MACHINE LEARNING PROJECT 1: OVERFITTING, UNDERFITTING AND METAPARAMETERS[1]

M-SECUC, INFOY-112: MACHINE LEARNING

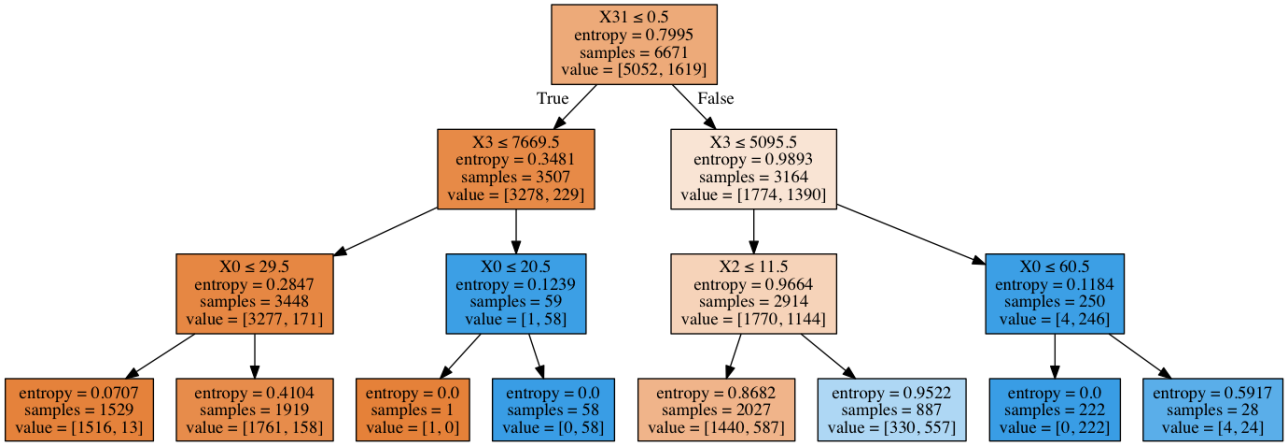
Muranovic Allan

[1] *Université de Namur*

1 TASK 1

1.Your first task is to train a DT with a maximal depth of 3. Show the DT in your report and comment on the result. **2**.What can you say about the entropy values and the class distributions at leaf nodes? **3**.Is the maximum depth of the decision tree large enough? **4**.Do you see over/underfitting cases by looking at the leaves? **5**.What is the training and test errors achieved by the DT?

1.1 Decision tree visualization



We do have an exact tree with a maximum depth of 3, with consistent values.

1.2 What can you say about the entropy values and the class distributions at leaf nodes?

We find 3 leaves with an entropy of 0, one relatively close to 0 (0.07) and the rest with values considered as a high level of disorder (on a scale from 0 to 1).We can conclude that there are several class types in our dataset.

1.3 Is the maximum depth of the decision tree large enough?

Thanks to the information obtained later on the prediction error greater than 16%, we can say that the depth of the tree is not important enough. This is felt through over- and under-fitting. Increasing the depth would reduce the significant error rate while reducing adjustments.

1.4 Do you see over/underfitting cases by looking at the leaves?

Yes :

- Over-fitting : low entropy and small sample (the leaf with entropy = 0 and samples = 1 for example)
- Under-fitting : High entropy with a big sample (the leaf with entropy = 0.9522 and samples = 887)

We notice that in the first case, we have a too specific leaf, and in the second case, we learn nothing from this high entropy with a large number of elements.

1.5 What is the training and test errors achieved by the DT?

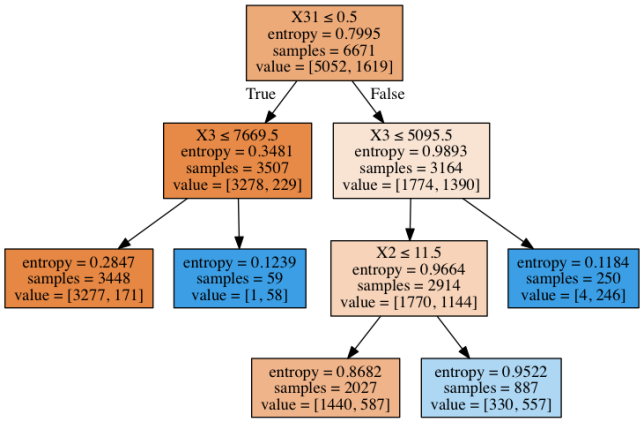
```
Training set evaluation:[[4718  334]
```

```
                        [ 758  861]]
             precision    recall  f1-score   support

      <=50K       0.86      0.93      0.90      5052
       >50K       0.72      0.53      0.61      1619

avg / total       0.83      0.84      0.83      6671
Training set error: 0.163693599161

Test set Evaluation:[[2005  136]
                     [ 330  388]]
             precision    recall  f1-score   support

      <=50K       0.86      0.94      0.90      2141
       >50K       0.74      0.54      0.62       718

avg / total       0.83      0.84      0.83      2859
Test set Error: 0.162994053865
```

2 TASK 2

1Your second task is to train a DT with another metaparameter than the maximal depth. Choose one metaparameter from the scikit-learn documentation, replace the maximal depth metaparameter and comment on the result. **2**How is this DT different with respect to the one from the first task? Compare the two DTs with respect to the over/underfitting at the leaves, the interpretability of the models (is one of the DTs easier/harder to understand?) and the training and test errors.

2.1 Your second task is to train a DT with another metaparameter than the maximal depth. Choose one metaparameter from the scikit-learn documentation, replace the maximal depth metaparameter and comment on the result.



Here we have a depth of 4, with the parameter of the number of maximum modified leaves. We arbitrarily chose a maximum number of leaves equal to 5 in order to obtain this result. We can see that the only node of depth 3 with leaves is the one with the largest entropy. This is naturally the best choice to try to correct the under-fitting problem of this node.

2.2 How is this DT different with respect to the one from the first task? Compare the two DTs with respect to the over/underfitting at the leaves, the interpretability of the models (is one of the DTs easier/harder to understand?) and the training and test errors.

Training and test errors :

```
Training set evaluation:[[4717  335]
 [ 758  861]]
             precision    recall  f1-score   support

      <=50K       0.86      0.93      0.90      5052
       >50K       0.72      0.53      0.61      1619

avg / total       0.83      0.84      0.83      6671
Training set error: 0.163843501724

Test set Evaluation:[[2005  136]
 [ 330  388]]
             precision    recall  f1-score   support

      <=50K       0.86      0.94      0.90      2141
       >50K       0.74      0.54      0.62       718

avg / total       0.83      0.84      0.83      2859
Test set Error: 0.162994053865
```

The first tree has leaves with generally lower entropy. So we have a second tree that is less accurate because each of these leaves has adjustment problems. The two are equally easy to understand, the second is just less precise. If we wanted more precision, we would have had to choose a depth of 4 or increase the maximum number of leaves.

3 TASK 3

1Your third task is to train DTs with a maximal depth of 1 to a maximal depth of 100 (meaning that 100 models are to be trained). Use the training set to train each DT and compute the training error (mean percentage of misclassification on training data). Use the test set to compute the test error (mean percentage of misclassification on test data). Show a plot of the training and test errors for each depth in your report and comment on the result. Does the plot correspond to the theory? Can you spot over/underfitting cases in the plot?

See :

`error_question_3.png`

We see the error rate on the training set decrease to a depth of about 30, once this point is reached, the error rate converges to 0. However, on the training set, the error rate increases towards depth 8, until it converges towards a value higher than that at the beginning of the ascent. It doesn't really make sense. The shaft of depth 8 is therefore more accurate than those above this value.

Concerning over/under-fitting, we can see one when a difference is created in the regression of the 2 curves. We therefore notice that there is an under-fitting up to a depth of about 5, once this depth level is exceeded it is a case of over-fitting that we observe.

4 TASK 4

Your fourth task is to train DT with an increasing value for the second meta- parameter you chose in Section 3. Use the training set to train each DT and compute the training error (mean percentage of misclassification on training data). Use the test set to compute the test error (mean percentage of misclassification on test data). Show a plot of the training and test errors in your report and comment on the result. Does the plot correspond to the theory? Can you spot over/underfitting cases in the plot? **1**What is the difference with the maximal depth plot?

`error_question_4.png`

Here, we notice directly that our tree is more sensible than the previous one. We have an under-fitting up to a tree of value 5. Once depth 5 is exceeded, the errors no longer follow the same regression, confirming over-fitting. The beginning of the graph is similar to that of task 3, this means underfitting to a depth of about 5, again, once the heading has passed the errors no longer follow the same regression, indicating a case of overfitting

4.1 What is the difference with the maximal depth plot?

We directly notice that the error rate of the test set based on a tree with a maximum depth is less stable than that based on a tree with a maximum number of leaves. But the 2 performances are globally equal, provided that we have chosen a good depth for tree with a maximum depth.

5 TASK 5

1Finally, choose the best maximal depth value (according to you) based on the corresponding plot of the third task. Justify your choice theoretically. Comment the tree very shortly (in terms of over/underfitting, interpretability and training/test errors). **2**Repeat the same procedure (choosing the best value based on the plot and analyzing the resulting tree) for the other metaparameter you chose in Section 3. Compare these two new DTs. How are they different with respect to the ones of Section 3?

5.1 Maximal depth value

Target files :

`dt_max_depth_of_5.png`
`evaluation_max_depth_of_5.txt`

Despite high entropies in some places, some sheets have an entropy equal to or close to 0, and the error ratio in the.txt file indicates lower error rates for our depth model 5. We have a little underfitting, leaf number 11 from the left, and also overfitting, leaf number 4.

5.2 Maximal leaves value

Target files :

`dt_max_leaf_node_of_33.png`
`evaluation_max_leaf_nodes_of_33.txt`

After a test period, we conclude on a maximum of 33 leaves in order to obtain the lowest error rate(the smallest obtained so far). It is in this model that we have the most leaves with the smallest entropy (close to 0), but we always face underfitting (leaf 12 from the left) and overfitting (leaf 6 and 7 for example).