CSC-40094: Assessment 1 –Performance Evaluation of Machine Learning Algorithms

Author:
Özgün Gizlenci

Contents

Introduction

This is a document to observe the evaluation performance of the machine learning algorithms such as the Decision Tree algorithm and Support Vector Machine algorithm over hyperparameters. This includes methods and techniques that have been used in machine learning algorithms.

Evaluation

In this assessment, the most known algorithms of machine learning have been used such as the Decision Tree and Support Vector Machine classifier algorithms. Decision Tree is a member of the supervised learning subcategory of machine learning and artificial intelligence. It is one of the modelling approaches in machine learning and to make a decision it uses a set of rules similar to humans make decisions. (Patel and Rana, 2014)

On the other hand, the Support Vector machine is again a supervised learning member and it procures classification, regression and outliers detection. This task mainly will be focused on the performance evaluation of those two algorithms. (Nguyen, 2019)

Many industries use tree-based methods and it successfully demonstrated the cases for the prediction. It is not necessary to do more data cleaning. The decision tree is known as the simplest method that can be used in classification and regression problems.(Nguyen, 2019)

Furthermore, in comparison of decision tree with the SVM classifier, it is observed that decision trees are superior to the support vector machine when it is used for categorical data and performs better with colinearity. In addition, decision trees use hyper-rectangles to solve the problems however, support vector machine classifiers utilize kernel tricks to solve non-linear problems.

In this assignment accuracy, precision & recall, ROC curve and mean square error are calculated for each algorithm and performance observed. Hyperparameters are used and for each iteration, those metrics are calculated again.

The decision tree is iterated with two different parameters first one includes max_leaf_nodes in the range of 2 to 500 along with min_samples_split. Another parameter consist of min_weight_fraction_leaf between 0 to 0.5 with addition a max_depth with [1,5,9] distinctly over first parameter.

Dataset is created from moons dataset with 10000 samples and 0.4 noise and divided into train and test sets and grid search with cross-validation is included. On the other hand, for the support vector machine classifier, the same dataset was included and divided into test and train sets. The generated dataset is iterated to prevent possible problems with the dataset. Thus, accuracy, precision, ROC curve, mean square error and confusion matrix are calculated for each algorithm with each iteration.

Decision Tree First Iteration Results

Accuracy	0.8695
Precision	0.821165
MSE	0.1305

Table 1- Accuracy, precision & recall and mean square error calculation results for the first iteration of the decision tree

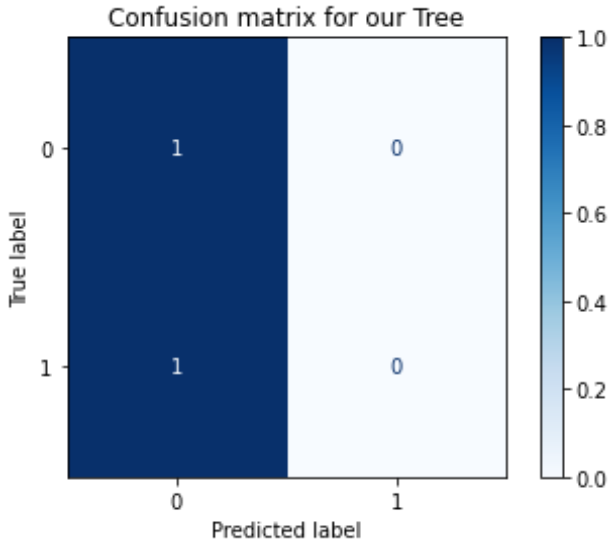

Figure 1 – Confusion matrix for the first iteration of the decision tree.

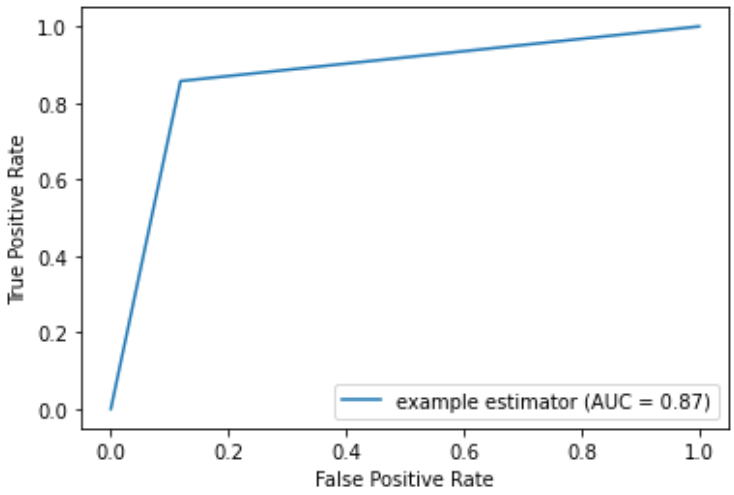

Figure 2 – ROC curve calculation for the first iteration of the decision tree.

Decision Second First Iteration Results

Accuracy	0.8605
Precision	0.813048
MSE	0.1395

Table 2- Accuracy, precision & recall and mean square error calculation results for the second iteration of the decision tree

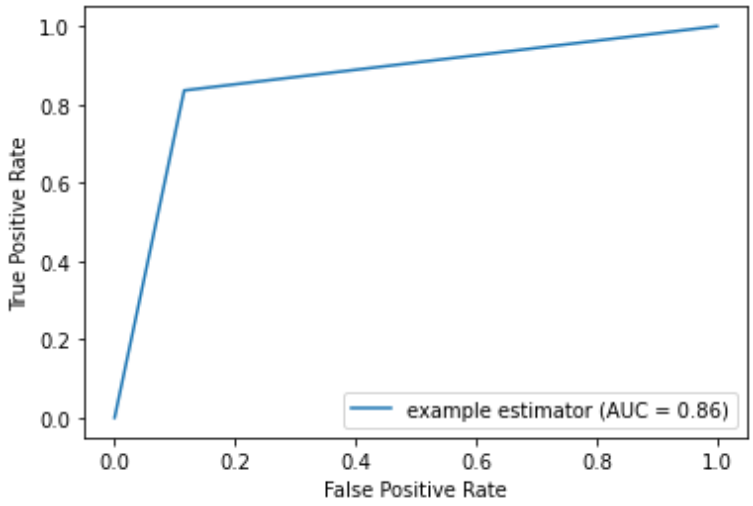

Figure 3 – ROC curve calculation for the second iteration of the decision tree.

Support Vector Machine Classifier Results

Accuracy	0.83
Acc. After Scaling	0.829375
Precision	0.814343
MSE	0.1325

Table 3- Accuracy, precision & recall and mean square error calculation results for SVM classifier

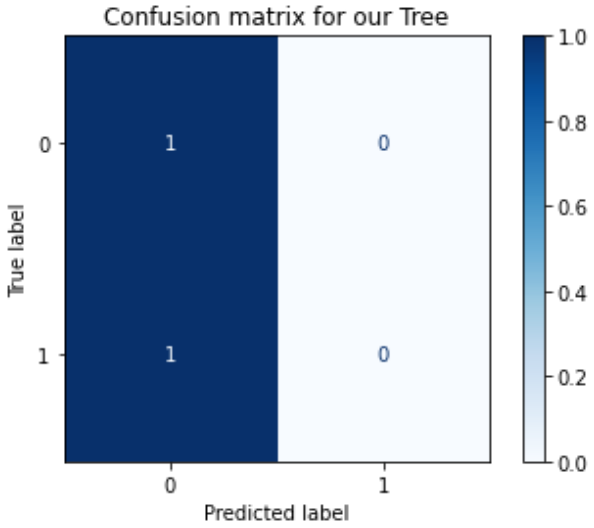

Figure 4 – Confusion matrix for the SVM classifier.

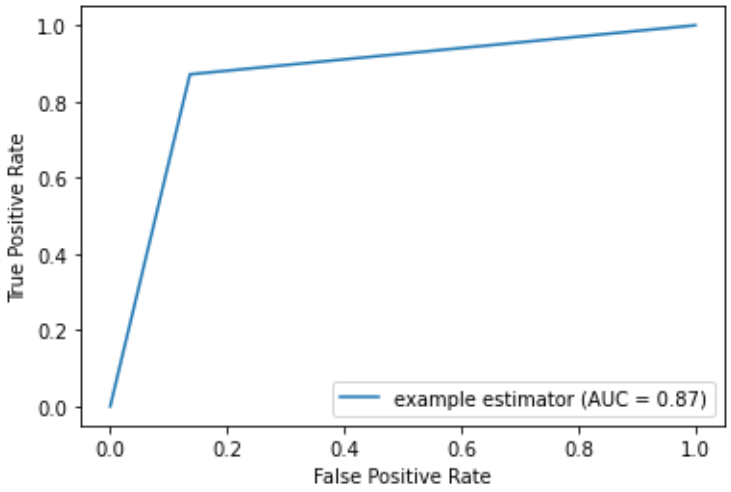

Figure 5 – ROC curve calculation for the SVM classifier.

As it is seen, the accuracy of the first and second iterations of the decision tree is almost the same. While accuracy is 0.8695 in the first iteration, it is decreased to 0.8605 by adding the min_weight_fraction_leaf parameter and mat_depth. On the other hand, precision is slightly decreased from 0.8211 to 0.8130. The mean square error is increased to 0.1395 in the second iteration from 0.1305.

	Decision Tree First Iteration	SVM Classifier
Accuracy	0.8695	0.83
Precision & Recall	0.821165	0.814343
Mean Square Error	0.1305	0.1325

Table 4- Accuracy, precision & recall and mean square error results comparison for two algorithms.

In other respects, if the decision tree and SVM classifier is compared, it can be observed result is almost identical with a small difference. Accuracy and precision results for the SVM are lower than the decision tree which means for the same model, the support machine vector is less accurate and precise. Besides, the mean square error comparison proves that SVM has a lower performance than the decision tree.

Conclusion

To sum up, a comparison of both algorithm's metrics results shows decision tree's performance evaluation proves it is more suitable for the classification than the support vector machine. Even if two algorithms produce close results, the decision tree is more accurate and precise, thus it gives high performance on the model.

References

Nguyen, H. H. (2019) 'A Complete View of Decision Trees and SVM in Machine Learning', *Medium*, pp. 1–10. Available at: https://towardsdatascience.com/a-complete-view-of-decision-trees-and-svm-in-machine-learning-f9f3d19a337b.

Patel, B. R. and Rana, K. K. (2014) 'A Survey on Decision Tree Algorithm For Classification', *Ijedr*, 2(1), pp. 1–5.