Advanced Machine Learning, 2023

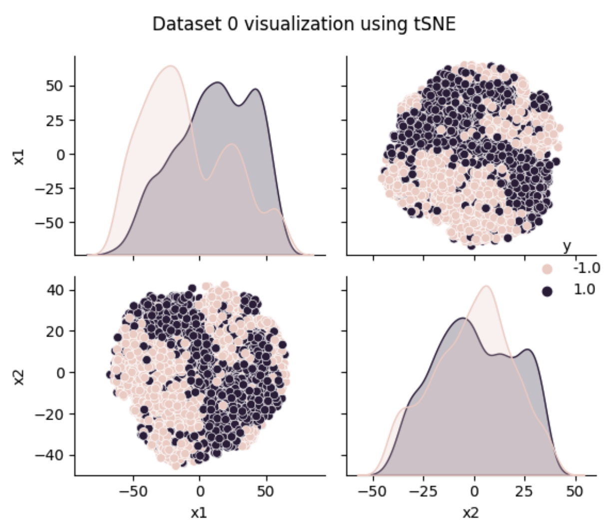
Home Assignment 2 – Ensemble Learning

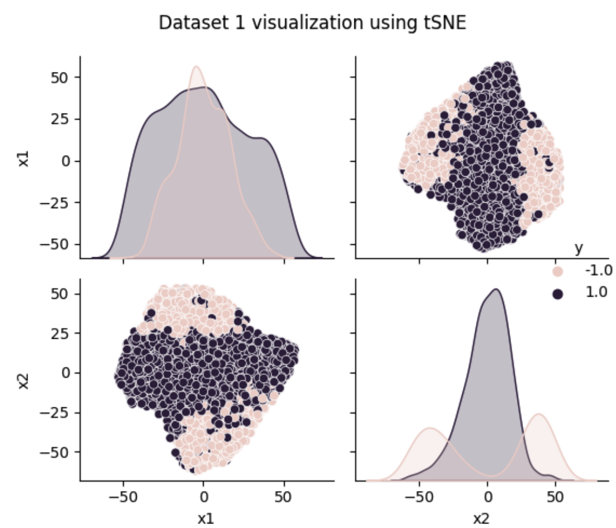
Abstract

In this assignment, we implemented two ensemble boosting algorithms: Gradient Boosted Regression Trees and AdaBoost. We generated datasets and experimented with different hyperparameters, evaluated the models' results, and compared them.

Part 0 - Dataset Generation

We generated two datasets using skleran’s `make_classification` method. Both are binary 5k-samples binary datasets (1/-1 labeled), where dataset 0 is balanced with 50 features and dataset 1 is imbalanced with 25 features. Here is a visualization of the datasets using tSNE dimensionality reduction, pair-plotted using seaborn:





Part 1 - Gradient Boosted Regression Trees

We implemented the algorithm as shown in class. We ran an experiment to evaluate the model’s performance over the generated datasets using multiple combinations of hyperparameters.

The overall performance was better on dataset 1 because it has fewer features and less noise. Moreover, the largest T we experimented with (200) performed best in both cases. A larger alpha (0.5) was better in dataset 1 than the 0.1

Dataset	Best Model	Precision	Recall	F1-Score	Support	Accuracy
0	GBRT - T: 200, alpha: 0.1, max_depth: 2	0.85 (class -1), 0.88 (class 1)	0.89 (class -1), 0.84 (class 1)	0.87 (class -1), 0.86 (class 1)	754 (class -1), 746 (class 1)	0.864
1	GBRT - T: 200, alpha: 0.5, max_depth: 2	0.92 (class -1), 0.95 (class 1)	0.91 (class -1), 0.96 (class 1)	0.92 (class -1), 0.95 (class 1)	448 (class -1), 802 (class 1)	0.94

chosen for dataset 0. The reason for that might be the noise in the data, where large learning rates might lead to bad performance in terms of convergence.

Part 2 - AdaBoost

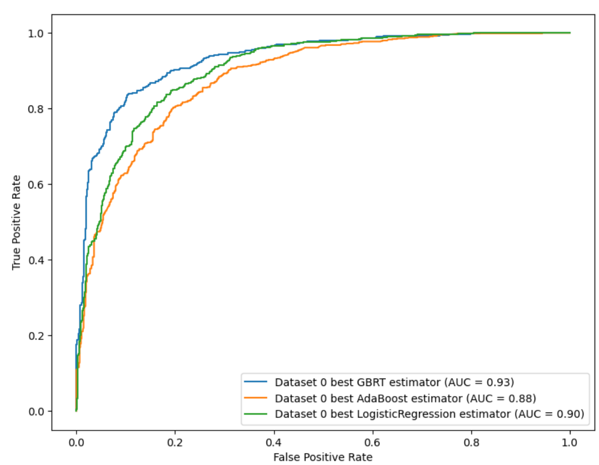
We implemented the algorithm as shown in class. We ran an experiment to evaluate the model's performance over the generated datasets using multiple combinations of hyperparameters.

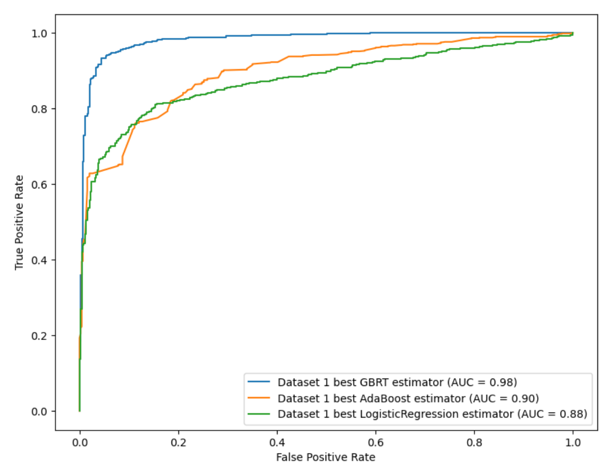
The overall performance was better on dataset 0. AdaBoost is known to be robust to noise, as it can effectively weight the importance of each training example and focus more on the informative examples. This ability to focus on informative examples may allow AdaBoost to perform well on datasets with more noise. In our case, dataset 0 has more informative features (10 vs. 5), which can lead to better classification results.

Dataset	Best Model	Precision	Recall	F1-Score	Support	Accuracy
0	AdaBoost - T: 100	0.78 (class -1), 0.81 (class 1)	0.82 (class -1), 0.77 (class 1)	0.80 (class -1), 0.79 (class 1)	754 (class -1), 746 (class 1)	0.7946666666666666
1	AdaBoost - T: 200	0.82 (class -1), 0.76 (class 1)	0.48 (class -1), 0.94 (class 1)	0.60 (class -1), 0.84 (class 1)	448 (class -1), 802 (class 1)	0.776

Part 3 - Base Model (Logistic Regression)

We fitted a baseline model (Logistic Regression) to validate our classifier implementations. After a quick hyper-params optimization, we plotted the ROC/AUC curves of all best models from each type:





Generally, GBRT wins in comparison to the others. Generally, there is an advantage to the boosting algorithms in dataset 1, and AdaBoost seems in-par with LR for dataset 0.