Question 5 : Gradient Tree Boosting

Parameters

We have performed a grid search on two parameters: max_depth and number of trees (n_estimators). But we weren't able to search the hyperparameter space in such a fine-grained manner as Random Forest because the Gradient Tree Boosting is very computationally expensive and it was taking hours to perform the search. Therefore, we analyzed the results of Random Forest heatmaps and decision trees to select a few good parameter settings :

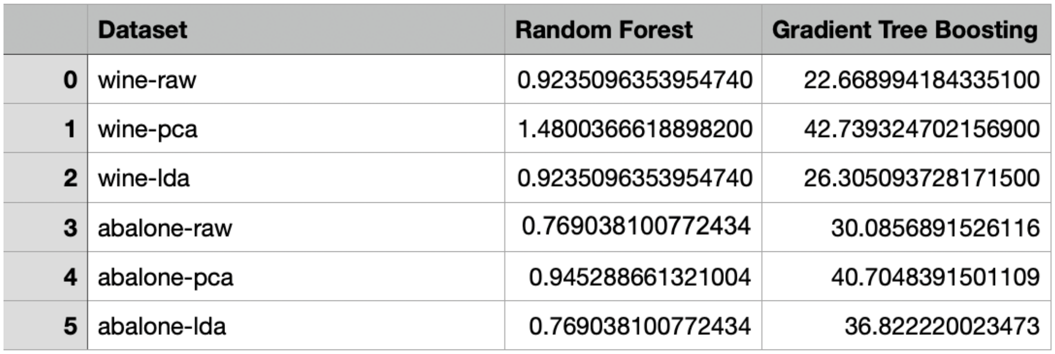
1. For the max_depth parameter, we have performed the search on these three values: **[6, 8 , 10]** for wine and **[3, 7 , 11]** for abalone
2. For the number of trees, we have performed the search on these three values: **[60, 100, 150**] for wine and **[80, 160, 180]** for abalone (in case of wine-raw, we added one more parameter i.e. 20)

Accuracy Scores for different parameter settings

- For abalone, the best accuracy of **25.35%** was obtained with **abalone-raw** dataset and the parameters: max_depth = 3 and 'n_estimators = 80.
- For wine, the best accuracy of **49.65%** was obtained with **wine-lda** dataset and the parameters: max_depth = 6 and 'n_estimators = 100.

Comparison of Runtime

- Random Forest seems to take much less time as compared to Gradient Tree Boosting for all the 6 datasets thus suggesting that the latter is a much much more computationally expensive model in comparison. On an average, it is taking more than 30 times more time. The values shown in Table I below are in *seconds* and the fit time is averaged over all the 9 grid search candidates, across all the 5-fold cross-validation fits.



	Dataset	Random Forest	Gradient Tree Boosting
0	wine-raw	0.9235096353954740	22.668994184335100
1	wine-pca	1.4800366618898200	42.739324702156900
2	wine-lda	0.9235096353954740	26.305093728171500
3	abalone-raw	0.769038100772434	30.0856891526116
4	abalone-pca	0.945288661321004	40.7048391501109
5	abalone-lda	0.769038100772434	36.822220023473

Table I : Runtime Comparison Wine and Abalone Dataset