Programming Assignment 2

Write a Python3 program that implements a decision tree using the ID3 algorithm presented in the lecture. Use the following entropy calculation:

$$entropy(S) = -\sum_{i=1}^{C} p_i log_C(p_i)$$

where p_i is the proportion of class i (with C being all classes in the data set). Use Information Gain as your decision measure and treat all features as discrete multinomial distributions.
Given are the two data sets[1] named *car* and *nursery* as csv files. Your program should be able to read both data sets and treat the last value of each line as the class. Your task is to correctly implement the ID3 algorithm and return the final tree without stopping early (both data sets can be learned perfectly, i.e. all leaves have an entropy of 0). The output of your algorithm should look like the example XML solution given for the *car* data set. With that, you can check the correctness of your solution. All features are unnamed on purpose, please number them according to the column starting from 0 (e.g. att0).
For each data set, you can acquire one point, if the solution of your program returns correct results. If the program fails, the data format is incorrect or I have to change source code, in order to make it work, you will get zero points. Machine learning libraries are not allowed. You can use libraries for handling the XML format and the input parameters.
Your program must accept the following parameters:

1. **data** - The location of the data file (e.g. /media/data/car.csv).
2. **output** - Where to write the XML solution to (e.g. /media/data/car_solution.xml).

Please prepare example statements on how to use your program. E.g. for a python program:

```
python3 decisiontree.py --data car.csv --output car.xml
```

The final program code must be sent via email until Sunday, 24th of November 2019, 23:59 to your respective tutor. Please format your e-mail header as follows:

[Exercise Group] ML Programming Assignment 2

Replace *Exercise Group* with the day and time of your exercise group. E.g for Monday from 13:00 to 15:00 it would be:

[Monday 13-15] ML Programming Assignment 2

Do not forget to include your names and matriculation numbers in the mail! Please also be prepared to present your solution shortly in front of the class. You will get one point for each data set, if the output of your ID3 algorithm is correct.

2 points

[1] http://wwwiti.cs.uni-magdeburg.de/iti_dke/Lehre/Materialien/WS2019_2020/ML/res/decisiontree.zip