CS/SE 4AL3 Homework 1: Regression

Fall 2025
Due: September 24

1 Overview

This assignment will get you started working with basic machine learning classifiers. You will build regression models to predict happiness for part 1 and abalone age in part 2.

1.1 Instructions

This assignment has two parts and will be graded out of 25 points. You must work on this assignment individually.

You may use ONLY the following libraries in your code. You will receive a zero if any other libraries are used besides the ones listed below.

Listing 1: Available Packages

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

Your code should be written in Python and should include enough documentation/instructions for someone else to be able to run. Your code should include a simple README file that explains the files/directories, and how to set up and run the code. You should specify the Python version that you are using and make sure that your file read/write operations use UTF-8 (this makes it much easier for us to run your code).

You are welcome to use code snippets from examples in class, things you find online, or from AI code generation tools. Just make sure to give proper attribution to code you did not write. Follow the syllabus instructions for how to report the use of AI tools. However, you may not copy code that does the entire assignment (e.g. someone who did this assignment in a previous semester).

DO NOT hard code any aspects of your model. Performance will be evaluated by passing the test CSV to your code. This means that your code should be able to take any CSV file as input with the same feature vector as your training data and output a graph.

2 Part 1

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: We studied gradient descent and how linear regression can be implemented. You will implement a linear regression model with gradient descent to find the relationship between happiness and richness using the below dataset.

Dataset: The dataset from 2018 has been preprocessed using the provided Linear_Regression.py file and is provided with the assignment.

Goal: Build a linear regression model that can reliably model the relationship between happiness scores of a country and their GDP. Your target variable is *happiness*, and your feature vector is comprised of the GDPs for each country. Implement the gradient descent based model and report the learned β' values. Plot a line to visually show how β' values from your gradient descent implementation generate a line that is a good fit to the dataset. Compare this line with the line derived using the OLS method in class. For gradient descent, experiment with five different learning rates and five different iteration counts. Select your best β' value derived from all of your experiments.

Output: Your code should produce the following outputs:

1. A graph with 4-8 different regression lines plotted using β' values derived from gradient descent and experiments with different learning rates and iteration counts (i.e. epochs). These lines must be overlayed on the data scatterplot (like the one shown in class) to show how well the line fits the data. The code should also print all β' values along with their corresponding epochs and learning rates.
2. A second graph with two lines. The first line should use the β' learned through OLS (code from class) and the second line as your best β' learned through gradient descent. Here you select your best β' value from your experiments. The code should also print both β' values, along with the learning rate and epoch you selected for your gradient descent approach.

3 Part 2

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 species is Norther 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 offspring are mature, they move to shallow waters, where harvesters can easily collect them. Abalone belongs to the marine Mollusca family and is highly valued for it's meat. This has led to significant illegal harvesting. The price of an abalone is anywhere between 28 and 45 CAD per kilogram and directly depends on it's 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 the abalone. The process is very time consuming, tedious, and prone to error. One way to lessen the human workload is by building a machine learning model to predict the age using factors such as height, weight, and shape. This is what you will accomplish for this assignment.

Dataset: The dataset has been downloaded from the UCI Machine Learning Repository (a popular and potentially useful database for your later projects). The dataset provided to you in the file *training_data.csv* has the following modifications from the original:

1. The column containing information about the sex has been dropped
2. The dataset has been divided into 5 subsets of 2000, 500, 500, 500 and 577 samples (100 were dropped). You have been given 2577 samples to train and test your model (2 sets). The remaining three sets have been held out. Your model performance will be evaluated on 1 of the 3 held out test sets (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 the rings to give the age in years). Therefore, your target variable is *rings*, and your feature vector has the seven features listed above.

Guidelines: The following steps will help you solve the problem

1. You should visualize your data by first plotting each feature and it's relation ship to abalone age. Take note of the relationship and if it is linear or polynomial.

2. Select your cost function (RMSE/MSE/MAE)
3. Choose your approach (gradient descent or OLS)
4. Train your model using a subset of the data and evaluate the data on a held out set

Output: Similar to part 1, your could should produce a plot for each feature and it's relationship to the target value along with the overlay of the line of best fit for each chart. Your script should print the β' values.

4 Deliverables

You should submit the code you used as well as a README.txt file to Avenue. You should include any instructions to run your code in the README.txt file and any disclaimers about the use of AI. Please submit all code in two python (*.py) files. Name these files <your_mac_id>_part1.py and <your_mac_id>_part2.py. Submit these files on Avenue. Do not upload any datasets.

5 Help

Please use the Teams channel and tutorials to ask questions about the assignment when you need guidance or pointers on this homework. You are free to discuss your approach and ideas with classmates, but should not share code or reuse data. You may use generative AI tools if you find them helpful, but please clearly document how they were used in the report and follow the guidelines in the syllabus for what you **must** include when using generative AI. If you use generative AI and do not report it, you may receive a 0 for the assignment. You take full responsibility for the deliverables you submit.