Homework 2: Optimization and Linear Regression
CPE 490/590 ST

Instructor: Rahul Bhadani

Due: Feb 16, 2025, 11:59 PM
175 points

You are allowed to use a generative model-based AI tool for your assignment. However, you must submit an accompanying reflection report detailing how you used the AI tool, the specific query you made, and how it improved your understanding of the subject. You are also required to submit screenshots of your conversation with any large language model (LLM) or equivalent conversational AI, clearly showing the prompts and your login avatar. Some conversational AIs provide a way to share a conversation link, and such a link is desirable for authenticity. Failure to do so may result in actions taken in compliance with the plagiarism policy.

Additionally, you must include your thoughts on how you would approach the assignment if such a tool were not available. Failure to provide a reflection report for every assignment where an AI tool is used may result in a penalty, and subsequent actions will be taken in line with the plagiarism policy.

Submission instruction:

Submission instruction for this homework supersedes one mentioned in the Syllabus.

This homework requires all answers recorded in a single `.ipynb` Python notebook. You can use a combination of text cell (i.e. markdown formatted cell) and code cell to provide your answer. To add equations you should be able to use Latex syntax in the text cells of your Python notebook. As a part of your submission, you must provide executed notebook with code, text, and outputs. Alternatively, you can also provide a url (whose permission you must change to 'anyone with link can view') of your Python notebook from Google Colab. The naming convention for your notebook should follow the format {firstname_lastname}_CPE 490/590 ST_hw02.ipynb. For example, if your name is Sam Wells, and you are enrolled in CPE 490 your file name should be sam_wells_CPE490_hw02.ipynb.

Please refer to https://github.com/rahulbhadani/CPE490_590_Sp2025/blob/master/Code/Chapter_04_Linear_Regression.ipynb

and

https://github.com/rahulbhadani/CPE490_590_Sp2025/blob/master/Code/Chapter05_Optimization_and_Gradient_Descent.ipynb for hands-on related to Chapter 04 and Chapter 05.

Theory

1 Linear Regression and Regularization (15 points)

In class we derived and discussed linear regression in detail. Find the result of minimizing the loss of sum of the squared errors after adding in a penalty called an L_2 penalty on the weights. More formally

$$\arg\min_{\mathbf{w}} \left\{ ||\mathbf{y} - \mathbf{x}\mathbf{w}||^2 + \lambda ||\mathbf{w}||^2 \right\} \tag{1}$$

which means we are interested in finding $\mathbf{w}$ that minimizes the overall expression. How does this change the solution to the original linear regression solution? What is the impact of adding in this penalty?

$||\cdot||$ is L2 norm.

Hint: First take the gradient with respect to $\mathbf{w}$ and then equate to zero you will obtain an expression for $\mathbf{w}$ which is the estimated $\hat{\mathbf{w}}$.

Also, remember in vector form the L2 norm $||v||$ can be written as $\mathbf{v}^\top \mathbf{v}$.

2 Optimizing a Function using Partial Differentiation (5 points)

Consider a simple cost function (also called as loss function or simply, loss) of one variable:

$$J(w_1) = (w_1 - 12)^2 + 32 \tag{2}$$

Using partial differentiation method to find the value of w_1 that minimizes the cost function $J(w_1)$.

3 Optimizing a Function using Stochastic Gradient Descent (5 points)

Use the Stochastic Gradient Descent Technique to calculate find the optimal value w_1 for the cost function in Equation (2). Use the learning rate parameter $\eta = 0.1$, and the initial value of w_1 as 0.0. Write the updated value of w_1 for up to 5 iterations. This question doesn't ask to write code.

4 Goodness of Fit (5 points)

What is goodness of fit in the linear regression and how do we quantify them?

5 Constraint Optimization (10 points)

Consider the function $f(x, y) = e^{xy}$ subject to constraint $x^3 + y^3 = 16$. Use Lagrange multipliers to find the coordinates (x, y) of any points subject to the given constraint where the function f could attain a maximum or minimum.

Python Practice: Pandas and Matplotlib

6 Electric Car Sales Data (10 Points)

Go to `https://github.com/rahulbhadani/CPE490_590_Sp2025/tree/master/Data/EVSales` and download the CSV file `IEA-EV-dataEVsalesHistoricalCars.csv`.

Perform the following task:

1. Read the `IEA-EV-dataEVsalesHistoricalCars.csv` into a pandas data frame.
2. Remove entries from the region column where value = "world".

3. Keep only those rows with "EV sales" and "EV stock" parameters.

4. Create a data frame showing the number of cars sold each year, in each country, with a BEV ("battery electric vehicle") powertrain.

7 Electric Vehicle Battery Data (10 Points)

Download electric vehicle data from `https://www.youtube.com/watch?v=C6eQ6VwTpxk&list=PLSK7NtBWwmpTS_YVfjeN3ZzIxItI1P_Sr`

The dataset contents electric vehicle battery log from Beijing, China. Each row represents a snapshot of data collected at a specific time, capturing various metrics related to the vehicle's battery and operational status.

Some information about specific columns:

- **Max Voltage Battery Pack Serial Number**: The serial number of the battery pack with the highest voltage.
- Max Voltage Battery Serial Number : The specific battery cell within that pack with the highest voltage. 7
- Max Voltage Value : The actual maximum voltage recorded (e.g., 3.94V).
- Min Voltage Battery Pack Serial Number : The serial number of the battery pack with the lowest voltage.
- Min Voltage Battery Serial Number : The specific battery cell within that pack with the lowest voltage.
- Min Voltage Value : The actual minimum voltage recorded (e.g., 3.94V).
- Total Voltage : The combined voltage of all battery packs in the vehicle (e.g., 359.50V).
- Max Temperature Battery Pack Serial Number : The serial number of the battery pack with the highest temperature.
- Max Temperature Probe Serial Number : The specific temperature probe within that pack recording the highest temperature.
- Max Temperature Value : The highest temperature recorded (e.g., 20 C).
- Min Temperature Battery Pack Serial Number : The serial number of the battery pack with the lowest temperature.

- Min Temperature Probe Serial Number : The specific temperature probe within that pack recording the lowest temperature.
- Min Temperature Value : The lowest temperature recorded (e.g., 16 C).
- State of Charge (SOC): Represents the current charge level of the battery as a percentage (e.g., 69.2%).
- Total Current : Indicates the current flow in the system, which can be positive (charging) or negative (discharging).
- Remaining Power (kWh) : Shows the amount of energy left in the battery, measured in kilowatt-hour
- Date and Time in Human Readable Format

Perform the following data analysis task:

1. Read a csv file `vehicle-PKUB7Y424-20150416134232.csv` from the folder specified into a pandas dataframe. Create new column called **Timestamp** in POSIX timeformat using Upload_Date, Upload_Time columns. The POSIX standard stores date time values internally as the number of seconds since midnight on January 1, 1970 in nanoseconds. For example 4/2/2015 7:31:41 is 1427977901 in POSIX format. **(2 Point)**
2. Create a scatterplot showing state of the charge (SOC) as a function of time (use Timestap) column for the x-axis you created in the previous step. **(2 Point)**
3. Is there any correlation (linear relationship) between the State of the Charge and Total Current? **(2 Points)**
4. Plot the probability distribution of the State of the Charge. **(2 Points)**
5. What kind of distribution do you observe in Min Temperature Value ? Explain after plotting the probability distribution. **(2 Points)**

Differentiation, Optimization, and Linear Regression

8 Computing Derivative (5 Points)

Consider a following function

$$\tau = 2r \cdot \sqrt{\frac{r}{r_s}} \cdot \sqrt{1 - \frac{r_s}{r}} \tag{3}$$

which is the time experienced by an observer traveling along a specific trajectory in curved space-time. r is the radial coordinate in a spherical coordinate system and r_s is the Schwarzschild radius.

Write a program in PyTorch to compute the derivative $\frac{d\tau}{dr}$ considering the value of $r = 30000.1$ for $r_s = 0.00887$.

9 Calculating the Partial Derivative (10 points)

Consider a function

$$f(x, y, z) = 20z^3 + 12yz - xy \tag{4}$$

Compute their partial derivative (i.e. gradient) at $x = 3.0$, $y = 9.0$, and $z = 2.0$ using PyTorch. Verify your answer by doing your own calculation by hand.

10 Optimization in PyTorch (20 points)

Minimize the cost function from Equation (2) with respect to w_1 using PyTorch. Choose the initial value of w_1 as 0.0 and the learning rate $\eta = 0.1$. Perform the training for 100 iterations. Plot the loss against the iteration as well as plot the w_1 against the iteration.

Now, redo the above, but set $\eta = 2.0$. What happened to the loss and w_1 as the training progressed? Make a comment on what you observed.

11 Constraint Optimization using PyTorch (20 Points)

Consider the function $f(x, y) = e^{xy}$ subject to constraint $x^3 + y^3 = 16$.

Create the contour plot visualizing the objective function f and the constraint.

Use Penalty method to find maximum/minimum value as well as x and y that maximizes and/or minimizes f.

1. Use penalty weight of 10, 100, 1000 and repeat the optimization procedure. Choose learning rate of 0.001 and perform 5000 epochs of training. What do you observe penalty weight is changed?

12 Constraint Optimization using PyTorch with Projected Gradient (20 Points)

Consider the objective function

$$f(x_1, x_2) = 2\pi x_1(x_1 + x_2) \tag{5}$$

which asks about minimizing surface area of cylinder with constraint on its volume as

$$g(x_1, x_2) = \pi x_1^2 x_2 - V = 0 \tag{6}$$

where x_1 and x_2 are raidus and height of cylinder and V is the required volume.

Choose a few values of V as 30.0, 40.0, 50.0, and 60.0

and perform constraint minimization using Project Gradient method in PyTorch.

Choose appropriate initial values of x_1 and x_2 and learning rate of 0.001.

Is there any relationship between different values of x_1 that minimizes f and V that you observe?

13 Single Predictor Variable Linear Regression using PyTorch (20 points)

Use Toluca Dataset from `https://github.com/rahulbhadani/CPE490_590_Sp2025/blob/master/Data/Toluca/toluca.txt` to implement a linear regression model to predict **Work Hour** given the **Lot Size**. Choose the appropriate value for learning parameter η and initial value of your model coefficients. Choose the number of iteration as 1000. (10 points)

(2 points each):

1. Plot the loss as a function of iteration.
2. Plot the coefficients as a function of iteration
3. Make a scatter plot of the data and overlay the fitted line. Make sure you label all plots properly.
4. How many number of **WorkHours** are required for the **LotSize** of 75?
5. Plot the WHS confidence band.

14 Solving a Differential Equation by Fitting a Polynomial using PyTorch (20 Points)

Following differential equation models radioactive decay

$$\frac{dN}{dt} + kN = 0 \tag{7}$$

where $N(t)$ is the number of radioactive atoms in some radioactive material.

Assume $k = 1.216 \times 10^{-2}$,

Its analytical solution is given by

$$N(t) = N(0)e^{-kt}$$

Compute $N(t)$ using PyTorch by fitting a Polynomial of degree 3 to the above differential equation. Choose the range of t as $t = 0.0$ to $t = 200.0$ sampled at $t = 0.004$ seconds. Consider $N(0) = 5.0$.

Choose learning rate as 0.001 and number of epochs as 500000.

The half life $t_{1/2}$ is given by $t_{1/2} = \dfrac{\ln 2}{k}$ at which number of radioactive atoms become half.

Compute the number of atoms at half life from the fitted model and compare with the one you obtain from the analytical solution.

Will you be able to make a prediction on $t = 5000.0$ correctly? Why or why not?