CSE422 Problem Set: Logistic Regression

Instructor: Ipshita Bonhi Upoma

Part 1: Mathematical Basics

The parameters of the hypothesis function, in logistic regression are updated in gradient descent algorithm as,

$$w_j = w_j + \alpha \frac{\partial Loss}{dw_i} \tag{1}$$

Calculate $\alpha \frac{\partial Loss}{dw_i}$ for,

1. Loss-function is of the MSE form,
$$\textbf{Loss} = \sum_{i}^{m} (h(x^{(i)}) - y^{(i)})^2 \tag{2}$$

2. Loss function is of the Binary-Cross Entropy form,
$$\textbf{Loss} = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \right] \tag{3}$$

   where:

   - m is the number of training examples.
   - $y^{(i)}$ is the actual class label of the i-th training example, which can be 0 or 1.
   - $h(x^{(i)})$ is the predicted probability that the i-th training example belongs to the class with label 1, calculated as $h(x^{(i)}) = \sigma(z)$), where σ is the sigmoid function, $z = \mathbf{W} \cdot \mathbf{X} + b$ and b is the bias.

Part 2: Simulation on Datasets

Task: Simulate the gradient descent algorithm for logistic regression, performing up to two iterations.
Instructions:

- Initial Parameters: Set all model parameters to an initial value of 1.
- Learning Rate: Use the learning rate $\alpha = \frac{1000}{1000+t}$ where, t is the iteration number.
- Loss Function: Employ the binary cross-entropy loss function to measure the model's performance.

Dataset 1: Email Spam Classification This dataset is used to classify emails as spam or not spam based on keyword frequency and email length.

Keyword Frequency (%)	Email Length (1000s of characters)	Spam (1) or Not Spam (0)
20	2	1
5	3	0
30	1	1
7	2	0
25	1	1
3	4	0
15	1.5	1
4	3.5	0

Table 1: Dataset for classifying emails as spam or not spam.

Dataset 2: University Admission Prediction This dataset predicts whether a student will be admitted to a university based on their GRE score and GPA.

GRE Score (out of 340)	GPA (out of 4.0)	Admitted (1) or Not Admitted (0)
330	3.9	1
315	3.5	1
300	3.2	0
320	3.8	1
310	3.0	0
305	3.3	0
325	3.7	1
318	3.6	1

Table 2: Dataset for predicting university admissions.

Dataset 3: Loan Default Prediction This dataset predicts whether a borrower will default on a loan based on their annual income and loan amount.

Annual Income ($1000s)	Loan Amount ($1000s)	Default (1) or Not Default (0)
45	15	0
85	20	0
50	30	1
60	22	0
30	25	1
100	35	0
40	18	1
75	28	0

Table 3: Dataset for predicting loan defaults.

Part 3: Find out the answers to these conceptual questions.

1. What is a decision boundary in logistic regression? Discuss how different parameter values can affect the position and shape of the decision boundary.
2. Why is it better to use binary-cross entropy as a loss-function instead of the mean squared error used in linear regression?
3. In logistic regression, how are the outputs of the model interpreted as probabilities?
4. Describe the role of the sigmoid function in logistic regression. Why is it used instead of a linear function?
5. What is logistic regression and in what type of machine learning problems is it most appropriately used? Explain how it differs from linear regression.