Random Forest
Epoch IIT Hyderabad

Chakka Surya Saketh
AI22BTECH11005

Introduction

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large.

The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favourably to Adaboost.

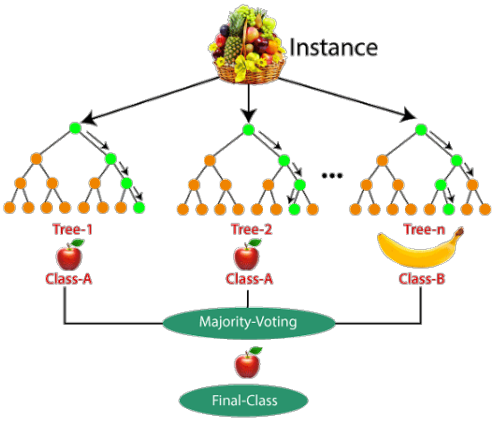
It is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. The algorithm's strength lies in its ability to handle complex datasets and mitigate over fitting, making it a valuable tool for various predictive tasks in machine learning.

Implementation

Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model. It uses two types of methods namely-

1) Bagging: It creates a different training subset from sample training data with replacement and the final output is based on majority voting.
2) Boosting: It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy.

Bagging is also known as Bootstrap aggregation, which is the technique used by Random forest. Now each model is trained independently with different, which generates results. It is called bootstrap.The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting, is known as aggregation.



Important features

1) Diversity
2) Immune to curse of dimensionality
3) Parallelization
4) No need for Train-Test split
5) Stability

Hyper parameters

- n_estimators: Number of trees the algorithm builds before averaging the predictions.
- max_features: Max features the forest considers splitting a node.
- mini_sample_leaf: Minimum number of leaves required to split a internal node.
- criterion: The criteria to split the node.(entropy/gini/log loss)
- max_leaf_nodes: Max leaf nodes in a tree.
- n_jobs: It tells the engine how many processors are allowed to be used. If the value is -1, it uses the max possible.

Conclusion

In conclusion, the Random Forest algorithm is a powerful and versatile machine learning tool that is widely used for both classification and regression problems. It is an ensemble method that combines the predictions of multiple decision trees to produce more accurate and robust results. One of the key strengths of the Random Forest algorithm is its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for a wide range of predictive tasks. Additionally, it is user-friendly and adaptable, with the ability to handle both continuous and categorical variables. Overall, the Random Forest algorithm is a highly effective and widely used tool in the field of machine learning.