COMP 4220 Machine Learning Final Project, Fall 2021
Due date: December 17, 11pm

This project is an opportunity for you to explore ideas that you see in the lectures, assignments, and other resources. You can think of your project as the first step towards doing research in machine learning. You will analyze the problem, design a machine learning solution, implement learning algorithms, and evaluate them on two data sets (**one for classification and one for regression**). Please review the full list of evaluation metrics at https://scikit-learn.org/stable/modules/model_evaluation.html. In this project, you are asked to (1) develop models that attain high accuracy and (2) explain the performance of the trained models. For example, let's say you designed a classifier with 99% accuracy. An important task is to investigate whether the studied data set is too simple, or the evaluation metric is not appropriately capturing the classifier performance (e.g., we discussed precision, recall, and ROC curves for classification problems). Recall that we should always use training, validation, and test data sets to properly evaluate the performance of machine learning algorithms to avoid underfitting and overfitting.

For each data set, you should use **three** machine learning algorithms. For example, for solving a regression problem, you may use linear regression, polynomial regression, regularized regression (either linear or polynomial). In this case, you should discuss how you tuned the regularization parameter with some evidence. For developing classifiers, we covered a variety of classification algorithms, such as logistic regression, support vector machines, decision trees, random forest, and neural network models. Also, you should think of utilizing preprocessing techniques, such as centering and scaling, to improve the performance of learning algorithms.

Milestones and Grading

The project is worth 20% of the class grade. The final report should be **five pages**. The report should be structured like a small research paper. Broadly speaking it should describe:

- What are the important ideas/methods you explored?
- Preprocessing techniques?
- Reporting the results (cross-validation and easy-to-read figures).
- Do the results make sense? Underfitting? Overfitting?
- Explain the behavior of models (e.g., does the model outperform a random classifier?)
- Please include the complete execution code to produce the reported results at the end of your report. (No page limit)

You will be assessed on the effort, the clarity of explanations, the evidence that you present to support your claims, and the performance of machine learning methods.

Data Sets

- Classification
  - Credit Card Fraud Detection
  - https://www.kaggle.com/mlg-ulb/creditcardfraud?select=creditcard.csv
  - Required to use the imbalanced-learn package: https://imbalanced-learn.org/stable/index.html (tutorial available at https://youtu.be/583SOiipnQ4)
- Regression
  - Energy Efficiency Data Set
  - https://code.datasciencedojo.com/datasciencedojo/datasets/tree/master/Energy%20Efficiency

An example of the final report from previous offerings of this course has been provided. Please note that requirements change every semester, and the attached document solely serves as an example.