HOMEWORK 1:
LINEAR REGRESSION AND LOGISTIC REGRESSION

ELEC 400M @ UBC
TAs: Sadegh Mahdavi (Primary), Wenlong Deng, Chun-Yin Huang

Instructions

- **Homework Submission:** Submit your code and report to Canvas. You will use Co-lab to implement the coding tasks. Please check Piazza for updates about the homework.
  - Upload a zip file containing two files: Your report in .PDF format and your notebook in .Ipynb format.
  - To ensure the reproducibility of your results: (1) set a seed for numpy and python random modules on top of your Colab notebook (2) restart and run all the cells of your notebook once before submission.
- **Collaboration policy**: The purpose of student collaboration is to facilitate learning, not to circumvent it. Studying the material in groups is strongly encouraged. It is also allowed to seek help from other students in understanding the material needed to solve a particular homework problem, provided no written notes (including code) are shared, or are taken at that time, and provided learning is facilitated, not circumvented. The actual solution must be done by each student alone. The presence or absence of any form of help or collaboration, whether given or received, must be explicitly stated and disclosed in full by all involved.
- **Description:** In Assignment 1, you will calculate the you will calculate the analytic solution of a linear model on a training dataset and report the error on both training and testing set. You will use Python an Sklearn to confirm your results. Next, you will practice using gradient descent to train a logistic regression model by coding the gradient descent algorithm by yourself. Again, you will use Python and Sklearn to confirm your results.

1 Linear Regression [10 pts]

Assume we have a training set and a testing set as follows:

Sample ID	x	y
1	6.94	2.0
2	0.19	-0.07
3	5.70	-0.32
4	6.74	1.17
5	4.49	-1.16
6	2.02	-0.71
7	1.78	-2.74
8	6.84	2.56
9	1.52	-1.37
10	0.80	-0.49

Table 1: Training set

Sample ID	x	y
1	4.88	-1.17
2	1.28	-0.01
3	3.36	-2.39
4	6.07	0.82
5	3.98	-2.18
6	3.91	-2.49
7	5.56	-0.38
8	4.62	-1.97
9	5.85	1.19
10	5.41	-0.50

Table 2: Testing Set

Please fit the linear model $\hat{Y} = f_\Theta(\mathbf{X})$ using RSS objective $J(\Theta) = \|\hat{Y} - Y\|_2^2$.

(a) Calculate the analytic solution of the linear model $f_\Theta(x) = \theta_0 + \theta_1 x$. Then, use scatter plot to plot the training data points and draw the fitted line on the same figure.

(b) Suppose we want to increase the model complexity, by considering y as a linear function of both x and x^2. Namely $f_\Theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$. In this case, calculate the analytic solution of model and plot the smooth curve of the model, together with the scatter plot of data points in training set.

(hint: in this case, $\mathbf{X} = \begin{pmatrix} 1 & x^{(1)} & x^{(1)^2} \\ \vdots & \vdots & \vdots \\ 1 & x^{(10)} & x^{(10)^2} \end{pmatrix}$)

(c) Let us further increase the model complexity by assuming y is related to higher-order forms of x, i.e., $f_\Theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$. Again, calculate the analytic solution of the model and plot the curve of the function, together with the data points in training set.

(d) Observe the above three functions, point out which could be faced with underfitting, which could be faced with overfitting, and which one is relatively a good fit? Then, calculate the values of prediction error on the test data to verify your thoughts.

(e) For question (a-c), verify your optimal Θ^* using the linear regression function in *sklearn*. The example code is provided in ELEC_400M_HW1_Example_Codes.ipynb.

2 Logistic Regression [10 pts]

Assume that we have a training set and a test set as follows: Use these data to implement a logistic classifer. We use the linear model $f_\Theta(x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ and the logistic regression function is written as $\sigma_\Theta(x_1, x_2) = \frac{1}{1+e^{-f_\Theta(x_1,x_2)}}$. We use cross-entropy error as the loss function.

As introduced in the lecture, we will use gradient descent method to update the model based on the data points in the training set. The model parameters are initialized as $\theta_0 = -1, \theta_1 = -1.5, \theta_2 = 0.5$. We set step size $\alpha = 0.1$.

(a) Write down the logistic model $P(\hat{y} = 1|x_1, x_2)$ and its cross-entropy error function.

(b) Use gradient descent to update θ_0, θ_1 and θ_2 for ONE iteration. Write down the equations, gradient values, and updated parameters. Then implement the iterative updating using your own Python algorithm till convergence. The example code is provided in ELEC_400M_HW1_Example_Codes.ipynb.

Sample ID	x_1	x_2	y
1	1.15	1.74	0
2	1.26	1.53	0
3	1.01	1.92	0
4	1.90	1.03	0
5	2.96	0.14	1
6	2.28	0.61	1
7	2.94	0.85	1
8	2.00	0.52	1

Table 3: Training set

Sample ID	x_1	x_2	y
1	1.77	1.51	0
2	2.18	0.93	0
3	2.13	1.95	0
4	2.80	1.38	1
5	2.06	-0.35	1
6	2.74	1.13	1

Table 4: Testing Set

(c) Visualize the training set and your model's decision hyperplane for your model's state at initialization, after one iteration, and after convergence. For each of the three plots, you may use a scatter plot for visualizing the training data, two different label colors for each class, and a line (derived from Θ of your model) indicating the decision boundary of your model.

(d) Use *Sklearn* to find the best θ_0, θ_1 and θ_2 till convergence. The example code is provided in ELEC_400M_HW1_Example_Codes You may leave *Sklearn*'s default solver and initialization unchanged to compute the optimal Θ.

(e) Use the above new model to make predictions for all the samples in the test dataset. Compare and verify your final results in (b) and the results generated by *Sklearn* in (c). Then, using your own calculations/implementations, compute the accuracy, precision, and recall of both models.

Note

1. Remember to submit your assignment by 23:59pm of the due date. Late submission will affect your scores.
2. If you submit multiple times, ONLY the content and time-stamp of the latest one would be considered.
3. We strictly follow the rules of UBC Academic Misconduct.