RUNI 3945: Advanced ML

Homework 2

Spring Semester 2022-23 TASHPAG

Submitted by:	Yarden Fogel ID: 011996279 (yarden.fogel@post.runi.ac.il)
Partner:	Roni Ben Dom ID: 207576463 (roni.bendomi@post.runi.ac.il)
Date:	May 28, 2023

Part 4: Report - Summary and Conclusions:

In this task we implemented both GBRT (Gradient Boosted Regression Trees) and AdaBoost (Adaptive Boosting), two popular machine learning algorithms used for regression and classification tasks.

GBRT is an ensemble method that combines multiple decision trees to create a powerful predictive model. It builds trees in a sequential manner, where each subsequent tree is trained to correct the mistakes made by the previous trees. This iterative process reduces the overall error and improves the model's accuracy. GBRT is known for its ability to handle complex relationships and capture non-linear patterns in the data.

AdaBoost, on the other hand, is a boosting algorithm that combines weak learners (typically decision trees) to create a strong classifier. In AdaBoost, each weak learner is trained on a modified version of the dataset, where more weight is given to the misclassified instances from the previous iterations. The final prediction is made by combining the predictions of all the weak learners, weighted by their accuracy. AdaBoost is particularly effective in situations where the weak learners are only slightly better than random guessing.

While both GBRT and AdaBoost are ensemble methods and involve combining multiple weak learners, there are several key differences between them:

1. Training Process: GBRT builds trees sequentially by minimizing a loss function at each step, correcting the mistakes made by previous trees by updating the residuals, while AdaBoost modifies the weights of the instances to emphasize the misclassified ones in subsequent iterations.

2. Weak Learners: GBRT typically uses decision trees as weak learners, whereas AdaBoost can use any weak classifier (e.g., decision trees, SVMs, or neural networks).

3. Error Correction: GBRT corrects the mistakes made by the previous trees by updating the residuals, while AdaBoost focuses on the misclassified instances by adjusting their weights.

4. Complexity: GBRT can capture complex relationships and handle non-linear patterns, making it suitable for a wide range of tasks. AdaBoost, on the other hand, is generally simpler and may struggle with highly complex datasets.

GBRT Advantages and Shortfalls:

Advantages: GBRT models are generally more accurate, as they minimize a loss function iteratively by fitting trees on the residuals. They can effectively handle mixed data types and

handle missing data well. GBRT can also be easily parallelized, improving efficiency.

Shortfalls: GBRT models can be sensitive to hyperparameters and require more tuning. The models may overfit if the number of iterations or depth of the trees is too large. Additionally, the sequential nature of boosting makes it harder to parallelize compared to bagging methods.

AdaBoost Advantages and Shortfalls:

Advantages: AdaBoost models are relatively simple and easy to implement. They are less prone to overfitting and can achieve good generalization with appropriate hyperparameter choices. AdaBoost models can perform well even with weak classifiers and can be easily parallelized, enhancing efficiency.

Shortfalls: AdaBoost models can be sensitive to noisy data and outliers, which can lead to degraded performance. They can also be slower to train due to the sequential nature of the algorithm.

Hyperparameters and step-size α:

Hyperparameters: In both GBRT and AdaBoost, the choice of hyperparameters such as the number of estimators, learning rate, and tree depth all impacted model performance. In our experiments, we chose hyperparameters based on their general effectiveness in practice and our parameter tuner, which allowed for fair comparisons between our custom models and the sklearn model classes.

Step-size / α: In AdaBoost, α determines the weight of the weak classifier in the final ensemble. Larger alpha α values lead to more aggressive boosting, which has the potential to lead to overfitting, while smaller alpha α values can result in underfitting if the alpha used is too small. It's for these reasons that choosing an appropriate α is essential to balance the all-important trade-off between underfitting and overfitting.

Dataset Design and Model Results:

In order to evaluate our algorithms, we created three datasets for experimentation and testing, with 20 features and 1000 samples each, and distinguished in order of separability of the clusters of labels, with the first dataset being relatively easy to separate, the second dataset being moderate, and the third dataset being very difficult to separate clusters with lots of overlap.

We ran both GBRT and AdaBoost using the sklearn classes, but for robustness, also created our own manual or custom GBRT and AdaBoost models. To ensure the correctness of our custom implementations and improve interpretability, we compared our results to those from the sklearn libraries, and generated the following results:

Original Results:

After much deliberation and debugging, we decided to take another crack at the datasets at the last-minute, scrambled up our original 3 datasets with varying seeds and separability, and then created a FOURTH dataset that was designed and inspired by volcano activity, to predict whether a volcano is dormant (it's all good) or active (run for your lives!!!). This dataset also

	Dataset1_Train	Dataset1_Test	Dataset2_Train	Dataset2_Test	Dataset3_Train	Dataset3_Test
sklearn_GBRT	0.99	0.98	0.99	0.97	0.91	0.84
custom_GBRT	1	0.98	1	0.96	0.94	0.86
sklearn_AdaBoost	0.99	0.98	0.99	0.97	0.81	0.77
custom_AdaBoost	0.99	0.98	0.98	0.97	0.8	0.72

was impossible to separate between the class clusters, as we wanted to make it particularly challenging on our ensembles of weak learners. We present those updated results below.

Latest Results with Updated and New Datasets:

	Dataset1_Train	Dataset1_Test	Dataset2_Train	Dataset2_Test	Dataset3_Train	Dataset3_Test	Dataset4_Train	Dataset4_Test
sklearn_GBRT	1	0.98	0.99	0.97	0.88	0.73	1	0.95
custom_GBRT	1	0.98	1	0.97	0.98	0.73	0.98	0.84
sklearn_AdaBoost	0.99	0.98	0.97	0.97	0.76	0.74	0.58	0.59
custom_AdaBoost	1	0.98	0.99	0.97	0.78	0.69	0.61	0.58

Final Thoughts:

As the above results demonstrate, the models had a harder time with the overlapping less-separated label clusters in dataset 3, particularly AdaBoost, where accuracy dropped from nearly perfect to somewhere in the 70s. The gap in performance was even more apparent on the overlapping complex volcano dataset we later introduced, with accuracies dropping across the board, but AdaBoost's scores of 58-61% were particularly underwhelming. Also of note is that the AdaBoost models (both sklearn and our own custom version) at times showed very little sensitivity to changing the n_estimators parameter, or number of weak learners, which is puzzling. Aside from that, GBRT performed better overall than AdaBoost, though that is largely influenced by AdaBoost falling well short on dataset 3 (and then also on the volcano dataset), which may or may not have been partiallly related to the issue with lack of sensitivity to n_estimator mentioned above.

[0]Compiled with LaTeX on May 28, 2023.