AdaBoost Algorithm

(ML)

00

What is AdaBoost Algorithm?

AdaBoost, also called Adaptive Boosting, is a technique in Machine Learning used as an **Ensemble Method(Boosting)**. The most common estimator used with AdaBoost is decision trees with one level which means Decision trees with only 1 split. These trees are also called **Decision Stumps.**

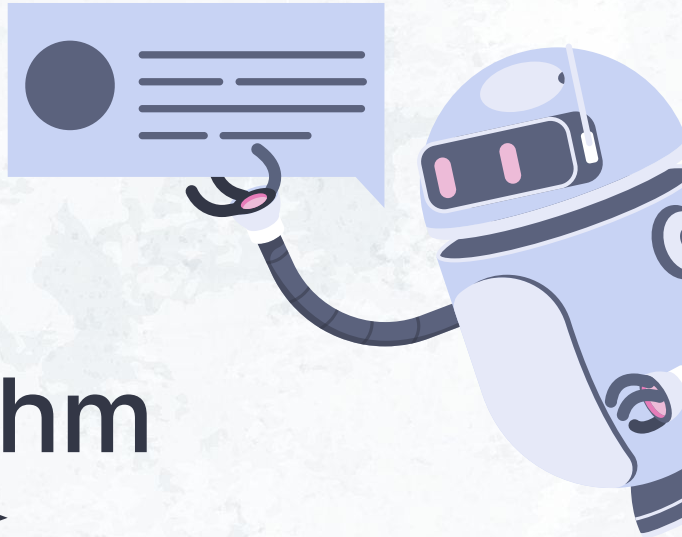
Three Ideas Behind AdaBoost

01 → AdaBoost combines a lot of "weak learners" to make classifications. The weak learners are almost aways stumps.

02 → Some stumps get more say in the classification than others.

03 → Each stump is made by taking the previous stump's mistakes

Understanding the Working of the AdaBoost Algorithm



Step 1 – Creating the First Base Learner

To create the first learner, the algorithm takes the first feature, i.e., feature 1 and creates the first stump, f1. It will create the same number of stumps as the number of features. We must calculate Gini Index the same way it is calculated for decision trees. The stump with the least value will be the first base learner.

Step 2 – Calculating the Total Error (TE)

The total error is the sum of all the errors in the classified record for sample weights.

Step 3 – Calculating Performance of the Stump (“**Amount of Say**”)

$$Performance\ of\ the\ stump\ =\ \frac{1}{2}\log_e(\frac{1-Total\ Error}{Total\ Error})$$

Note: Total error will always be between 0 and 1.

0 Indicates perfect stump, and 1 indicates horrible stump.

Step 4 – Updating Weights

The wrong predictions will be given more weight, whereas the correct predictions weights will be decreased. Now when we build our next model after updating the weights, more preference will be given to the points with higher weights.

$$New\ sample\ weight\ =\ old\ weight\ *\ e^{\pm Amount\ of\ say\ (\alpha)}$$

The amount of, say (alpha) will be **negative** when the sample is correctly classified.

The amount of, say (alpha) will be **positive** when the sample is miss-classified.

Step 4 – Updating Weights

We know that the total sum of the sample weights must be equal to 1, but here if we sum up all the new sample weights it is not equal to 1. To bring this sum equal to 1, we will normalize these weights by dividing all the weights by the total sum of updated weights

Step 5 – Creating a New Dataset

We are almost done. Now, what the algorithm does is selects random numbers from 0-1. Since incorrectly classified records have higher sample weights, the probability of selecting those records is very high. Size of New dataset is must same with old dataset

Iterate through these steps until and unless a low training error is achieved.

Understand with Data

Step 1 - Find first stump & assign sample weight

Row No.	Gender	Age	Income	Illness	Sample Weights
1	Male	41	40000	Yes	1/5
2	Male	54	30000	No	1/5
3	Female	42	25000	No	1/5
4	Female	40	60000	Yes	1/5
5	Male	46	50000	Yes	1/5

The formula to calculate the sample weights is:

$$w(x_i, y_i) = \frac{1}{N}, \; i = 1, 2, \ldots . n$$

Gender has The lowset gini index

Step 2 & 3 - Calculate TE & Amount of Say

Here in our dataset, let's assume there is 1 wrong output, so our total error will be 1/5, and the alpha (performance of the stump) will be:

$$Performance\ of\ the\ stump\ =\ \frac{1}{2}\log_e(\frac{1-Total\ Error}{Total\ Error})$$

$$\alpha\ =\ \frac{1}{2}\log_e\left(\frac{1-\frac{1}{5}}{\frac{1}{5}}\right)$$

$$\alpha\ =\ \frac{1}{2}\log_e\left(\frac{0.8}{0.2}\right)$$

$$\alpha\ =\ \frac{1}{2}\log_e(4)\ =\ \frac{1}{2}*(1.38)$$

$$\alpha\ =\ 0.69$$

Step 4 - Updating Weight

There are four correctly classified samples and 1 wrong. Here, the ***sample weight*** of that datapoint is *1/5,* and the ***amount of say/performance of the stump*** of ***Gender*** is *0.69*.

New weights for *correctly classified* samples are:

$$New\ sample\ weight\ =\ \frac{1}{5}\ *\ \exp(-0.69)$$

$$New\ sample\ weight\ =\ 0.2\ *\ 0.502\ =\ 0.1004$$

For *wrongly classified* samples, the updated weights will be:

$$New\ sample\ weight\ =\ \frac{1}{5}\ *\ \exp(0.69)$$

$$New\ sample\ weight\ =\ 0.2\ *\ 1.994\ =\ 0.3988$$

Step 4 - Updating Weight

Row No.	Gender	Age	Income	Illness	Sample Weights	New Sample Weights
1	Male	41	40000	Yes	1/5	0.1004/0.8004 =0.1254
2	Male	54	30000	No	1/5	0.1004/0.8004 =0.1254
3	Female	42	25000	No	1/5	0.1004/0.8004 =0.1254
4	Female	40	60000	Yes	1/5	0.3988/0.8004 =0.4982
5	Male	46	50000	Yes	1/5	0.1004/0.8004 =0.1254

New Weight after Normalize

Step 5 - Make New data set

Row No.	Gender	Age	Income	Illness	New Sample Weights	Buckets
1	Male	41	40000	Yes	0.1004/0.8004= 0.1254	0 to 0.1254
2	Male	54	30000	No	0.1004/0.8004= 0.1254	0.1254 to 0.2508
3	Female	42	25000	No	0.1004/0.8004= 0.1254	0.2508 to 0.3762
4	Female	40	60000	Yes	0.3988/0.8004= 0.4982	0.3762 to 0.8744
5	Male	46	50000	Yes	0.1004/0.8004= 0.1254	0.8744 to 0.9998

Pickled tilapia will be selected from 0-1 as many as 5 to get a new data set. Values 0-1 will be referred to the bucket and the corresponding row will be instantiated. misclassified records have a larger sample size

New data set

Row No.	Gender	Age	Income	Illness
1	Female	40	60000	Yes
2	Male	54	30000	No
3	Female	42	25000	No
4	Female	40	60000	Yes
5	Female	40	60000	Yes

Suppose the 5 random numbers our algorithm take is 0.38, 0.26, 0.98, 0.40, 0.55.

Source:

- AdaBoost, Clearly Explained - Statquest (https://www.youtube.com/watch?v=LsK-xG1cLYA&t=455s)
- The Ultimate Guide to AdaBoost Algorithm | What is AdaBoost Algorithm? - Great Learning (https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/)
- Master the AdaBoost Algorithm: Guide to Implementing & Understanding AdaBoost - Analytics Vidhya (https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/)

Thanks!

Any questions?

azrhza9@gmail.com
linkedin.com/in/azri-ahza
https://github.com/Ahzaazr

CREDITS: This presentation template was created by **Slidesgo** and includes icons by **Flaticon**, infographics & images by **Freepik** and content by **Eliana Delacour**