SML Assignment Report

Vikranth Udandarao
2022570

Question 1:

1. Introduction:

- The goal of this project is to:
    - Implement AdaBoost.M1, a boosting algorithm, on the MNIST dataset using digits 0 and 1 and labelling them as -1 and 1, respectively. AdaBoost.M1 is a powerful ensemble learning method that combines weak learners (decision stumps in this case) to create a strong classifier. The project aims to demonstrate the effectiveness of AdaBoost.M1 in classifying handwritten digits and to evaluate its performance on both validation and test sets.
    - Implement gradient boosting using absolute loss as a regression problem and evaluate its performance on the test set. The key tasks include data preprocessing, PCA dimensionality reduction, training decision stumps, updating labels based on negative gradients, and computing mean squared errors (MSE) on validation and test sets.

2. Approach:

- **Data Preparation:** The MNIST dataset is loaded, and samples for digits 0 and 1 are filtered and labeled as -1 and 1, respectively. The dataset is then divided into training, validation, and test sets, with 1000 samples from each class reserved for validation.
- **PCA Dimensionality Reduction:** Principal Component Analysis (PCA) is applied to reduce the dimensionality of the data to p=5. This reduction helps in capturing the most relevant features while reducing computational complexity.
- **Training Decision Stumps:**
    - Decision stumps are trained using AdaBoost.M1. For each stump, the best split is determined by minimizing the weighted miss-classification error. Stumps are sequentially trained, and their predictions are combined using weighted alphas to form the final classifier.
    - Decision stumps are trained using the train set. Each decision stump is grown as a decision tree with a maximum depth of 1. The best split for each dimension is found by minimizing the Sum of Squared Residuals (SSR).
- **Updating Labels and Training Stumps:** Labels are updated based on negative gradients for absolute loss. Stumps are sequentially trained and updated for 300 iterations, with labels and weights adjusted at each iteration.
- **Evaluation:**

    - The accuracy of the classifier is evaluated on the validation set after each iteration to monitor its performance. The best stump with the highest accuracy on the validation set is selected for evaluation on the test set.
    - Mean Squared Error (MSE) is computed on the validation set after each iteration to track the model's performance. The tree with the lowest MSE on the validation set is selected as the best tree, and its performance is evaluated on the test set.

3. Results:

- Accuracy of h1(x) on the validation set: <Accuracy Value>
- Best accuracy on the validation set: <Best Accuracy Value> at iteration <Best Stump Index>
- Accuracy on test set using the best stump: <Test Accuracy Value>
- MSE on Training Set using Decision Stump h1(x): <MSE value>
- MSE of Residue using y - 0.01 * h1(x): <MSE value>
- MSE of Residue using y - 0.01 * h1(x) - 0.01 * h2(x): <MSE value>
- MSE on Validation Set vs. Number of Trees: (Refer to the plotted graph)
- Lowest MSE on Validation Set: <MSE value> (Tree <tree index>)
- MSE on Test Set using the Best Tree: <MSE value>

4. Discussion:

- The results demonstrate the effectiveness of AdaBoost.M1 in improving classification accuracy through ensemble learning. The iterative training of decision stumps with updated weights allows the model to focus on difficult-to-classify instances, leading to improved performance. The accuracy achieved on the test set validates the robustness of the selected best stump and its generalization to unseen data.
- The project demonstrated the effectiveness of gradient boosting with absolute loss for regression tasks. The iterative training of decision stumps and updating of labels based on negative gradients improved model accuracy. The MSE analysis on the validation set helped in selecting the best-performing tree for evaluation on the test set.

5. Conclusion:

- AdaBoost.M1 applied to the MNIST dataset for binary classification of digits 0 and 1 yielded promising results. The project highlights the importance of ensemble learning techniques in enhancing classification accuracy and generalization. Future work could explore the application of AdaBoost.M1 to multiclass classification tasks and investigate other boosting algorithms for comparison.
- Gradient boosting with absolute loss is a powerful technique for regression problems, providing accurate predictions with careful tuning of hyperparameters and iterative training of weak learners. The project's results highlight the importance of model selection and evaluation using validation and test sets in machine learning tasks.