Decision Tree with Scikit-learn

LAB 2 - Introduction to Artificial Intelligence

19127216 - Đặng Hoàn Mỹ

Evaluate the Decision Tree in different training/ testing sets

19127216 | Lab 2

The classification report shows a representation of the main classification metrics on a per-class basis. It displays the precision, recall, F1 and support scores for the model.

Using this terminology the metrics are defined as follows:

- Precision can be seen as a measure of a classifier's exactness. For each class, it is defined as the ratio of true positives to the sum of true and false positives which means **"What percent of your predictions were correct?"**

- Recall is a measure of the classifier's completeness; the ability of a classifier to correctly find all positive instances. For each class, it is defined as the ratio of true positives to the sum of true positives and false negatives. Said another way, **"What percent of the positive cases did you catch?"**

- The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Generally speaking, F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

- Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

From the above definitions, we can make some conclusions when training the model on different data sets as follows:

Evaluate the Decision Tree in different training/ testing sets

19127216 | Lab 2

Interpret the classification report and the confusion matrix => make your comment about those decision tree classifiers.

In 4 different classification reports, we will receive the same precision, recall and f1-score in 1.00. Because there is no limit of max_depth in this section, the deeper the trees grow, the more complex the models will become. Without restrictions, the decision tree might simply overfit, this will cause the testing accuracy to decrease.

The confusion matrix also evaluates the performance of a classification model on a set of test data for which the true values are known. In this case, the target value has 2 classes p and e, so it seems like a binary classifier 0 and 1 after LabelEncoder fitting.

There are a small number of basic terms which represent for 4 cells in the confusion matrix,

- **true positives (TP)**: These are cases in which we predicted yes (they have the disease), and they do have the disease.
- **true negatives (TN)**: We predicted no, and they don't have the disease.
- **false positives (FP)**: We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- **false negatives (FN)**: We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

From these confusion matrices, we are able to compute a list of rates such as accuracy, precision (like in the classification report).

Through 4 different models, their performances are slightly similar. The number of testing objects predicted is half and half in two disparate classes 0 and 1. The data division is very uniform, there is no situation where there are too many elements of a class in the test sets.

Evaluate the Decision Tree in different maximum depths

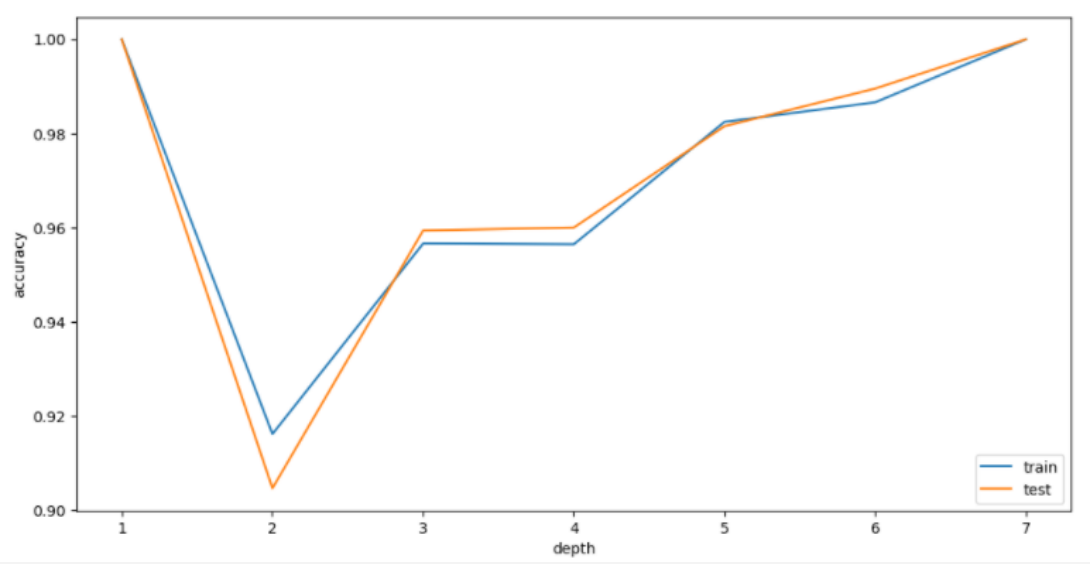
19127216 | Lab 2

The deeper the tree expands, the higher accuracy the model has. Nevertheless, 100% accuracy is not definitely the best model. The model does not categorize the data correctly if there are too many details and noise. A solution to avoid overfitting is using the parameters like the max_depth for stopping and some for pruning methods.

Therefore, the depth of the decision tree plays a vital role in the accuracy score. The ideal max_depth for this problem is the max_depth 3. In 0.2 test_size as the requirement of the assignment, the difference in max_depth 3 and 4 is slightly similar, that is the point of starting from depth 4 the model is overfitted. The accuracy_score is recorded and plotted to a line graph, shown at the bottom of the page. The 1 depth is the None max_depth.

The accuracy in the table below is rounded by 3 digits.

max_depth	**None**	**2**	**3**	**4**	**5**	**6**	**7**
Accuracy	**1.0**	**0.905**	**0.959**	**0.96**	**0.982**	**0.989**	**1.0**



The accuracy_score on the test set when changing the max_depth

References

- Classification Report — Yellowbrick v1.3.post1 documentation. (n.d.). Yellowbrick: Machine Learning Visualization. Retrieved August 27, 2021, from https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html
- D. (2020, June 14). Impact of decision tree depth on the accuracy. Kaggle. https://www.kaggle.com/dmytrol/impact-of-decision-tree-depth-on-the-accuracy
- Decision Tree Sklearn -Depth Of tree and accuracy. (2018, March 14). Stack Overflow. https://stackoverflow.com/questions/49289187/decision-tree-sklearn-depth-of-tree-and-accuracy/49289462
- How to interpret scikit's learn confusion matrix and classification report? (2015, June 10). Stack Overflow. https://stackoverflow.com/questions/30746460/how-to-interpret-scikits-learn-confusion-matrix-and-classification-report
- Kaggle: Your Home for Data Science. (n.d.). Kaggle. Retrieved August 27, 2021, from https://www.kaggle.com/prashant111/decision-tree-classifier-tutorials
- Kohli, S. (2021, June 1). Understanding a Classification Report For Your Machine Learning Model. Medium. https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397
- Markham, K. (2020, February 3). Simple guide to confusion matrix terminology. Data School. https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
- Mithrakumar, M. (2019, November 12). How to tune a Decision Tree? - Towards Data Science. Medium. https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680
- Teggi, P. (2020, February 15). Chapter 3 — Decision Tree Learning — Part 2 — Issues in decision tree learning. Medium. https://medium.com/@pralhad2481/chapter-3-decision-tree-learning-part-2-issues-in-decision-tree-learning-babdfdf15ec3