Exercise on Decision Tree

We are going to use a dataset from the Sloan Digital Sky Survey (SDSS) which contains several images of the sky collected with a wide-angle optical telescope in New Mexico, United States. The final data release covers over 35% of the sky and it is publicly available. More information can be found on Wikipedia: `https://en.wikipedia.org/wiki/Sloan_Digital_Sky_Survey`. We are going to use a small excerpt of that dataset containing data related to approximately 10000 objects in the sky. We are going to focus on the task of classifying sky objects as stars, galaxies or quasars. In our dataset, the class value 0 corresponds to star, 1 corresponds to galaxy and 2 to quasar. You should write a report (max 3-5 pages) which contains the answers to your questions and your findings. You should submit both the code and the report on the moodle website until **October 19th** at 5pm.

Plagiarism Discussion with the students and teachers is encouraged, however, you should write your own code. If we believe that a student has copied his/her code from another student or from the internet (e.g. github or other repositories), that student will fail the exam!

1. Given the dataset we provided to you, build a decision tree using the default input parameters. Include the decision tree you built in the report.

2. compute the generalization error of the decision tree you built. To this end, you might use the array $clf.tree\_.children\_left$ where $clf.tree\_.children\_left[i] = -1$ if i is a leaf while clf is the tree you built with $DecisionTreeClassifier$ in sci-kit learn.

3. The decision tree you built in the first part of the question might not be ideal for our task. You should try to change the input parameters of $DecisionTreeClassifier$, so as to build a decision tree with *minimum generalization error*. Specify in your report which parameters you considered and how you expect that a given parameter will affect the generalization error. It should be clear from your answer that you understood what is the role of each parameter and how it might affect the generalization error. Include the decision tree you built in the report.

4. Compare the decision trees you built in point 1 and the best one you obtained in point 2. Which one would you recommend to use to classify sky objects? Motivate your answer.

5. Consider the decision tree you considered to be best in the previous point. Predict the class value of an object of your choice. Which insights you obtained by looking at the decision tree?

6. Do you think that the best decision tree you built could be pruned so as to improve the generalization error? Motivate your answer (you are supposed to answer this question by only looking at the tree, no implementation is required).

7. The library we recommend (sci-kit learn) does not support post-pruning, yet. However this could be implemented by using the variables of the $tree\_$ object computed by the DecisionTreeClassifier in sci-kit learn. See [1] to see some examples. In particular, $clf.tree\_.children\_left[i]$ specifies the index of the left children of i, $clf.tree\_.children\_right[i]$ specifies the index of the right children of i, while $clf.tree\_.value[i]$ specifies the class distribution of i. Implement a post-pruning strategy (among the ones we considered in our course) and run it on the best decision tree so far. Does this improve the generalization error?

8. (Bonus question, it will give extra points). In case your implementation of the post-pruning strategy is very good and efficient we might check whether it can be integrated in sci-kit learn. In case you are interested, please contact the teacher to determine which post-pruning strategy could be implemented. Some extra time to solve this exercise will be granted.

[1] `http://scikit-learn.org/stable/auto_examples/tree/plot_unveil_tree_structure.html#sphx-glr-auto-examples-tree-plot-unveil-tree-structure-py`