CS 422/622
INTRODUCTION TO MACHINE LEARNING

FALL 2022
Assignment #4

Due Date/Time: **11/14/2022 @ 11:59PM**
Total Points: **100**

CS 422 students may complete this assignment individually or in teams of two; CS 622 students are to complete it individually.

Description:

- For this assignment, you will implement two ***linear regression algorithms*** in Python or MATLAB to solve a ***regression problem***.
- ***Ordinary least squares (OLS) solution:*** Implement the ordinary least squares solution ***from scratch*** as discussed in class.
- ***Linear regression with gradient descent:*** Implement gradient descent for linear regression as discussed in class. You may use all built-in functions. The built-in linear regression models in Python and MATLAB that can be used are:
  - Python: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html
  - MATLAB: https://www.mathworks.com/help/stats/incrementalregressionlinear.html with the "learner" parameter set to 'leastsquares'
- For both algorithms, don't forget to include the bias term in the parameter vector.
- Train and test your model with a dataset of your choosing that meets the following criteria:
  - Number of input features: 3+
  - Input characteristics: Continuous real-valued
  - Output characteristics: Continuous real-valued

  Note: If you find a dataset that has categorical input features, you may remove them as long as you end up with 3+ continuous input features.
- Use 80% of the dataset for training and 20% for testing your model.
- If you'd like, you may use one of the following sources to find a dataset:
  - University of California, Irvine Machine Learning Repository https://archive.ics.uci.edu/ml/index.php
  - Kaggle https://www.kaggle.com/
  - Awesome Public Datasets https://github.com/awesomedata/awesome-public-datasets
  - Google Dataset Search Engine https://datasetsearch.research.google.com/

- Microsoft Research Open Data https://msropendata.com/
- U.S. Government's Open Data https://www.data.gov/
- Registry of Research Data Repositories https://www.re3data.org/
- CMU Libraries https://guides.library.cmu.edu/machine-learning/datasets

- Summarize your approach and results in a report that includes at least the following:
  - Name(s) of the student(s) completing the assignment
  - The dataset you used, its source and characteristics.
  - The data preprocessing steps you took (if any).
  - The solution $\boldsymbol{w}$ (parameter vector) for both the OLS and the linear regression with gradient descent algorithms. This vector should include the intercept (bias term).
  - Relevant evaluation metrics for OLS (MSE, MAE, R^2) for BOTH the training and test datasets.
  - Relevant evaluation metrics for linear regression with gradient descent (MSE, MAE, R^2) for BOTH the training and test datasets.
  - Any additional details you would like to include.
- Submit your report along with your dataset and source code. Feel free to include your code in the report, but you also need to submit your source code files (.py or .m) and your dataset separately, so that your results can be replicated for grading.

Submission Instructions:

Compress all files and submit it through Canvas as a **.zip** file.

If you are completing the assignment as a team of two, only one of the team members needs to submit the assignment.

I will set Canvas to allow unlimited number of submissions and will only grade the last submission. So, please do not wait until the last minute to submit as you can always submit an updated version.