Tutorial Assignment I - Regression

Instructions: Please read the below instructions carefully to complete this assignment.

1. This assignment has two parts (Part I and Part II) and will be graded out of 25 points.
2. Please work on this assignment individually.
3. Please submit all your code in two python (*.py) files; *one for each part*. Name these files as <*your_mac_id*>_part_I.py and <*your_mac_id*>_part_II.py.
4. Submit these files on Avenue.
5. Please DO NOT upload ANY datasets. The datasets you upload will NOT be used for assignment evaluation.
6. Please import only the below libraries in your code. The assignment will NOT be graded if any other libraries are used besides the ones listed below.

```
import os
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sn
import random
```

7. The coding assignment will be evaluated on both **functionality** and **code quality**.
    a. Functionality refers to correctness of solution, sanity checks, stress tests, robustness, catching edge cases, comprehensiveness in code architecture, conciseness, and performance on the held-out test set (where applicable).
    b. Code quality refers to cleanliness/ readability, well-chosen data structures, program design for optimal efficiency, and clear and meaningful comments existing for *at least* 75% of the lines of code.

8. Please do not hard code any aspects of your model. The model performance will be evaluated by passing the test csv file to your code. This means, your code should be able take any CSV file as an input with the same feature vector as your training data and produce a graph.

9. Please do not use AI to generate your code. For this class, it will count as academic integrity violation. While AI is good at generating text, it is still not good at producing quality code for building ML models. I have observed several students make fundamental mistakes on their ML models because they prompted AI to complete their task. *Most importantly, to know whether AI has given you the correct solution, you should be able to know what the correct solution is*. Using AI to learn how to build AI is a dystopian idea that will harm your learning and mastering AI.

Part I (Max 10 points)

Context: Understanding what factors lead to a happy life are important for building meaningful societal structures. This understanding not only helps us prioritize factors that are important, but also tackle challenges that get into our way of achieving high life satisfaction.

Challenge: In the class, we studied gradient descent and how linear regression can be implemented using gradient descent. For this part of the assignment, you will implement a linear regression model using gradient descent to find the relationship between happiness and richness using the below dataset.

Dataset: Please use the dataset uploaded in Week 1 folder in Lectures on Avenue. The dataset has been downloaded from this source and preprocessed using the code given in Linear_Regression.py (uploaded in the same folder).

Goal: Build a Linear Regression model that can reliably model the relationship between happiness scores of a country and their GDP (a measure of richness of the country) in 2018. Therefore, your target variable is "happiness", and your feature vector comprises of the GDP. Implement the Gradient Descent (GD) based Linear Regression model and report the learned β' values. Plot a line to visually show how β' values from your GD implementation generate a line that is a "good" fit to the dataset. Compare this line with the line derived using OLS method in class (refer to the lecture slides). For Gradient Descent, experiment with 5 different learning rates and 5 different iteration counts. Select your best β' value derived from all your experiments.

What should be the output of your code: Your code must produce the following outputs.

1. A graph with 4-8 different regression lines plotted using β' values derived from the GD approach and experiments with *different learning rates and iteration counts* (also referred to as "epochs"). These lines must be overlayed on the dataset scatterplot (like the one shown in class) to show "goodness" of fit. The code should also print all β' values along with their corresponding epochs and learning rates.
2. A graph with 2 plotted lines; the first line plotted using β' learned through OLS (code shared in class) and second line is your *best* β' value learned through GD. Here you select your best β' value from your experiments. The code should also print both β' values, along with learning rate and epoch you selected for GD approach.

Part II (Max 15 points)

Context: Abalone is a type of marine snail found in sheltered bays, exposed coastlines, shallow sea waters, and even in freshwaters. In Canada, the more dominant specie is Northern Abalone popularly harvested in areas of British Columbia, and along the coastal regions of North America, stretching all the way to Alaska. As soon as Abalone offsprings mature, they move to shallow waters, where harvesters can easily collect them. Abalone belongs to marine Mollusca family and is highly valued for its meat. This has led to significant illegal harvesting. The price of an abalone is anywhere between CAD 28 to CAD 45 per kilogram and directly depends on its age.

Challenge: To determine the age of abalone, the shell is first cut through the cone, stained and the number of rings are counted through a microscope. The rings indicate the age of abalone. The process is very time-consuming, prone to error and very tedious. One way to lessen the human workload is by building an ML model that can predict the age of abalone using factors such as its height, weight and shape. This is precisely what you will accomplish in this tutorial assignment.

Dataset: The dataset has been downloaded from UCI Machine Learning Repository, a popular database for ML research. The dataset provided to you in the file *training_data.csv* (uploaded in the folder) has following modifications from the original:

1. The column containing information about the Sex of abalone has been dropped.
2. The dataset has been divided into 5 subsets of 2000, 500, 500, 500 and 577 samples (100 samples were dropped). You have been given 2577 samples to train and test your ML model (2 sets). The remaining 3 sets have been held out. Your model performance will be evaluated on 1 of the 3 held out test sets (same set used for all submissions).

Goal: Build a Linear or Polynomial Regression model that can reliably predict the age of abalone using features of Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight and Shell weight. The age is computed by counting the rings of abalone (adding 1.5 to rings give the age in years). Therefore, your target variable is “rings”, and your feature vector comprises of the above listed 7 features.

Guidelines: The following steps may be taken to solve the problem.

1. First visualize all features and their relationship to age of the abalone.
2. See what kind of relationship might exist and select a model (linear or polynomial)
3. Select your cost function (RMSE / MSE / MAE)
4. Select your approach (GD or OLS)
5. Select a subset from 2577 to train your model on.
6. Train your model
7. Evaluate the performance of your model.
8. Report your β' values (printing them is fine).
9. Visualize your fit using your β' values by overlaying the line on the charts you created in step 1.