CS 412 - HW3 Report

Logistic Regression Model Performance Analysis

28.04.2024

İpek Akkuş

30800

The purpose of this report is to explain the results of building a logistic regression model to predict survival outcomes using dataset from the Titanic.

Data were preprocessed by partitioning into distinct sets for training (60%), validation (20%), and testing (20%). Random seed is set to 42 as requested in the homework document to standardize the results. Different functions such as sigmoid, cost and gradient descent are implemented as requested to be used in implementing the model.

Feature standardization was applied to normalize the data. MinMaxScaler() is used to transform each feature to the range of [0,1]. The feature columns "Age," "Sex," and "Pclass" are subject to scaling; the target column "Survived" is not as it is dropped already. Only the training set of data is used to fit the scaler and determine the proper scaling parameters (the lowest and maximum values for each feature). This prevents data leakage by making sure that no information from the test or validation data affects the scaling parameters. The training, validation, and test datasets are transformed using the same scaler after fitting. This ensures consistent preparation free from bias from unseen data since the validation and test data are scaled using just the parameters determined from the training data.

In order to build a logistic regression model, sigmoid activation functions, cost computation functions, and gradient descent optimizers for fine-tuning model weights have developed as requested in the homework document. For the formulas and implementations for each function, please refer to the .ipynb file.

Firstly, the parameters 0.1 for learning rate and 100 for iterations are experimented to measure the model configuration. *Figure 1* illustrates the model's performance with an initial configuration using a learning rate (α) of 0.1 over 100 training iterations. The blue dashed line represents the training loss, and the solid orange line denotes the validation loss as understood from the legend of the graph. A declining loss curve suggests the model is learning and improving its predictions over time.

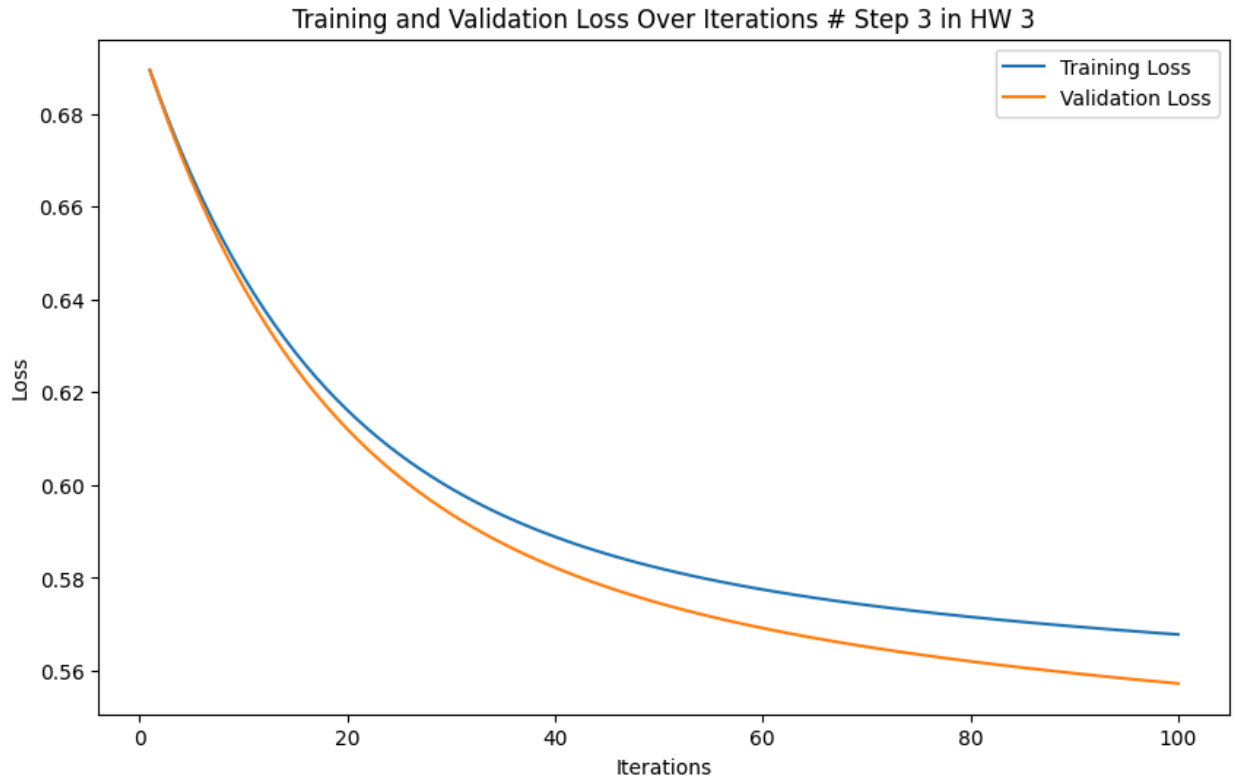

Figure 1: Training and Validation Loss Over Iterations with Learning Rate α = 0.1 and 100 Iterations

We experimented with different hyperparameters to get the best model configuration. The following learning rates were investigated: 0.001, 0.01, 0.1, 0.5, 0.8, and 1. The number of iterations was set at 50, 100, 200, and 500. At a learning rate of **0.8 over 200** iterations, the model produced the lowest validation cost, indicating sufficient generalization and avoiding overfitting. This was the optimal performance. *Figure 2* presents the loss curves after hyperparameter tuning, showing the best training and validation losses achieved with a learning rate (α) of 0.8 across 200 iterations. The

model achieved the lowest validation cost with these hyperparameters, indicating an optimized balance between learning efficiency and generalization capability.

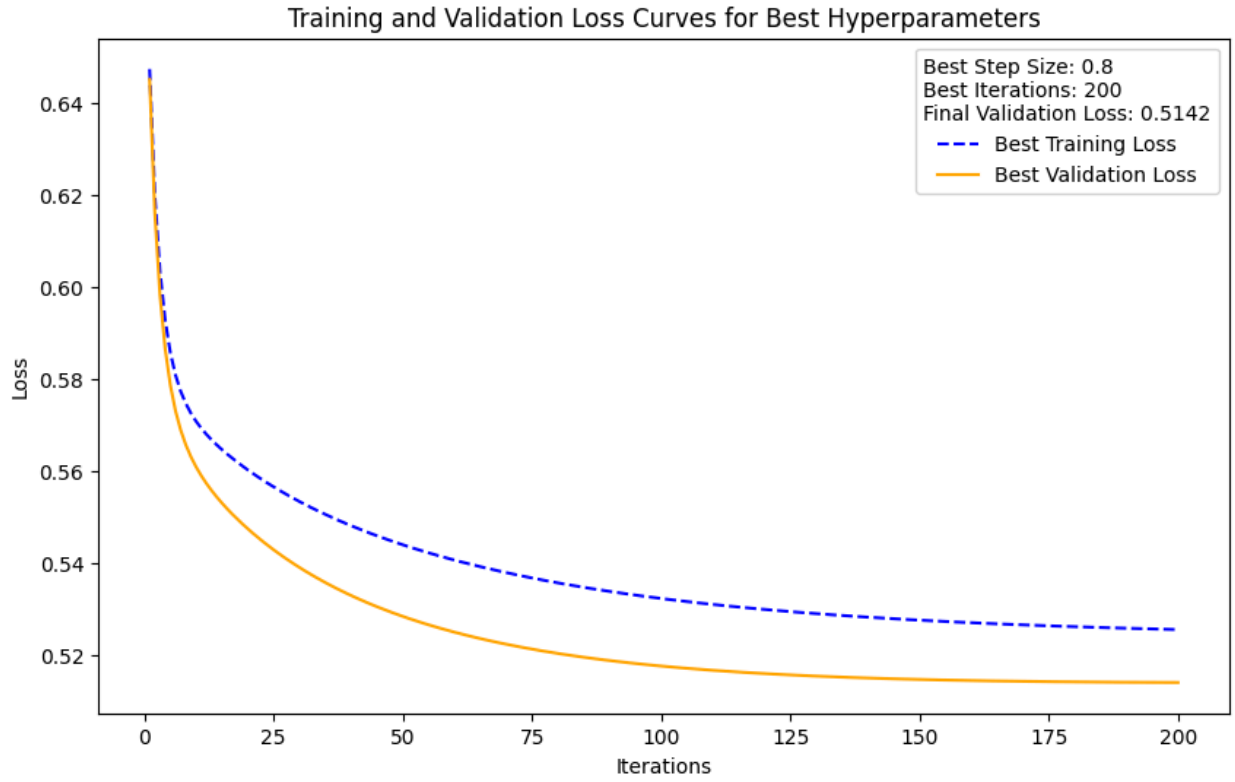

Figure 2: Optimal Training and Validation Loss Curves with Learning Rate α = 0.5 and 50 Iterations

The training and validation datasets were combined after optimization, and the model was retrained with the obtained hyperparameters to improve its generalization performance in step 5.

In the step 6, the completed model was tested on a separate test set, yielding an accuracy value of **75.98%**, which indicates the effectiveness of the chosen hyperparameters and the competency of the logistic regression model. Retrained logistic regression model demonstrated great prediction accuracy, suggesting that it is robust. Enhancing the model's performance required careful consideration of its hyperparameters.