2023-Apr-27

BGU

Computational Learning

Assignment 1

Roei Zaady 318747946

Omer Yanai 024093866

This report presents a summary of the results of building a Decision Tree Classifier, and a Bagging Classifier in Python.

A Decision Tree Classifier, is a model which attempts to classify data into groups based on previous observations. It works by building a tree-like model of conditions that will divide the data into groups which have shared attributes. Each node in the tree represents a decision, based on a feature in the input data, and each split represents the outcome of that decision. At the bottom of the tree there are leaves, which represent the final predicted classification. The Decision Tree Classifier is easy to interpret and visualize, making it a popular algorithm in machine learning. In this assignment, all features are assumed to be binary.

A Bagging Classifier (Bootstrap Aggregating), is a model that uses a group of Decision Tree Classifiers to reach a prediction of classification. The Bagging Classifier works by training multiple base classifiers, such as Decision Trees, on different subsets of the training data, and then combining the predictions of these classifiers to make a final prediction. By using an ensemble of Decision Trees in which each tree saw only parts of the data, the Bagging Classifier can help reduce overfitting and improve the generalization of the model.

Our work compares the performance of two bagging classifiers in a binary classification task (target classes are 0/1), on 5 different datasets, using 5 different metrics (accuracy, precision, recall, F1, AUC-ROC).

The Classifiers:

1. MyID3: MyBaggingID3: Our Bagging ensemble classifier using “MyID3” - Our Decision Tree developed from scratch.
2. SKLearn BaggingClassifier using SKLearn’s Decision Tree Classifier as base estimators.

The Datasets (source and shape) :

1. Tic-Tac-Toe: UCI (958,27)
2. Breast Cancer Wisconsin: UCI (699, 90)
3. Mushroom: Kaggle (8124, 117)
4. Haberman: UCI (306, 94)
5. Monk: UCI (432, 17)

Columns which contain no classification value (e.g. unique identifier per row), or columns with wide numeric range and no significant value to the predictions were removed. All other columns were preprocessed using Pandas’ get_dummies to “One Hot Encode” columns of binary values.

In order to evaluate the generalization capability of the different classifiers, we ran repeated K-Fold cross validation, subdividing the dataset into several folds, averaging the results from the different iterations in order to get a more robust measurement of performance by reducing both variance and bias caused by random data sampling.

Our assumption prior to running the evaluation on the different models, was that both bagging classifiers will perform about the same in all metrics, but that SKlearn's Bagging Classifier will be faster, as it is better adapted for a more efficient data handling.

During the evaluations of the classifiers and experimentation with the various datasets, we observed that two of the datasets are yielding perfect classification scores when allowed to use unlimited tree depth, therefore a limit (e.g max_depth=3) was introduced. By making the classification task more challenging, the models' predictions accuracy was impaired and would therefore allow us to better compare between the different classifiers.

Evaluation:

We ran 60 different experiments in Google Colab, using different configurations (defined in an external CSV file "runs.csv"). During the evaluation, we collected the results using the "Weights & Biases" platform. Link to table.

Name (10 visualized)	classifier	dataset	depth	n_estimato	max_featur	max_sample	Runtime	fit_time	test_accura	test_auc-ro	test_f1_sco	test_precision	test_recall
tic-tac-toe_MyBaggingID3	MyBaggingID3	tic-tac-toe	3	7	1.0	1.0	7s	0.4443	0.7787	0.7007	0.8493	0.765	0.9565
tic-tac-toe_BaggingClassifier	BaggingClassifier	tic-tac-toe	3	7	1.0	1.0	2m 42s	0.13	0.7631	0.6858	0.8381	0.7577	0.9395
mushroom_MyBaggingID3	MyBaggingID3	mushroom	None	7	1.0	1.0	6s	10.045	0.9997	0.9997	0.9997	1	0.9994
mushroom_BaggingClassifier	BaggingClassifier	mushroom	None	7	1.0	1.0	8s	0.1015	0.9997	0.9997	0.9997	1	0.9994
monk_MyBaggingID3	MyBaggingID3	monk	8	7	1.0	1.0	4s	0.5181	0.8137	0.7755	0.6942	0.7426	0.6599
monk_BaggingClassifier	BaggingClassifier	monk	8	7	1.0	1.0	18m 1s	0.01615	0.8345	0.8129	0.7448	0.758	0.7425
haberman_MyBaggingID3	MyBaggingID3	haberman	3	100	0.5	0.5	7s	1.059	1	1	1	1	1
haberman_BaggingClassifier	BaggingClassifier	haberman	3	100	0.5	0.5	11s	0.2871	1	1	1	1	1
breast-cancer-wisconsin_MyBaggingID3	MyBaggingID3	breast-cancer-wiscons	3	7	0.5	0.5	4s	0.2568	0.9485	0.9499	0.9255	0.9051	0.95
breast-cancer-wisconsin_BaggingClassifier	BaggingClassifier	breast-cancer-wiscons	3	7	0.5	0.5	21s	0.01528	0.9435	0.9434	0.918	0.9017	0.9378

The table below shows a partial snapshot of the collected results. Each run is showcased with the means of each metric, as well as the hyperparameters used in the models.

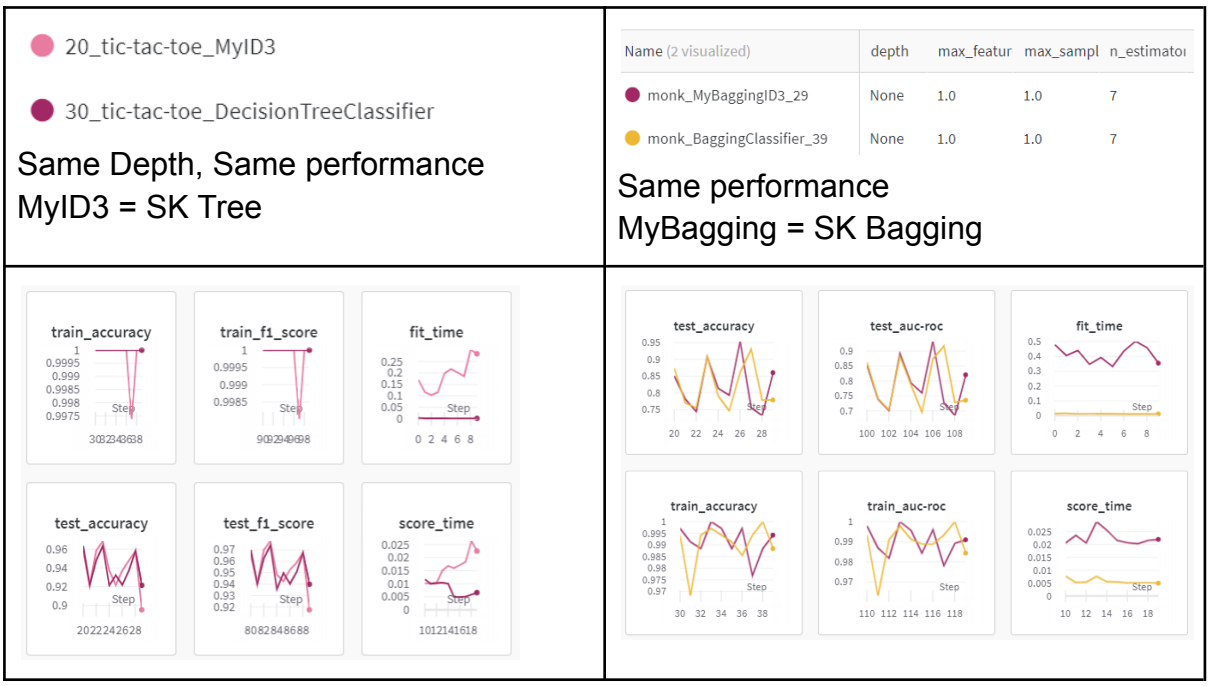
As can be seen, our "MyBaggingID3" Classifier performs similarly to SKlearn's "BaggingClassifier" in terms of Accuracy, Precision, Recall, F1 Score and AUC-ROC. As expected, SKlearn's Bagging Classifier is better adapted for speed and efficiency.

The full table of configurations and results can be seen in Appendix A.

Additional Findings from running all 60 configurations:

1. Comparison of MyBaggingID3 to SKLearn :

When applying the same hyper parameters on both models - we got similar performance on all datasets and all metrics, except for fit time in which SKlearn is faster.



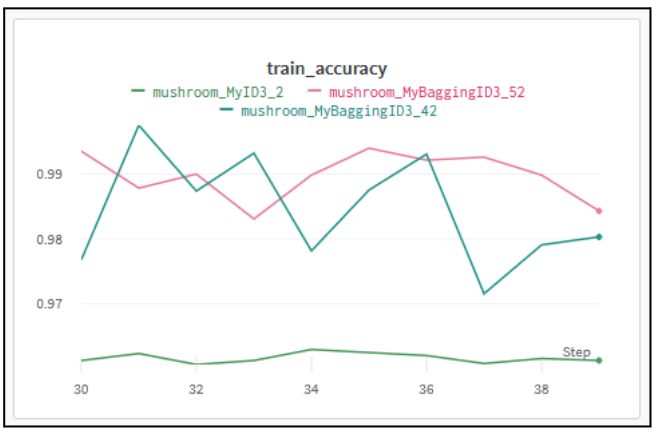
Example run numbers: run 20 vs run 30 ,run 29 vs run 39.

2. Impact of using a Bagging classifier with a subset of features/samples:

When setting max_samples/max_features to 0.5 in the ensembles, running many trees can still get better results than single trees, as variability is reduced, and the model becomes more resilient to over-fitting.

Name (3 visualized)	classifier	dataset	depth	max_featur	max_sampl	n_estimato	Runtime	fit_time	score_time	test_accura	test_auc-ro	test_f1_sco	test_precisi	test_recall
mushroom_MyID3_2	MyID3	mushroon	3	1.0	1.0	1	6s	0.04122	0.02705	0.9631	0.9649	0.9625	0.9276	1
mushroom_MyBaggingID3_52	MyBaggingID3	mushroon	3	0.5	0.5	100	20s	24.162	4.658	0.9889	0.989	0.9883	0.987	0.9896
mushroom_MyBaggingID3_42	MyBaggingID3	mushroon	3	0.5	0.5	7	6s	1.5	0.363	0.9803	0.9802	0.9792	0.9792	0.9792

We can see that run number 52 (with n_estimators=100 trees) outperforms both run number 42 (with 7 trees) and run number 2 (with a single tree and 100% of the features).



Appendix A: configuration and results
Bagging Classifiers

Name (40 visualized)	classifier	dataset	depth	max_featur	max_sampl	n_estimato	fit_time	score_time	test_accura	test_auc-ro	test_f1_sco	test_precisi	test_recall
tic-tac-toe_MyB...	MyBaggingID3	tic-tac-toe	3	0.5	0.5	100	1.314	0.3667	0.7749	0.6871	0.8502	0.7439	0.9919
tic-tac-toe_MyB...	MyBaggingID3	tic-tac-toe	3	1.0	1.0	7	0.2722	0.04107	0.733	0.6579	0.8159	0.7338	0.9187
tic-tac-toe_MyB...	MyBaggingID3	tic-tac-toe	3	0.5	0.5	7	0.08643	0.03281	0.712	0.6022	0.8148	0.6954	0.9837
tic-tac-toe_MyB...	MyBaggingID3	tic-tac-toe	None	1.0	1.0	7	0.728	0.0475	0.9686	0.9657	0.9756	0.9756	0.9756
tic-tac-toe_Bag...	BaggingClassifier	tic-tac-toe	3	0.5	0.5	100	0.1856	0.0178	0.7592	0.6683	0.8403	0.7333	0.9837
tic-tac-toe_Bag...	BaggingClassifier	tic-tac-toe	3	0.5	0.5	7	0.01384	0.006305	0.6911	0.5695	0.8053	0.6778	0.9919
tic-tac-toe_Bag...	BaggingClassifier	tic-tac-toe	None	1.0	1.0	7	0.01858	0.00643	0.9581	0.9609	0.9669	0.9832	0.9512
tic-tac-toe_Bag...	BaggingClassifier	tic-tac-toe	3	1.0	1.0	7	0.02563	0.009497	0.7487	0.6668	0.8298	0.7358	0.9512

Name (40 visualized)	classifier	dataset	depth	max_featur	max_sampl	n_estimato	fit_time	score_time	test_accura	test_auc-ro	test_f1_sco	test_precisi	test_recall
mushroom_My...	MyBaggingID3	mushroon	3	1.0	1.0	7	6.133	1.001	0.9631	0.9649	0.9625	0.9276	1
mushroom_My...	MyBaggingID3	mushroon	3	0.5	0.5	100	24.162	4.658	0.9889	0.989	0.9883	0.987	0.9896
mushroom_My...	MyBaggingID3	mushroon	3	0.5	0.5	7	1.5	0.363	0.9803	0.9802	0.9792	0.9792	0.9792
mushroom_My...	MyBaggingID3	mushroon	None	1.0	1.0	7	7.687	0.568	1	1	1	1	1
mushroom_Bag...	BaggingClassifier	mushroon	3	0.5	0.5	100	0.4942	0.04582	0.9951	0.9948	0.9948	1	0.9896
mushroom_Bag...	BaggingClassifier	mushroon	3	0.5	0.5	7	0.05805	0.01627	0.9803	0.9804	0.9792	0.9767	0.9818
mushroom_Bag...	BaggingClassifier	mushroon	None	1.0	1.0	7	0.09467	0.01155	1	1	1	1	1
mushroom_Bag...	BaggingClassifier	mushroon	3	1.0	1.0	7	0.1227	0.01865	0.9575	0.9587	0.9563	0.9321	0.9818

Name (40 visualized)	classifier	dataset	depth	max_featur	max_sampl	n_estimato	fit_time	score_time	test_accura	test_auc-ro	test_f1_sco	test_precisi	test_recall
monk_MyBaggi...	MyBaggingID3	monk	8	1.0	1.0	7	0.7069	0.04469	0.8372	0.7953	0.7407	0.8696	0.6452
monk_MyBaggi...	MyBaggingID3	monk	8	0.5	0.5	100	2.498	0.3988	0.6395	0.5	0	0	0
monk_MyBaggi...	MyBaggingID3	monk	8	0.5	0.5	7	0.1386	0.02051	0.7209	0.6199	0.4	0.8889	0.2581
monk_MyBaggi...	MyBaggingID3	monk	None	1.0	1.0	7	0.617	0.02507	0.7907	0.766	0.7	0.7241	0.6774
monk_Bagging	BaggingClassifier	monk	8	0.5	0.5	100	-	-	-	-	-	-	-
monk_Bagging...	BaggingClassifier	monk	8	0.5	0.5	7	0.01427	0.006388	0.6512	0.5584	0.3182	0.5385	0.2258
monk_Bagging...	BaggingClassifier	monk	None	1.0	1.0	7	0.01876	0.01041	0.8256	0.8144	0.7619	0.75	0.7742
monk_Bagging...	BaggingClassifier	monk	8	1.0	1.0	7	0.0252	0.01169	0.8023	0.7821	0.7213	0.7333	0.7097

Name (40 visualized)	classifier	dataset	depth	max_featur	max_sampl	n_estimato	fit_time	score_time	test_accura	test_auc-ro	test_f1_sco	test_precisi	test_recall
haberman_MyB...	MyBaggingID3	haberman	3	1.0	1.0	7	0.2415	0.03998	1	1	1	1	1
haberman_MyB...	MyBaggingID3	haberman	3	0.5	0.5	100	0.8435	0.1087	1	1	1	1	1
haberman_MyB...	MyBaggingID3	haberman	3	0.5	0.5	7	0.06152	0.01418	1	1	1	1	1
haberman_MyB...	MyBaggingID3	haberman	None	1.0	1.0	7	0.1244	0.01904	1	1	1	1	1
haberman_Bag...	BaggingClassifier	haberman	3	0.5	0.5	100	0.177	0.01692	1	1	1	1	1
haberman_Bag...	BaggingClassifier	haberman	3	0.5	0.5	7	0.01318	0.00605	1	1	1	1	1
haberman_Bag...	BaggingClassifier	haberman	None	1.0	1.0	7	0.01308	0.007415	1	1	1	1	1
haberman_Bag...	BaggingClassifier	haberman	3	1.0	1.0	7	0.01265	0.006083	1	1	1	1	1

Name (8 visualized)	classifier	dataset	depth	max_featur	max_sampl	n_estimato	Runtime	fit_time	score_time	test_accura	test_auc-ro	test_f1_sco	test_precisi	test_recall
breast-cancer-w...	MyBaggingID3	breast-car	3	1.0	1.0	7	2m 13s	0.6191	0.0389	0.9568	0.9486	0.9434	0.9804	0.9091
breast-cancer-w...	MyBaggingID3	breast-car	3	0.5	0.5	100	13m 3s	2.515	0.2755	0.964	0.9577	0.9533	0.9808	0.9273
breast-cancer-w...	MyBaggingID3	breast-car	3	0.5	0.5	7	1m 0s	0.2106	0.02717	0.964	0.9577	0.9533	0.9808	0.9273
breast-cancer-w...	MyBaggingID3	breast-car	None	1.0	1.0	7	2m 27s	2.061	0.05316	0.9424	0.9335	0.9245	0.9608	0.8909
breast-cancer-w...	BaggingClassifier	breast-car	3	0.5	0.5	100	11s	0.1959	0.01855	0.9712	0.9668	0.963	0.9811	0.9455
breast-cancer-w...	BaggingClassifier	breast-car	3	0.5	0.5	7	5s	0.01425	0.006388	0.9784	0.9759	0.9725	0.9815	0.9636
breast-cancer-w...	BaggingClassifier	breast-car	None	1.0	1.0	7	6s	0.01929	0.00653	0.9496	0.9395	0.9333	0.98	0.8909
breast-cancer-w...	BaggingClassifier	breast-car	3	1.0	1.0	7	6s	0.01633	0.006266	0.9568	0.9486	0.9434	0.9804	0.9091

Results of Single Tree Classifiers

Name (20 visualized)	classifier	dataset	depth	max_featur	max_sampl	n_estimato	Runtime	fit_time	score_time	test_accura	test_auc-ro	test_f1_sco	test_precisi	test_recall
tic-tac-toe_Myl...	MyID3	tic-tac-toe	None	1.0	1.0	1	4s	0.003507	0.00534	0.9215	0.916	0.9388	0.9426	0.935
tic-tac-toe_Myl...	MyID3	tic-tac-toe	3	1.0	1.0	1	4s	0.004131	0.005311	0.733	0.6809	0.8061	0.7571	0.8618
tic-tac-toe_Deci...	DecisionTreeClassil	tic-tac-toe	None	1.0	1.0	1	4s	0.003653	0.005544	0.8639	0.8581	0.8926	0.9076	0.878
tic-tac-toe_Deci...	DecisionTreeClassil	tic-tac-toe	3	1.0	1.0	1	4s	0.002966	0.005239	0.712	0.6449	0.797	0.7297	0.878
mushroom_Myl...	MyID3	mushroon	None	1.0	1.0	1	4s	0.0128	0.007314	1	1	1	1	1
mushroom_Myl...	MyID3	mushroon	3	1.0	1.0	1	6s	0.04122	0.02705	0.9631	0.9649	0.9625	0.9276	1
mushroom_Dec...	DecisionTreeClassil	mushroon	None	1.0	1.0	1	5s	0.01571	0.01097	1	1	1	1	1
mushroom_Dec...	DecisionTreeClassil	mushroon	3	1.0	1.0	1	4s	0.0135	0.007994	0.9631	0.9649	0.9625	0.9276	1

Name (20 visualized)	classifier	dataset	depth	max_featur	max_sampl	n_estimato	Runtime	fit_time	score_time	test_accura	test_auc-ro	test_f1_sco	test_precisi	test_recall
monk_MyID3_4	MyID3	monk	8	1.0	1.0	1	11s	0.01217	0.01741	0.6279	0.6106	0.5152	0.4857	0.5484
monk_MyID3_24	MyID3	monk	None	1.0	1.0	1	16s	0.004718	0.007539	0.6047	0.5572	0.4138	0.4444	0.3871
monk_Decision...	DecisionTreeClassil	monk	None	1.0	1.0	1	4s	0.004878	0.008025	0.6977	0.6862	0.6061	0.5714	0.6452
monk_Decision...	DecisionTreeClassil	monk	8	1.0	1.0	1	5s	0.002955	0.004975	0.593	0.5411	0.386	0.4231	0.3548
haberman_Myl...	MyID3	haberman	3	1.0	1.0	1	12s	0.03089	0.02081	1	1	1	1	1
haberman_Myl...	MyID3	haberman	None	1.0	1.0	1	4s	0.002638	0.004996	1	1	1	1	1
haberman_Deci...	DecisionTreeClassil	haberman	None	1.0	1.0	1	5s	0.003242	0.006073	1	1	1	1	1
haberman_Deci...	DecisionTreeClassil	haberman	3	1.0	1.0	1	4s	0.004576	0.008205	1	1	1	1	1
breast-cancer-w...	MyID3	breast-car	None	1.0	1.0	1	6s	0.004023	0.005449	0.9424	0.9335	0.9245	0.9608	0.8909
breast-cancer-w...	MyID3	breast-car	3	1.0	1.0	1	6s	0.005579	0.00818	0.9496	0.9458	0.9358	0.9444	0.9273
breast-cancer-w...	DecisionTreeClassil	breast-car	None	1.0	1.0	1	4s	0.003649	0.005315	0.8993	0.8759	0.8571	0.9767	0.7636
breast-cancer-w...	DecisionTreeClassil	breast-car	3	1.0	1.0	1	6s	0.004919	0.007599	0.9424	0.9304	0.9231	0.9796	0.8727