Cairo University
Faculty of Computers and Artificial Intelligence
Machine Learning Course

Assignment 1: Linear and Logistic Regression

Most vehicles on the road today use petrol or diesel fuel engines. When these fuels burn, they release energy that powers the vehicle. However, this process also produces waste in the form of carbon dioxide gas (CO_2). The amount of CO_2 emission is a basic indication of a vehicle's impact on the environment and air quality. So, with climate change at the forefront of global issues, it is crucial to understand and predict the vehicles' CO_2 emission amounts.

In this assignment, you are required to build a linear regression model and a logistic regression model to predict the amounts of CO_2 emission and their respective classes based on the vehicles' data.

Dataset:

The attached dataset **"co2_emissions_data.csv"** contains over 7000 records of vehicles' data with 11 feature columns in addition to 2 target columns. The features are: the vehicle's make, model, size, engine size, number of cylinders, transmission type, fuel type, and some fuel consumption ratings columns. The targets are the CO_2 emission amount (in g/km) and the emission class.

Note: This dataset is a modified version of the "CO2 Emission by Vehicles" dataset. The original dataset was obtained from Kaggle.

Requirements:

Write a Python program in which you do the following:

a) Load the **"co2_emissions_data.csv"** dataset.

b) **Perform analysis** on the dataset to:
   i) check whether there are missing values
   ii) check whether numeric features have the same scale
   iii) visualize a pairplot in which diagonal subplots are histograms
   iv) visualize a correlation heatmap between numeric columns

c) **Preprocess** the data such that:
   i) the features and targets are separated
   ii) categorical features and targets are encoded
   iii) the data is shuffled and split into training and testing sets
   iv) numeric features are scaled

d) **Implement linear regression using gradient descent from scratch** to predict the CO_2 emission amount.

   -> Based on the correlation heatmap, select **two features** to be the independent variables of your model. Those two features should have a strong relationship with the target but not a strong relationship with each other (i.e. they should not be redundant).

   -> Calculate the **cost** in every iteration and illustrate (with a plot) how the error of the hypothesis function improves with every iteration of gradient descent.

   -> Evaluate the model on the test set using Scikit-learn's **R^2 score.**

e) **Fit a logistic regression model** to the data to predict the emission class.

   -> Use the **two features** that you previously used to predict the CO_2 emission amount.

   -> The logistic regression model should be a **stochastic gradient descent classifier**.

   -> Calculate the **accuracy** of the model using the test set.

Remarks:

- You can use functions from **data analysis and computing libraries** (e.g. Pandas and NumPy) as you please throughout the entire code.

- You can use **machine learning libraries** such as Scikit-learn for preprocessing and metrics **but NOT for "from scratch" requirements.**

- The **train/test split** has to be **performed before** the feature scaling step.

- **The numeric features of the test set should be scaled using the statistics of the train set that were used to scale it.**

- You should use the **R^2 score** to evaluate the linear regression model as it **provides a measure of how well observed outcomes are replicated by the model**. In general, **the best possible score is 1**, but the score can be negative as the model can be arbitrarily worse.

Deliverables:

- **You are required to submit ONE zip file containing the following:**
  - Your code (.py) file.

    If you have a (.ipynb) file, you have to save/download it as (.py) before submitting.
  - A report (.pdf) containing the team members' names and IDs, and the code of each requirement with screenshots of the output of each part.

- **The zip file MUST follow this naming convention:**
  Group_A1_ID1_ID2_ID3_ID4

Submission Remarks:

- The **maximum** number of students in a team is **5** and the **minimum is 3**.
- Team members must be from the **same lab** (or have the same TA).
- All team members must understand all parts of the code.
- **No late submission** is allowed.
- Stick to uploading **ONLY** the **required files** following the **naming convention:** Group_A1_ID1_ID2_ID3_ID4
- A **penalty** will be **imposed for violating** any of the assignment rules.
- **Cheaters will get ZERO** and no excuses will be accepted as per the **"Plagiarism Scope"** document.

Grading Criteria:

Both the code and the report must include:	
Analysis	**4 marks** (1 mark per step)
Preprocessing	**4 marks** (1 mark per step)
Linear Regression	
Selected features	**1 mark**
Gradient descent	**3 marks**
Cost function and plot	**1 mark**
R^2 score	**1 mark**
Logistic Regression	
Classifier	**3 marks**
Accuracy	**1 mark**
The total is 18 marks (will be scaled to 6 marks)	