Gradient Boosting Based Optimization Algorithm Using Model Selection to Reduce Overfitting

Moitra, Agnij (School: Birla Vidya Niketan)

Purpose — Gradient Boosting uses Decision Trees as estimators to create a feed forward loop in order to minimize the residuals errors, which is prone to overfitting and high variance. Thereby, it is proposed that a variational feed forward loop of a combination of many different estimators should be implemented. Method — Gradient Boosting based Optimization Algorithm (GBOA), makes an initial prediction as the mean of the dependent feature. Then it calculates the residuals and uses these residuals as the dependent feature for the next layer. In each subsequent layer besides just relying on Decision Trees, it simultaneously trains many linear and non-linear models and chooses the model with least validation mean squared error (MSE). This feed forward loop is repeated until the MSE value converges to zero or it has exhausted the iterations. Further, it is extended to Multioutput and Multiclass problems using Scikit-Learn’s Multi-Output and One-vs-the-rest classes. Results — XGBoost and GBOA were trained on the Higgs Boson Dataset, Covtype Dataset for classification, Accelerometer Data Set for regression, and Microsoft LTR Dataset for ranking. GBOA had a higher F1 score (for classification) and NDCG 10 (for ranking), and lower MSE (for regression). Further it had a lower difference of validation and training metrics when compared to XGBoost. Conclusion — The above benchmarking proves that testing various different machine learning models and making a feed forward loop would combat high variance and overfitting thereby also improving the overall metrics and leading to a more generalized model.