MATH 456 — Mathematical Modeling

Assignment #5: Linear Regression **Due Date:** March 11, 2022, 11:59 PM

Question 1. (Simple Linear Regression) In this question, you will implement linear regression with one variable to predict profits for a food truck. Suppose you are the CEO of a restaurant franchise and are considering different cities for opening a new outlet. The chain already has trucks in various cities and you have data for profits and populations from the cities.

The file ex1data1.txt contains the dataset for our linear regression problem. The first column is the population of a city and the second column is the profit of a food truck in that city. A negative value for profit indicates a loss.

(1.a) Visualize the data by showing the scatter plot (Hint: X-axis — Population of City in 10,000s, Y-axis — Profit in $10,000s).

(1.b) Consider the linear regression model

$$y = \theta_0 + \theta_1 x$$

in the population-profit prediction problem. Compute the least square cost function $J(\theta)$ where $\theta = (\theta_0, \theta_1)$, and find the optimal parameter θ that minimize the cost. To understand the cost function $J(\theta)$ better, you are recommended to visualize $J(\theta)$ over a 2-dimensional grid of θ_0 and θ_1 values.

(1.c) Use the result in (1.b) to make predictions on profits in areas of 35,000 and 70,000 people.

Question 2. (Linear Regression with Multiple Variables) In this question, you will implement linear regression with multiple variables to predict the prices of houses. Suppose you are selling your house and you want to know what a good market price would be. One way to do this is to first collect information on recent houses sold and make a model of housing prices.

The file ex1data2.txt contains a training set of housing prices in Portland, Oregon. The first column is the size of the house (in square feet), the second column is the number of bedrooms, and the third column is the price of the house.

(2.a) **Pre-processing of data**. Note that house sizes are about 1000 times the number of bedrooms. When features differ by orders of magnitude, the feature matrix can be highly ill-posed. To avoid this issue, one often perform feature normalization before solving the regression problem. Specifically, normalize the data in the following steps.

- Subtract the mean value of each feature from the dataset.
- After substraction, divide the feature values by their respective standard deviations.

(2.b) Consider the linear regression model

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

in the house-price prediction problem. Find the optimal $\theta = (\theta_0, \theta_1, \theta_2)$ that minimizes the least square cost function.

(2.c) Use the result in (1.b) to make predictions on housing price a 1650-square-foot house with 3 bedrooms.

Note: Both questions need to be done using either Python or Matlab. Document your results for both questions in Word/PDF/Markdown/Jupyter Notebook. You need to upload the document along with the codes.