CS 6375.004: Machine Learning - Spring 2023

Project #3

Kartikey Gupta (kartikey.gupta@utdallas.edu)

March 27, 2023

1 Scores of various tree/ensemble models on the given dataset

1.1 sklearn.tree.DecisionTreeClassifier

Clauses	Samples	criterion	splitter	min_sample_split	min_sample_leaf	**Accuracy**	**F1**
300	100	gini	best	2	2	0.605	0.603
	1000	log_loss	best	2	1	0.652	0.648
	5000	gini	best	2	10	0.734	0.741
500	100	entropy	best	2	10	0.625	0.611
	1000	entropy	best	2	1	0.683	0.682
	5000	entropy	best	2	1	0.736	0.738
1000	100	entropy	best	2	1	0.75	0.75
	1000	log_loss	best	2	1	0.774	0.774
	5000	entropy	random	4	5	0.844	0.844
1500	100	gini	best	2	5	0.86	0.855
	1000	gini	best	2	5	0.913	0.912
	5000	entropy	random	4	2	0.9.484	0.482
1800	100	entropy	random	4	5	0.955	0.955
	1000	entropy	best	2	1	0.97	0.97
	5000	log_loss	best	2	1	0.98	0.980

1.2 sklearn.ensemble.BaggingClassifier

Clauses	Samples	n_estimators	max_samples	max_features	oob_score	**Accuracy**	**F1**
300	100	100	1.0	1.0	False	0.725	0.744
	1000	100	1.0	1.0	False	0.845	0.842
	5000	100	1.0	1.0	False	0.900	0.905
500	100	50	1.0	1.0	False	0.72	0.728
	1000	100	1.0	1.0	False	0.846	0.844
	5000	100	1.0	1.0	False	0.916	0.917
1000	100	50	0.5	0.25	False	0.94	0.941
	1000	100	1.0	1.0	False	0.968	0.968
	5000	100	1.0	1.0	False	0.989	0.989
1500	100	50	1.0	0.25	False	1.0	1.0
	1000	100	0.25	0.25	False	0.99	0.99
	5000	50	1.0	0.25	False	0.9997	0.997
1800	100	50	1.0	1.0	False	0.98	0.98
	1000	50	0.25	0.25	True	1.0	1.0
	5000	100	0.5	0.25	True	1.0	1.0

1.3 sklearn.ensemble.RandomForestClassifier

Clauses	Samples	criterion	min_sample_split	min_sample_leaf	max_features	bootstrap	**Accuracy**	**F1**
300	100	gini	2	1	sqrt	True	0.725	0.744
	1000	gini	2	1	None	True	0.845	0.842
	5000	gini	2	1	None	True	0.9005	0.905
500	100	gini	2	5	sqrt	True	0.835	0.829
	1000	log_loss	2	5	sqrt	True	0.928	0.928
	5000	gini	2	1	sqrt	False	0.9455	0.9458
1000	100	log_loss	4	2	log2	False	0.965	0.9651
	1000	gini	2	5	sqrt	True	0.987	0.987
	5000	log_loss	4	5	log2	False	0.995	0.995
1500	100	gini	2	1	sqrt	True	1.0	1.0
	1000	gini	2	1	sqrt	True	0.999	0.9989
	5000	gini	2	1	2	True	0.999	0.9989
1800	100	gini	2	1	sqrt	True	1.0	1.0
	1000	gini	2	1	sqrt	True	1.0	1.0
	5000	gini	2	1	sqrt	True	0.999	0.9989

1.4 sklearn.ensemble.GradientBoostingClassifier

Clauses	Samples	learning_rate	n_estimators	subsample	criterion	min_sample_split	min_sample_leaf	max_features	**Accuracy**	**F1**
300	100	0.1	100	1.0	friedman_mse	2	5	sqrt	0.75	0.7641
	1000	0.1	400	0.5	squared_error	4	10	None	0.9625	0.9628
	5000	0.1	100	1.0	friedman_mse	2	1	None	0.9789	0.9793
500	100	0.1	400	1.0	friedman_mse	2	5	log2	0.9	0.8958
	1000	1	400	1.0	squared_error	3	5	None	0.971	0.9711
	5000	0.1	100	1.0	friedman_mse	2	1	None	0.9815	0.9817
1000	100	1	100	1.0	friedman_mse	2	1	sqrt	0.945	0.946
	1000	0.1	400	1.0	friedman_mse	2	5	sqrt	0.9955	0.9955
	5000	0.1	400	1.0	friedman_mse	2	1	sqrt	0.9983	0.99830
1500	100	0.1	100	1.0	friedman_mse	2	1	sqrt	1.0	1.0
	1000	0.1	100	1.0	friedman_mse	2	1	sqrt	0.999	0.9989
	5000	1	100	1.0	friedman_mse	2	1	sqrt	1.0	1.0
1800	100	0.1	100	1.0	friedman_mse	2	1	sqrt	1.0	1.0
	1000	0.1	100	1.0	friedman_mse	2	1	sqrt	1.0	1.0
	5000	0.1	100	1.0	friedman_mse	2	1	sqrt	1.0	1.0

2 Evaluation of models

2.1 Which classifier (among the four) yields the best overall generalization accuracy/F1 score? Based on your ML knowledge, why do you think the "classifier" achieved the highest overall accuracy/F1 score?

On the basis of the results, we can confidently say that the `GradientBoostingClassifier` yields the best overall generalization and accuracy. This is because `GradientBoostingClassifier` brings the best of all tree/ensemble methods. It uses the "wisdom-of-the-crowd" to reduce variance, and unlike `BaggingClassifier` and `RandomForestClassifier`, each successive tree is added only when it improves the performance of the overall classifier.

2.2 What is the impact of increasing the amount of training data on the accuracy/F1 scores of each of the four classifiers?

Increasing the training data increases the accuracy of the all the classifiers. This is because more examples allow the classifiers to "study" the data more effectively, thereby better understanding the relations between the features.

2.3 What is the impact of increasing the number of features on the accuracy/F1 scores of each of the four classifiers?

Similary, increasing the number of features also increases the accuracy across all classifiers.

3 Evaluation of various tree/ensemble models on the MNIST dataset

3.1 Which classifier among the four yields the best classification accuracy on the MNIST dataset and why?

Similar to the results that we get for the randomly-sampled dataset, we see that the accuracy of the classifiers is this order (lowest to highest): `DecisionTreeClassifier` (87.97%), `BaggingClassifier` (94.22%), `GradientBoostingClassifier`(94.59%), `RandomForestClassifier`(96.94%).

3.2 Compare the classification accuracy of tree and ensemble based classifiers with the (best) accuracy you obtained using the MLPClassifier, SVMs and nearest-neighbors in Project 2 (best as in after tuning the hyperparameters). Which classifier (or classifiers) among the seven has (have) the highest accuracy on the test set and why?

We see that the `MLPClassifier` wins among all the classifiers that we have tested so far with an accuracy of **97.5%**. This is because neural networks have more parameters than tree-based methods, which gives them more flexibility for capturing complex relationships between features.