HW2 - Advanced Machine Learning Report

This summary provides an overview of the implementation and evaluation of Gradient Boosting and AdaBoost models.

The aim of this analysis was to explore the performance of the two ensemble learning algorithms and compare their results. Additionally, various evaluation metrics and visualization techniques were employed to assess the models' performance and understand the impact of hyperparameter tuning.

Data Generation and Visualization

Firstly, we generated two datasets named "Blobs" and "Circles," each comprising binary labels. Our intention was to construct datasets characterized by non-linear associations, deliberately avoiding the presence of a definitive decision boundary. This design choice was motivated by our desire to generate datasets in which the labels overlap, making their separation impossible.

To gain insights into the data and facilitate model visualization we created a 2d scatter plot to examine how the labels spread across the euclidean space, for plotting higher dimensions we used the first two components of PCA exclusively.

Gradient Boosting Model

After data generation and dimensionality reduction, the Gradient Boosting Regression Trees (GBRT) Estimator was implemented. This ensemble learning algorithm sequentially builds an ensemble of weak decision tree models to improve the predictive accuracy. The weak decision tree was utilized using Scikit-Learn Decision Tree and The performance of the GBRT Estimator was evaluated using the Mean Squared Error (MSE) loss metric.

AdaBoost Model

In addition to the GBRT estimator, the AdaBoost algorithm was implemented and evaluated on the same datasets. AdaBoost is another popular ensemble learning algorithm that iteratively combines multiple weak decision stumps to create a strong model. The performance of the AdaBoost model was compared with that of the GBRT model to assess its efficiency.

Demonstration and Evaluation

To gain further insights into the Estimator's performance, several evaluation metrics were employed.

1. A classification report was generated, providing information on precision, recall, F1-score, accuracy and support for each class. This report helped assess the estimator's ability to correctly classify instances from the generated datasets.
2. A confusion matrix was created to visualize the distribution of predicted and actual labels, enabling a deeper understanding of the model's predictive behavior.
3. A decision boundary plot was generated to visualize the model's separation of classes. This plot allows us to explore if overfitting occurs and to validate the correctness of the model (e.g. to see if the separation is non-linear).
4. A learning curve plot was constructed to compare the test scores with the training scores, allowing an assessment of the model's generalization performance.

Hyperparameter Tuning

To optimize the performance of both models, a grid search was conducted to tune the hyperparameters. Grid search is a technique that systematically explores different combinations of hyperparameters to find the optimal configuration. By fine-tuning the hyperparameters, it was possible to enhance the estimators' predictive capabilities and improve overall performance.

Conclusion

In conclusion, this analysis focused on implementing and evaluating Gradient Boosting and AdaBoost models on generated datasets with binary labels. The estimators' performances were assessed using various evaluation metrics, including MSE loss, classification reports, confusion matrices, decision boundary plots, and learning curves. The results indicated that the GBRT estimator exhibited slightly better performance compared to that of the AdaBoost in this particular scenario. Hyperparameter tuning through grid search barely enhanced the models' performance.