**under-sampling**) and/or adding more examples from the minority class (**over-sampling**). The most popular of such algorithms is called **SMOTE** or the Synthetic Minority Over-sampling Technique.
5. **Ensemble techniques**: This approach leverages the idea that a group of models can produce a more reliable result than a single model alone, which can suffer from issues like overfitting or bias. Popular methods include bagging (like Random Forests), boosting (like AdaBoost and Gradient Boosting), and stacking, which combine models in different ways to reduce error and variance.
6. **Feature Scaling**: Feature scaling, like normalization and standardization, is crucial for algorithms sensitive to feature magnitude, while ML scalability ensures a system can grow in capacity.
7. Any other topics that we covered in the class.

Note: You shouldn't worry about getting "great" results. The idea and your understanding of the machine learning issues involved are much more important than getting "great" results.

Some data repositories you might find useful:

the UC Irvine Machine Learning Repository — https://archive.ics.uci.edu/
KDD Repository (Various) — http://kdd.ics.uci.edu/
Protein data bank (Genome) — https://www.rcsb.org/
Protein structural database (Genome) — http://scop.mrc-lmb.cam.ac.uk/scop/
Cancer classification data (Medical) — https://www.kaggle.com/code/shubhankartiwari/cancer-classification
Newsgroups (Text) — http://www.ai.mit.edu/people/jrennie/20 newsgroups/
Reuters Documents (Text) — https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
4 Universities (Text) — http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/
kaggle —https://www.kaggle.com/