Assignment 2: Adaboost

TAK YIN PANG
University of Adelaide
a1796036@adelaide.edu.au

Abstract

Adaboost

1. Introduction

Boosting is an ensemble method to improve the performance of weaker learner by training a sequence of these weak model and combine them into a strong learner. In this assignment, I will go into the details of Adaboost, short for adaptive boosting, from the implementation to the retionale behind.

2. Description of the algorithm

Adaboost combines a weighted sum of a sequence of weak learning algorithms to a strong learning algorithm. Weak learning algorithms is a weak learner that predict slightly better than a random guess. Through adaboost, a sequence number of weak learner can ensemble into a strong learner that the prediction has a relatively high accuracy.

2.1. Adaboost algorithm

The idea of adaboost is to train a serie of weak learners with a number of iteration N by adaptively changing the weight distribution D_t over iteration of the data points$(x_1, y_1), ..., (x_m, y_m)$, $x_i \in X, y_i \in \{-1, 1\}$ and combine these weak learners $h(x)$ with respect to a weight α into a final strong learner $H(x)$.

The pseudocode mentioned in the paper Explaining AdaBoost [5] is as below:

Given Data: $(x_1, y_1), ..., (x_m, y_m), x_i \in X, y_i \in -1, 1$
Initialize: $D_1(i) = \frac{1}{m} for i = 1, ..., m$
For $t = 1, ..., N$:
Train weak learner h_t using D_t
$h_t : X \to \{-1, +1\}$
Select the h_t with lowest weighted error ϵ_t

$$\epsilon_t = Pr_(i \sim D_t)[h_t(x_i) = y_i]$$

Choose the weak learner weight α_t

$$\alpha_t = \frac{1}{2} ln \frac{1-\epsilon_t}{\epsilon_t}$$

Update the weight distribution $D_{t+1}(i)$ for i,...m:

$$D_{t+1}(i) = \frac{D_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

where Z_t is a normalization factor,
such that $\sum_i D_{t+1}(i) = 1$

The final prediction:

$$H(x) = sign(\sum_t \alpha_t * h_t(x)) \forall t \in N$$

2.2. Explaination of the algorithm

Adaboost is a simple, fast and elegant algorithm that boost a weak learning algorithm to a strong learning algorithm by altering the weights of the data points during training a serie of weak learners so as to address the prediction error in the previous weaker learner.

Initially, the weight of the data point are equal and sum up to 1, $D_1(i) = \frac{1}{m}$. After using D_1 to train the weak learner h_1, the weight of the datapoints that are incorrectly classified will increase by a factor of e^{α_t} and decrease by a factor of $e^{-\alpha_t}$ if correctly classified. This update penalizes the next weaker learner if it misclassifies the same data points as the previous weaker learner. Hence the next weak learner trained on the weighted disturbation will try to minimize the weighted error as much as possible. This process will repeat for a number of iteration N.

The final prediction aggregates the weak learners with the corresponding weights α_t. Every weak learners vote for the prediction and α_t determines how much they contribute to the final output.

2.3. Classifier weight - Alpha

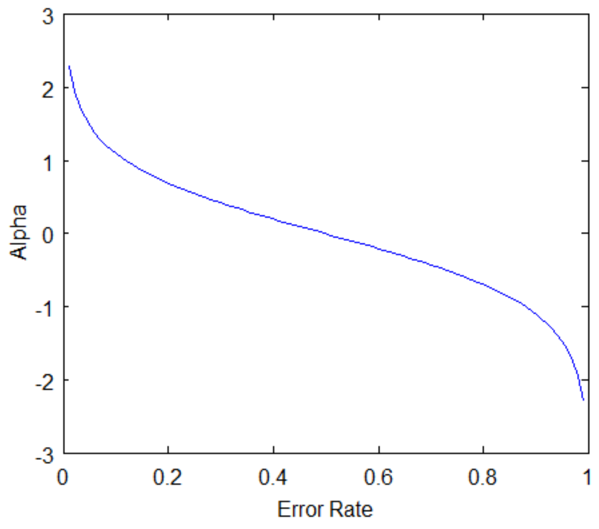

Figure 1. Alpha values against Error Rate [3]

$$\alpha_t = \frac{1}{2} ln \frac{1-\epsilon_t}{\epsilon_t}$$

The alpha value is defined as the above formula.From Figure 1, there is three intuitions to note. First, the alpha value is zero when the error rate is 0.5. Since $\epsilon = 0.5$ means that the classifier is the same as random guess, it is reasonable to assign zero weight to the classifier. Second, the alpha values grow exponentially when the error approaches zero. More weight is assigned to the classifier if the error rate is low. Third, the alpha values become negative when the error is larger than 0.5. Negative weight means that do the opposite prediction for these poorly performed classifiers.

2.4. Weak learner and decision stump

Weak learner is defined as the classifier with a prediction error $\epsilon <= 0.5$. Decision stump is the classic default choose of the weak classifier for the adaboost algorithm but the weak learner choice in adaboost algorithm can be any classifier perform better than randon guess.

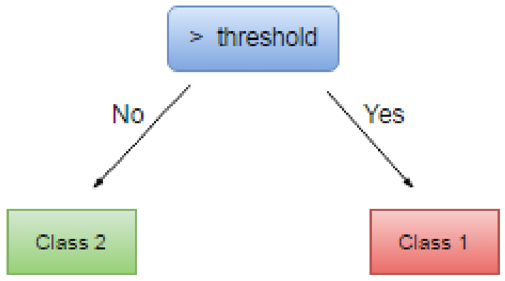

Figure 2. Decision Stump [2]

Decision stump is a decision tree with 1 depth which means that it splits the data points into two class nodes based on one attribute(feature) value threshold. In Figure 2, it illustrates how a decision tree looks like.

To determine the goodness of the spilt, there is several criterions including gini impurity and information gain.

For gini impurity, the split attribute value is selected if the impurity is the least.

$$gini = 1 - \sum_c p(c)^2$$
$$impurity = \frac{leftSample}{totalSample} * leftGini + \frac{rightSample}{totalSample} * rightGini$$

For information gain, the split attribute value is selected if the gain is the highest.

$$entropy = \sum_c p(c) log(p(c))$$
$$informationGain = \frac{leftSample}{totalSample} * leftEntropy + \frac{rightSample}{totalSample} * rightEntropy$$

3. My understanding of Adaboost

In this section, I will talk about several properties about the adaboost algorithm.

3.1. Hyperparameter

Adaboost, unlike the other machine learning algorithms, has less hyperparameter to choose during the training process. For adaboost, there are two hyperparameters to tune. The frist one is the choice of weak learner. Although decision stump is a default choice for the adaboost algorithm, every classifier that is better than a random guess can be selected as the weak classifier. The second one is the number of iteration. The number of iteration highly depend on the dataset. I will talk about them one by one.

3.1.1 Choice of classifier

For a weak classifier like decision stump, it does not fit the training data well and has a high bias. Adaboost algorithm works well on reducing the bias by boosting the weak learners in sequencial order. The weight put on the misclassified data points helps the next classifier corrects the mistakes made by the last classifier and hence help to reduce the bias.

In theory, every classifier that is better than random guess (prediction error ¡ 0.5) can be applied to adaboost algorithm. However, the result of ensembling strong classifiers in adaboost may or may not perform well. This may be because the strong classifier has already low bias and it does not help to push the bias lower by boosting.

3.1.2 Number of iteration

The number of iteration depends on the problem and the training data. The default number of iteration in sklearn adaboost classifier is 50. Some suggest early stopping [1] should be applied while Mease and Wyner (2008) [4] argue adaboost should run for a long time until it converges.

3.2. Overfitting

Although adaboost might overfit, it is quite robust to overfitting. It depends on several factors [7] whether it will overfit or not. The first one is the ”strength” of the ”weak” learners. It is less likely to overfit if the weak learner is very simple like decision stump. The second one is the noise level in the data. Adaboost is prone to overfitting on the noisy datasets. The last one is the dimensionality of the data. It is easier to overfit in high dimensional spaces.

3.3. Outliner

Adaboost is excellent in identifying the outliners in the training data. [6]This is becasue weight will repeatly put on the outliners during the iteration process. The final weight on the those outliners will become very high because most weak classifiers fail to classify them correctly.

4. Analysis

4.1. Introduction

In this assignment, I decide to experiment on the different setting of the adaboost algorithm on breast cancer dataset. Frist, I will compare the performace on the adaboost algorithm with the sklearn AdaboostClassifier as well as the parameter. Then, I will explore the impact of number of iteration and the depth of the decision tree classifier used in adaboost algorithm. Lastly, I will compare the performace of SVM classifier with different kernels with adaboost.

4.2. Experiment set up

In the experiment, Wisconsin Diagnostic Breast Cancer dataset is used for testing the adaboost algorithm. The first 300 samples are the training data and the remaining 269 samples are the testing data. The target label is changed from M,B to -1,1 for adaboost. Sklearn DecisionTreeClassifier is used as the base weak learner for both implemented adaboost and 3rd party adaboost. Gini coefficient is used for the split criterion.

4.3. Comparison with 3rd party adaboost

	adaboost	sklearn
train accuracy	1	1
test accuracy	0.977695	0.977695
train time(s)	0.185496	0.257749
test time(s)	0.0143885	0.0128233
alpha ratio	0.5	1

Table 1. Performance comparison with 3rd party sklearn adaboost

As shown in table 1, the train and test accuracy performance of both algorithm are exactly the same. However, the time performance of my implementation is faster than the sklearn model. I guess the reason is that my impletation did not include parameter like learning rate and random state which are in sklearn adaboost classifier. Then I further compare the alpha values and error rate in each weak classifier.

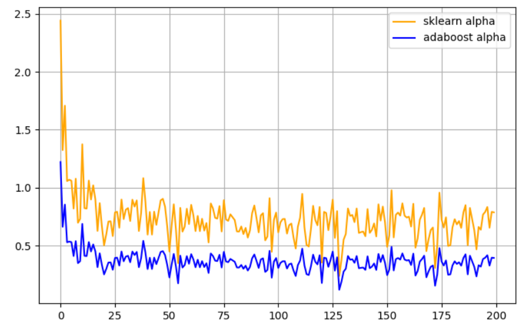

Figure 3. Alpha values

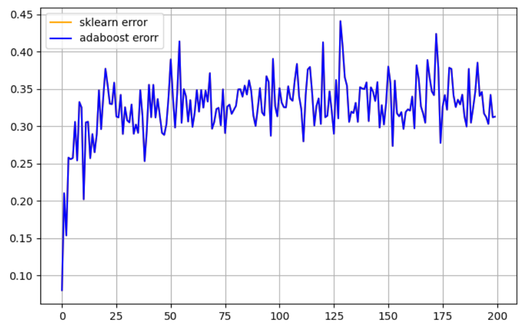

Figure 4. Error rate of the weak classifiers in each iteration

In figure 3, it shows the values of alphas of each iteration. The values in sklearn model is double to the values of my implementation but it will not affect the prediction of the final ensemble models.

In figure 4, the error rates of each weak classifier are exactly the same. Therefore, I can conclude that my implementation perform as well as the sklearn implementation. And yet the run time of my implementation is even slightly faster.

4.4. Performance with different setting

In this section, I experiment the impact on the performance of the adaboost algorithm of the depth of decision tree as well as number of iteration. I run adaboost with decision tree depth from 1 to 3 and number of iteration from 1 to 200 and plot the train error vs test error agaist number of iteration with respect to different depth.

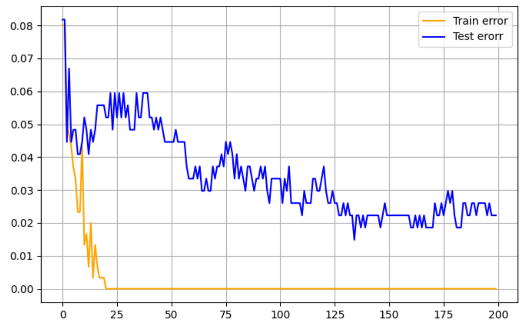

Figure 5.Train vs Test error with decision tree Depth 1

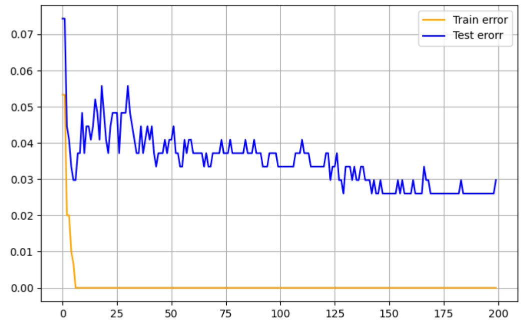

Figure 6.Train vs Test error with decision tree Depth 2

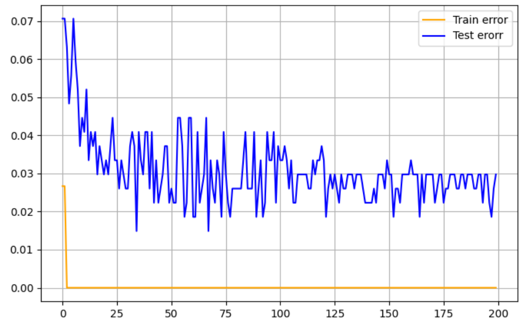

Figure 7.Train vs Test error with decision tree Depth 3

The above figures show the the training error as well as the testing error with different depth against the number of iteration. All of them show that the train error reduce to zero very quickly while the test error drop significantly at frist then slowly converage at around 150 iterations. In this dataset, overfitting does not seem to occur and the test error still fluctuate for a long time after the train error dropped to zero.

The training error of the one with decision tree depth of 1 decreased to zero much slower then the other one with higher depth. However, the test error reached by the simpler model is the lowest at 0.015 at around 130 iteration. Meanwhile the test error with decision tree depth of 3 decrease faster but fail to converage in a low error rate than the one with lower depth.

4.5. Performance comparison with SVM classifier

kernel	adaboost	linear	poly	rbf
train acc	1	0.96	0.89	0.9
test acc	0.977695	0.955390	0.933085	0.914498
train time	0.185496	0.557233	0.001250	0.001836
test time	0.014388	0.000352	0.000705	0.001233

Table 2. Performance comparison with SVM classifiers

In table 2, it shows that adaboost algorithm outperform all the svm classifier with different kernels and parameter C = 1. Among three kernels, SVM classifier with linear kernel performs the best and reach a 0.96 train accuracy and 0.95539 test accuracy. Also, it has the slowest train time but the fastest test time.

4.6. Boosting on SVC

I tried to applied the above svm classifier on the adaboost algorithm. Since they are not weak classifier, I would like to experiment on the behavior applying strong classifier on adaboost algorithm. I run them on adaboost for 10 iteration and plot the error curve for each of them.

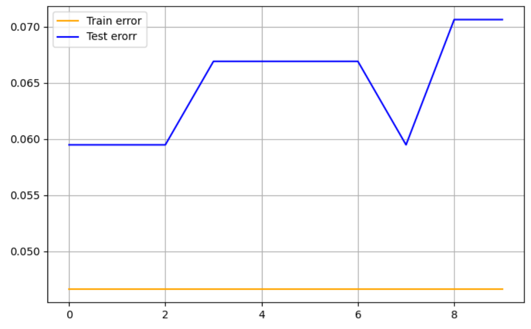

Figure 8. linear kernel

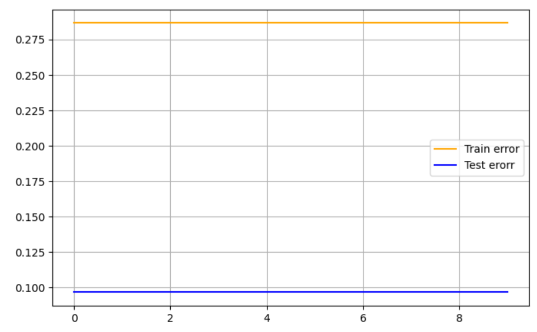

Figure 9. poly kernel

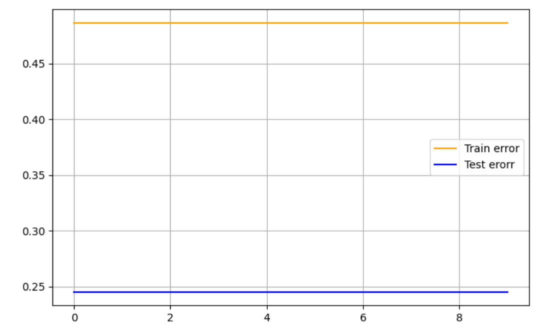

Figure 10. rbf kernel

The results show a very weird behavior on each of those kernel. I can hardly draw any conclusion from the results but it is very interesting.

5. Code

The code can be simply run by:
python3 Adaboost.py

6. Conclusion

The adaboost algorithm perfroms well on this dataset and achieve a very high train and test accuracy. Also, overfitting does not happen in this dataset and the algorithm converages slowly with more than 100 iterations. My implementation performs as well as the 3rd party adaboost classifier.

Adaboost has a better boosting performace on weak classifers. When using decision tree with depth of 3, it does not perform as well as the simpler model. Also it suffers a very strange bahavior when appling strong classifiers like svm classifier on the adaboost.

References

[1] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000, 1998.

[2] Jason Leaster. Machine learning with boosting.

[3] Chris McCormick. Adaboost tutorial.

[4] David Mease and Abraham Wyner. Evidence contrary to the statistical view of boosting. *J. Mach. Learn. Res.*, 9:131156, June 2008.

[5] Robert Schapire. Explaining adaboost. pages 37–52, 10 2013.

[6] Robert E. Schapire. A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, page 14011406, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[7] tdc (https://stats.stackexchange.com/users/7365/tdc). Is adaboost less or more prone to overfitting? Cross Validated. URL:https://stats.stackexchange.com/q/20774 (version: 2012-01-08).