ISTANBUL TECHNICAL UNIVERSITY

COMPUTER ENGINEERING DEPARTMENT

BLG 454E

LEARNING FROM DATA HOMEWORK REPORT

150200060 : ÖMER YILDIZ

SPRING 2024

Contents

1 Logistic Regression Method (15 points)

1.1 Theoretical Background

Logistic Regression is a statistical method used for binary classification problems. It models the probability of a target variable belonging to one of two classes using the logistic (sigmoid) function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where $z = w^T x + b$ is a linear combination of the input features x, weights w, and bias b.

To optimize the model, loss functions such as Mean Squared Error (MSE) or Cross-Entropy Loss are used. The Cross-Entropy Loss, commonly used for classification tasks, is given by:

$$\text{Cross-Entropy Loss} = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability.

1.2 Results and Analysis

Figure 1 illustrates the effect of hyperparameters on training accuracy using Cross-Entropy Loss, while Figure 2 shows the results for MSE Loss.

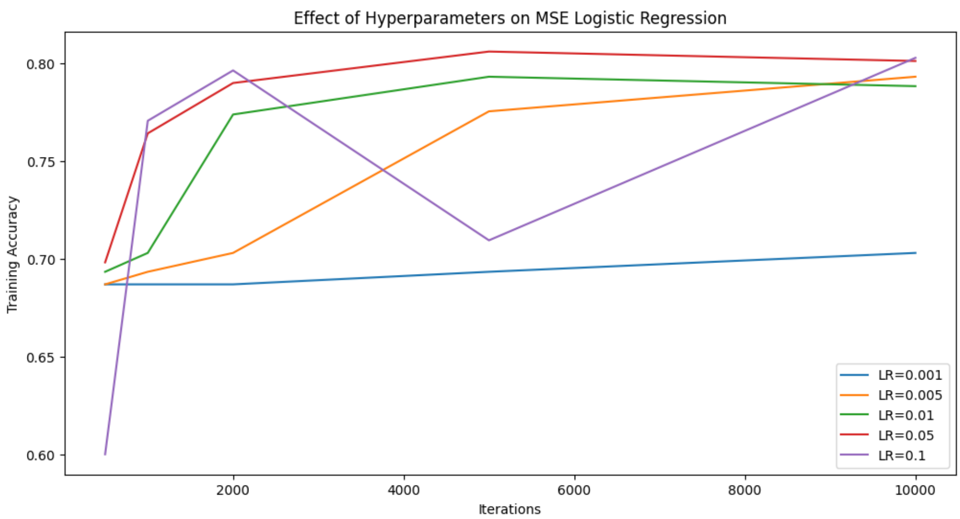

Figure 1: Effect of Hyperparameters on Cross-Entropy Logistic Regression.

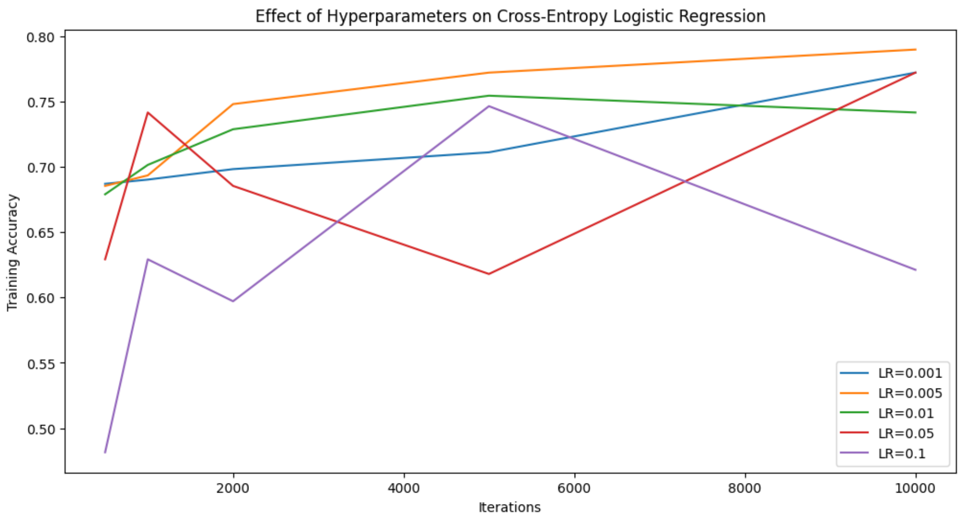

Figure 2: Effect of Hyperparameters on MSE Logistic Regression.

The best results were achieved with the following hyperparameters:

- Cross-Entropy Loss: Learning Rate = 0.005, Iterations = 10,000, Accuracy = 0.7897.
- MSE Loss: Learning Rate = 0.05, Iterations = 5,000, Accuracy = 0.8058.

1.3 Discussion

The learning rate and number of iterations significantly affect model performance. Smaller learning rates ensure gradual convergence, while higher rates may lead to instability. Iterations control the extent of optimization, with diminishing returns observed after a certain threshold.

Figure 3: Impact of learning rate and number of iterations on Logistic Regression accuracy.

2 Decision Tree Method (15 points)

2.1 Theoretical Background

Decision Trees are supervised learning algorithms used for classification and regression tasks. A decision tree splits the data into subsets based on feature values, creating a tree-like structure. Each node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome.

The depth of the tree, controlled by the `max_depth` hyperparameter, plays a crucial role in the model's performance. A shallow tree may underfit the data, while a deep tree can lead to overfitting.

2.2 Results and Analysis

The model's performance was evaluated by varying the `max_depth` hyperparameter. The results showed:

- Small `max_depth` values result in underfitting, with low training and test accuracies.
- Large `max_depth` values lead to overfitting, with high training accuracy but lower test accuracy.

2.3 Hyperparameter Tuning Analysis

Using hyperparameter tuning, the optimal `max_depth` was identified. The highest test accuracy of **80.97%** was achieved at `max_depth` = **3**. Figure 4 illustrates the impact of `max_depth` on the model's accuracy.

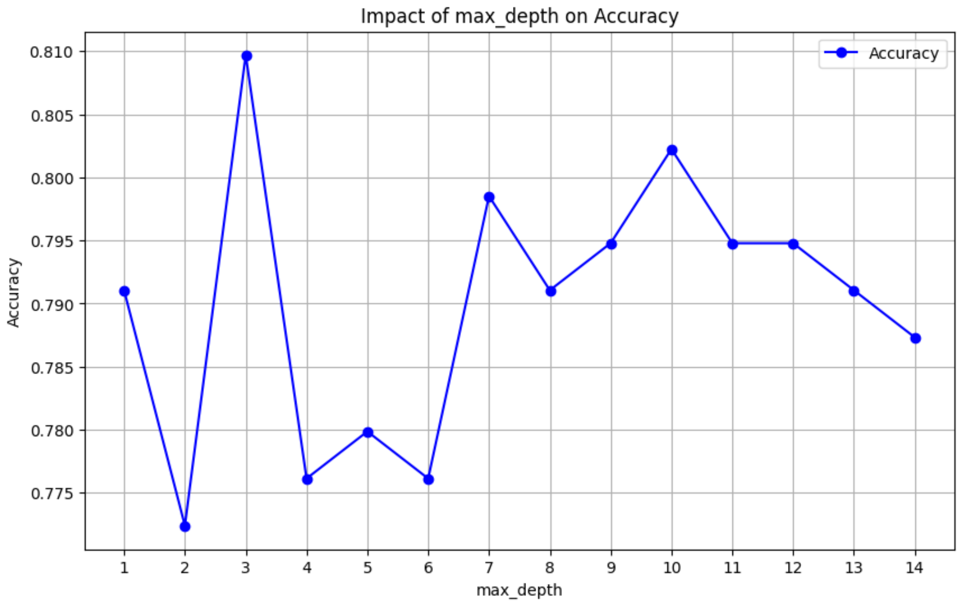

Figure 4: Impact of `max_depth` on Decision Tree accuracy.

3 Conclusions

This report analyzed Logistic Regression and Decision Tree methods, focusing on their theoretical foundations, hyperparameter tuning, and the impact of key parameters on model performance. Logistic Regression is sensitive to the learning rate and number of iterations, while Decision Trees are influenced by the `max_depth` parameter. Proper tuning of these parameters is essential for achieving optimal performance.