Decision Tree Learning Algorithm

0. About the script.

Some parts of the logic: recursion, numpy methods, shape of DT, prediction function are adopted from:

https://www.python-course.eu/Decision_Trees.php

Impurity calculations, splitting branches, probability count and other logic are based on the lectures.

1. Report the constructed decision tree classifier printed by your program. Compare the accuracy of your decision tree program to the baseline classifier

{'ASCITES': {True: {'SPIDERS': {True: {'VARICES': {True: {'STEROID': {True: ['live', 0.54], False: {'SPLEENPALPABLE': {True: {'FIRMLIVER': {True: ['live', 0.55], False: {'BIGLIVER': {True: {'SGOT': {True: ['live', 0.38], False: {'FEMALE': {True: ['live', 0.4], False: {'ANOREXIA': {True: ['die', 0.67], False: ['live', 0.33]}}}}}}}, False: ['live', 0.38]}}}}, False: {'ANOREXIA': {True: ['live', 0.67], False: ['die', 0.33]}}}}}, False: ['die', 0.01]}}, False: {'FIRMLIVER': {True: {'ANOREXIA': {True: {'SGOT': {True: ['live', 0.25], False: ['die', 0.75]}}, False: ['live', 0.33]}}, False: {'SGOT': {True: {'BIGLIVER': {True: ['live', 0.62], False: ['die', 0.38]}}, False: ['live', 0.6]}}}}}}, False: {'BIGLIVER': {True: {'VARICES': {True: {'FIRMLIVER': {True: {'STEROID': {True: ['die', 0.5], False: {'BILIRUBIN': {True: ['live', 0.67], False: ['die', 0.33]}}}}, False: ['live', 0.14]}}, False: ['die', 0.5]}}, False: ['live', 0.07]}}}}

Accuracy is: 76.0 %

2. Apply 10-fold cross-validation to evaluate the robustness of your algorithm

hepatitis-training-run-0 hepatitis-test-run-0

Accuracy is: 86.67 %

assign1-2.py hepatitis-training-run-2

Accuracy is: 66.67 %

assign1-2.py hepatitis-training-run-3 hepatitis-test-run-3

Accuracy is: 70.0 %

hepatitis-training-run-1 hepatitis-test-run-1

Accuracy is: 80.0 %

hepatitis-training-run-4 hepatitis-test-run-4

Accuracy is: 80.0 %

hepatitis-training-run-5 hepatitis-test-run-5

Accuracy is: 66.67 %

hepatitis-training-run-6 hepatitis-test-run-6

Accuracy is: 80.0 %

hepatitis-training-run-7 hepatitis-test-run-7

Accuracy is: 76.67 %

hepatitis-training-run-8 hepatitis-test-run-8

Accuracy is: 66.67 %

assign1-2.py hepatitis-training-run-9 hepatitis-test-run-9

Accuracy is: 73.33 %

Average accuracy is 74.6%

3. "Pruning" (removing) some of leaves of the decision tree

(a) how you could prune leaves from the decision tree

Possible methods:

1. Finding most irrelevant for the final decision attributes and disable it on particular node.

2. Set limits for DT.

Impurity criteria

Maximum amount of leaf nodes and minimum amounts of leaf nodes

Depth of the tree

Some random splitting

In the programme it could be done if remove attributes which are still did not became the nodes if the previously listed conditions are met

(b) why it reduces accuracy on the training set

The decision tree could be overfitting to training data set which can contain inaccuracy and it is being constructed consuming the noise from the data. During the training this noisy data also results in final decision. Removing some of the occurrences from DT will directly impact on performance rate. As there is going to be less correct decisions.

(c) why it might improve accuracy on the test set.

However, this will increase the performance on the tree on training data especially if there is a large amount of noise which could be tolerated with pruning. Also smaller size trees easier to understand.

4.Explain why the impurity measure (from lectures) is not an appropriate measure to use if there are three or more classes in the dataset.

In the node we always have split into two.

We have a formula to calculate impurity with the method given on the lectures.

$$P(A)\times P(B))$$

In this formula we multiply occurrences of each class. If one of the classes is not represented in the node the impurity of other class will be equal 0. Which is correct.

In case of three and more classes the formula will be.

$$P(A)\times P(B)xP(C)x\ldots$$

If one off the classes is be absent in the node. It will make us consider impurity of two (or more) others as 0. But it might not be true.

Therefor all the calculation will be incorrect with this method with three and more classes.

5. What three conditions must be met if a function (such as P(A)×P(B)) is used as an impurity measure in building a decision tree?

1. Maximum impurity only when all the probabilities are equal.

2. The minimum of the impurity function corresponds to the situation when probability of some class is 1 , while the probabilities of other classes 0.

3 .Swapping of probabilities does not change the value of the impurity function.

Source https://online.stat.psu.edu/stat508/lesson/11/11.2