School of Information Technologies and Engineering, ADA University

CSCI4734 – Machine Learning

Spring 2024

Assignment 1

Deadline: February 25, 2024, 23:59

Dataset

In this programming homework, we will use a *turbo.az* dataset which is a collection of cars (Mercedes C-class). It contains information about 1328 cars with features of (*Sheher, Marka, Model, Buraxilish ili, Ban novu, Reng, Muherrikin hecmi, Muherrikin gucu, Yanacaq novu, Yurush, Suretler qutusu, Oturucu, Yeni, Qiymet, Extra Information, Seller's comment*).

Grading Components

Homework is divided into several parts:

1. Loading data (10%)

Using pandas library in Python is recommended.

You will need to read the data from data file (*turboaz.csv*) and extract only 3 columns for your model:

X_1 = *Yurush* (milage). If samples of *Yurush* are described by string format, remove *km* from the string and convert it to the number format.

X_2 = *Buraxilish ili* (model year).

Y = "Qiymet" (price). If prices of car are given in dollar ($) convert them to manat (AZN).

2. Visualization (10%)

Using matplotlib library (scatter, Axes3D) in Python is recommended.

You will need to provide 3 visualizations of data.

a) *Qiymet* (Y) vs *Yurush* (X_1)

b) *Qiymet* (Y) vs *Buraxilish ili* (X_2)

c) 3D plot of all three values (Y, X_1, X_2)

3. Implementation of Linear Regression from scratch (40%)

a) Calculate cost function. Implement a function which returns cost given true y values, x values and coefficients (w).

$$J(w_0, w_1, \dots, w_d) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_w\left(x^{(i)}\right) - y^{(i)}\right)^2$$

b) Normalize data using Z score normalization (Recommended).

$$z = \frac{x - \mu}{\sigma}$$

Mean	Standard deviation
$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

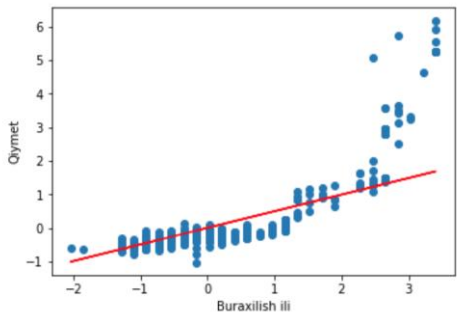
c) Implement gradient descent algorithm to minimize the cost function.

- Assign initial values of W=(w_0, w_1, w_2) to zero or choose randomly
- Learning rate: alpha=0.001, you can change it in different experiments
- Number of iterations: 10000 or you can stop it when two sequential values are too close.
- Calculate values of parameters using gradient descent formula.

$$\boldsymbol{w_j} := \boldsymbol{w_j} - \boldsymbol{\alpha}\frac{\boldsymbol{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(\boldsymbol{h_w}\left(\boldsymbol{x}^{(\boldsymbol{i})}\right) - \boldsymbol{y}^{(\boldsymbol{i})}\right)\boldsymbol{x}_{\boldsymbol{j}}^{(\boldsymbol{i})}$$

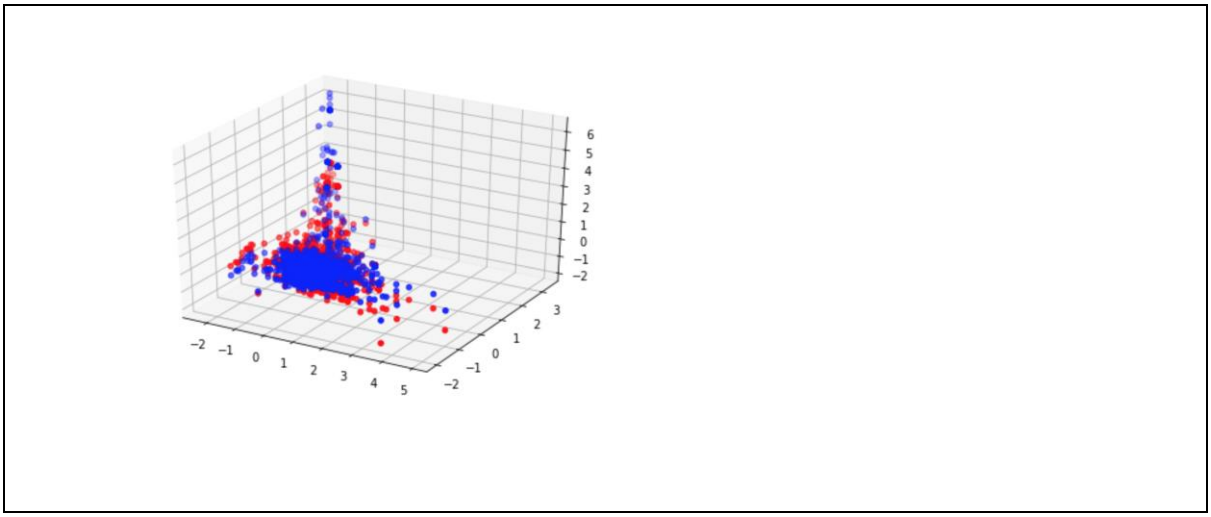
d) Plot graph of Cost function and describe how it changes over iterations.

e) Plot points of Y (*Qiymet*) vs X1 (*Buraxilish ili*) and draw a line of predictions made with parameters you got from gradient descent.



f) Plot points of Y (*Qiymet*) vs X_2 (*Yurush*) and draw a line of predictions made with parameters you got from gradient descent.

g) Plot 3D graph of points of Y (*Qiymet*), X_1, X_2 and predicted Y (*Qiymet*) using the same X_1 and X_2. It should look like this (blue points are true values, red points are predicted values):



h) Given are two new cars which are not in the dataset:

Car 1 {*Yurush*: 240000, *Buraxilish ili*: 2000, *Qiymet*: 11500}

Car 2 {*Yurush*: 415558, *Buraxilish ili*: 1996, *Qiymet*: 8800}

Predict the prices of these cars by using your parameters and compare your predictions with the actual prices.

4. Linear Regression using library (20%)

Use a library to fit perform linear regression on the data. You should use the same features (*Yurush* and *Buraxilish ili*) as input to this model and perform the same testing (3(h)) as above (using scikit-learn library is recommended).

5. Report (20%)

Write codes, their explanation and achieved results for each step (1-4) of homework in the report. Codes must be in the text format. You can take screenshots only for output of the code.

Submission

Please follow the instructions below when you make your submissions to the Blackboard System:

- The language of choice for assignments is Python. Consider using Jupyter Lab for your own convenience.
- Submit your homework solution in 2 separate files:
  1. Code (.py/.ipynb).
  2. Report (.doc/.docx/.pdf) – containing explanations and achieved results.

- Name the file according to the template (all capital):
  1. CSCI4734_2024S_20252_A1_FIRSTNAME_LASTNAME_CODE
  2. CSCI4734_2024S_20252_A1_FIRSTNAME_LASTNAME_REPORT
- Do NOT submit files in an archive file format.
- You can submit your assignment 3 times before the deadline.
- Latest submission will be considered for grading.
- For one day delay, your grade will be deducted 25%. No submission is accepted after one day delay.
- Furthermore, any student whose solution may arise a question, will be asked for some explanations as well.