Diabetes Classification using Machine Learning

Dexson John D'Souza
Department of Computer Science
University at Buffalo
Buffalo, NY 14260
dexsonjo@buffalo.edu

Results

I. Experimental Setup

To classify whether a person has diabetes, we have used Gradient Descent with Logistic Regression. Pima Indians Diabetes Database dataset was used. The dataset was split into train(60%), test(20%) and validation(20%) sets. We have performed normalization on the dataset. We have used Jupyter Notebook which is an open-source software. We have used Scikit-learn which is open-source software for machine learning library in Python programming language to perform data partitioning. We have used Numpy which is library for the Python programming language. We have also used the Python library Pandas.

II. Result Analysis

The cost function with respect to the number of iterations for the diabetes classification system using Logistic Regression is depicted in Fig. 1. The X-axis denotes the iterations and Y-axis denotes the cost at a specific iteration. The model was able to achieve 79.2% accuracy for diabetes classification. The hyperparameters: learning rate was 0.001 and number of epochs were 10000.

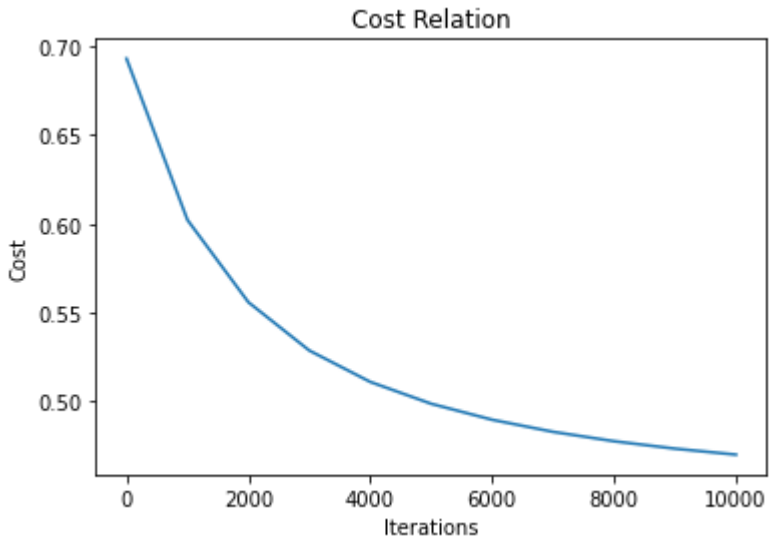

Figure 1: Cost-Iteration Relation for Logistic Regression

III. Comparative Study

Performance of the model for different hyper-parameter and model parameter is depicted in Table 1.

Table 1: Comparison of the model's performance for different parameter settings.

Learning Rate	No of Iteration	Accuracy Achieved
0.0001	1000	75.97
0.01	1000	79.2
0.0001	100000	79.2
0.01	100000	77.9