HOMEWORK 2:
LOGISTIC REGRESSION

CPEN 355 @ UBC
TAs: Wenlong Deng (Primary for HW2), Chunyin 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. Note if you use ChatGPT, please provide your prompts and answers provided by chaGPT.
- **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 2, you will implement a logistic regression model and evaluate its performance.

Logistic Regression [15 pts]

Assume that we have a training set and a test set as follows: Use these data to implement a logistic classifer.

Sample ID	x_1	x_2	y
1	-1.63	-1.36	0
2	-1.67	-1.36	0
3	-1.81	-2.73	0
4	-0.82	-1.40	0
5	-0.71	2.95	1
6	-0.25	1.78	1
7	-0.61	0.79	1
8	0.49	0.56	1

Table 1: Training set

Sample ID	x_1	x_2	y
1	-1.90	-0.95	0
2	-0.27	-0.87	0
3	0.14	-2.23	0
4	0.49	-0.18	1
5	0.97	1.43	1
6	1.07	1.30	1

Table 2: Testing Set

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 loss 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 loss 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 CPEN_355_HW2_Example_Codes.ipynb.

(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, and plot the decision boundary of the final parameters. The example code is provided in CPEN_355_HW2_Example_Codes.ipynb. 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 (d). Then, using your own calculations/implementations, compute the accuracy, precision, and recall of both models.

Short Answer Questions [5 pts]

Provide your answer to the following questions.
Question 1: In a binary classification problem with Logistic Regression, what is the primary purpose of the decision boundary?

(A) It separates the training data into two clusters.

(B) It defines the line where the predicted probability equals 0.5.

(C) It is a hyperplane that maximizes the margin between classes.

(D) It represents the line with the highest gradient of log-odds.

Question 2: For linear separable data, Logistic Regression can find parameters that achieve *exact zero* loss on the training set?

(A) Yes

(B) No

(C) Depends

Note

1. Remember to submit your assignment by 23:59pm of the due date (Feb 14, 2024). 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.