Sanjivani Rural Education Society's

College of Engineering, Kopargaon-423603

DEPARTMENT OF COMPUTER ENGINEERING

Instruction No. 01 and 02 ML Lab / Sr. No.01 and 02 Rev 00 Date: 27/12/17	**Title: Assignment on Gradient Boost Classifier**

Aim: Implement **Gradient Boost Classifier** Algorithm on a given dataset.

Objectives: Implement the GBC on any dataset.

Input: .csv File

Theory:

Gradient Boost Classifier Algorithm:

Gradient Boosting is a functional gradient algorithm that repeatedly selects a function that leads in the direction of a weak hypothesis or negative gradient so that it can minimize a loss function. Gradient boosting classifier combines several weak learning models to produce a powerful predicting model.

Gradient Boosting in Classification

Gradient Boosting consists of three essential parts:

Loss Function

The loss function's purpose is to calculate how well the model predicts, given the available data. Depending on the particular issue at hand, this may change.

Weak Learner

A weak learner classifies the data, but it makes a lot of mistakes in doing so. Usually, these are decision trees.

Additive Model

This is how the trees are added incrementally, iteratively, and sequentially. You should be getting closer to your final model with each iteration.

Steps to Gradient Boosting

Gradient boosting classifier requires these steps:

Fit the model

Adapt the model's Hyperparameters and Parameters.

Make forecasts

Interpret the findings

Visualizing Gradient Boosting

1. The method will obtain the log of the chances to make early predictions about the data. Typically, this is the ratio of the number of True values to the False values.

2. If you have a dataset of six cancer occurrences, with four people with cancer and three who are not suffering, then the log(odds) is equal to log(4/3) 1.3, and the person who is free of cancer will have a value of 0. The person who has cancer will have a value of 1.

3. To make predictions, you must first convert the log(odds) to a probability with the help of a logistic function. Here, it would be around 1.3, the same as the log(odds) value of 1.3

4. Since it is greater than 0.5, the algorithm will use 1.3 as its baseline estimate for each occurrence.

e * log(odds) / (1 + e * log(odds))

5. The above formula will determine the residuals for each occurrence in the training set.

6. After completing this, it constructs a Decision Tree to forecast the estimated residuals.

7. A maximum number of leaves can be used while creating a decision tree. This results in two potential outcomes:

Several instances are into the same leaf.

The leaf is not a single instance.

You must use a formula to modify these values here:

ΣResidual / Previous Prob (1 - Previous Prob)]

8. You must now complete two things:

Obtain the log forecast for each training set instance.

Transform the forecast into a probability.

9. The formula for producing predictions would be as follows:

base_log_odds + (learning_rate * predicted residual value)

Advantages and Disadvantages of Gradient Boost

Advantages:

Frequently has remarkable forecasting accuracy.

Numerous choices for hyperparameter adjustment and the ability to optimize various loss functions.

It frequently works well with numerical and categorical values without pre-processing the input.

Deals with missing data; imputation is not necessary.

Disadvantages:

Gradient Boosting classifier will keep getting better to reduce all inaccuracies. This may lead to overfitting and an overemphasis on outliers.

Costly to compute since it frequently requires a large number of trees (>1000), which can be memory and time-consuming.

Due to the high degree of flexibility, numerous variables interact and significantly affect how the technique behaves.

Less interpretative, even though this can be easily corrected with several tools.

Output:

```
{'Gradient-Boost Default Test Score': 0.8708736373407032,
'Gradient-Boost GridSearch Test Score': 0.8785505911254414}
```

Conclusion:
Studied about Gradient Boost Classification and implemented it on income_evaluation.csv dataset for classification.

Prepared by:

Dr.T.Bhaskar

Subject Teacher

Approved by:

Dr. D.B. Kshirsagar

HoD-Computer