Cairo University
Faculty of Computers & Artificial Intelligence
Machine Learning

Assignment 1

(Total 6 Marks)

➔ A certain car company is planning to manufacture and launch a new car. So, the company’s consultants need to study the factors which the pricing of cars depends on. Based on various market surveys, the consultants gathered a dataset of different types of cars across the market. They would like to know which of the car features are significant in predicting the price of a car.

➔ In addition, they would like to predict whether customers will be interested in purchasing the new car. That’s why they collected a few records of some of the company’s previous customers who either purchased a new car from the company as an upgrade or didn’t purchase a new car.

➔ You are required to build a linear regression model and a logistic regression model for this company to predict car prices and purchases based on some features.

1. Data:

There are 2 attached datasets:

- The first dataset “car_data.csv” contains 205 records of cars with 25 features per record in addition to 1 target column. These features include the car size and dimensions, the fuel system and fuel type used by the car, the engine size and type, the horsepower, etc. The final column (i.e., the target) is the car price (in some monetary unit).
- The second dataset “customer_data.csv” contains 400 records representing some of the company’s previous customers. The customer data is composed of the customer’s age and salary. The final column (i.e., the target) is a boolean value (0 if the customer didn’t purchase a new car and 1 if he/she purchased a new car).

2. Requirements:

Write 2 python programs (*2 separated .py files*) in which you work on each dataset (each model) separately as follows:

- In the first program:
    A. Load the “car_data.csv” dataset.
    B. Use scatter plots between different features (7 at least) and the car price to select 5 of the numerical features that are positively/negatively correlated to the car price (i.e., features that have a relationship with the target). These 5 features are the features that will be used in linear regression.
    C. Split the dataset into training and testing sets.
    D. Implement linear regression **from scratch** using gradient descent to optimize the parameters of the hypothesis function.
    E. Print the parameters of the hypothesis function.
    F. Calculate the cost (mean squared error) in every iteration to see how the error of the hypothesis function changes with every iteration of gradient descent.
    G. Plot the cost against the number of iterations.
    H. Use the optimized hypothesis function to make predictions on the testing set and calculate the accuracy of the final (trained) model on the test set.

- In the second program:
    A. Load the “customer_data.csv” dataset.
    B. Normalize the feature data (the customer’s age and salary) before applying regression. You can use minmax normalization where z is the normalized value and z = (x – min) / (max – min).
    C. Use scatter plot between the customer’s age and salary and differentiate between the purchased by colors (ex: red for y=0 and blue for y=1)
    D. Split the dataset into training and testing sets.
    E. Run logistic regression (from **sklearn**) to optimize the parameters of the hypothesis function. Use the 2 features (age & salary) as input and the output to be predicted is “purchased” (Do not implement the logistic regression from scratch).
    F. Print the parameters of the hypothesis function.
    G. Use the optimized hypothesis function to make predictions on the testing set.
    H. Calculate the accuracy of the final (trained) model on the test set.

3. Submission Remarks:

- Deadline will be Saturday 25 March @11:59 PM.
- The number of students in a team is 3.
- Team members can be from different labs (but they must all attend the discussion together). If anyone did not attend the discussion will take zero in the assignment. You should understand every point in your code.
- No late submission is allowed.
- You should submit one .zip file with naming convention: **ID1_ID2_ID3.zip**
  This zip file contains 2 .py file: **ID1_ID2_ID3_first.py** and **ID1_ID2_ID3_second.py** , first one contains the first program (linear regression) and second one contains the second program (logistic regression).
  You will lose 0.5 mark from the assignment grade if you did not write the correct naming convention.
- We will run a plagiarism tool to check any cheating. Cheaters will take ZERO in the assignment and no excuses will be accepted.

4. Grading Criteria:

First Program (11 marks)	
Load the dataset/Splitting the data	1
Scatter plots for feature selection	1
Linear regression	6
MSE (calculation and plot)	2
Test the hypothesis function	1
Second Program (7 marks)	
Load the dataset/Splitting the data	1
Scatter plot	1
Normalization	1
Logistic regression	3
Predictions on testing set/Accuracy	1
Total (18/3) => **6 marks**	