

Intro: Artificial Intelligence (CS331)

Assignment - 3

Instructions

- Due Date: 24th March, 2021 at 11:55PM
- For the written part, you need to submit a scanned .pdf file of your **handwritten** solution named <RollNumber>.pdf
- The dataset given in the coding part is only for your testing purposes, your code should be general and should work on any dataset (with same formatting and a yes/no output)
- Your submission should be a **.zip** file (not .rar or any other format) containing the scanned pdf for part 1, python code, encodings.txt and dataset.txt for part 2.
- You **must** name your python file as <RollNumber>.py before submitting the assignment. For example, if your roll number is 22100279, you will name your python file as 22100279.py.
- Start early. This assignment should not take you more than a day to complete (max two), please do not leave it for the last minute. You can, however, use your remaining grace days if you fail to complete it before the deadline. Submissions ahead of the deadline by even a second will mean your submission is late.
- If all grace days are used and you still submit late, then the assignment will **NOT** be accepted.
- **Any plagiarism case will be strictly dealt with according to the course policy provided on LMS without any further warnings.**

Written Part [25]

The given table displays different conditions on the basis of which Zafir decides whether he will go to a restaurant for dinner or not.

Perform the ID3 algorithm to generate the decision tree which corresponds to the table.

Budget	Hungry?	Restaurant Reviews	Goes to restaurant
low	no	bad	no
high	yes	bad	no
medium	no	good	no
low	no	good	no
low	yes	bad	yes
high	yes	good	yes
medium	yes	bad	yes
medium	no	bad	no
high	no	good	yes

Coding Part [75]

In this part, you will be implementing the ID3 Algorithm in python.

We have provided you with a dataset.txt file that you can work on to test your code.

Each row's values in the dataset corresponds to the following:

outlook,temperature,humidity,wind,playTennis

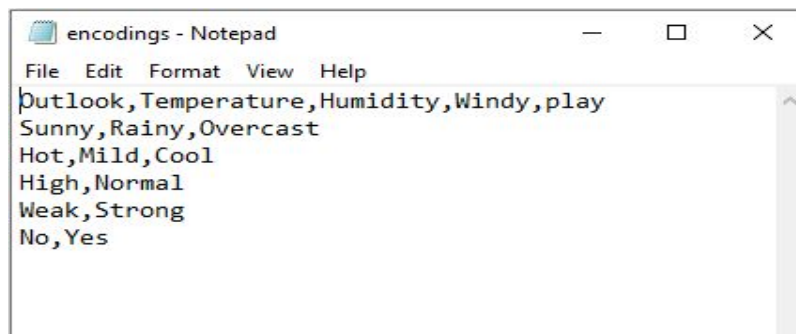
The values of each attribute are as follows:

Outlook	Temperature	Humidity	Wind	Plays Tennis
Sunny: 0	Hot: 0	High: 0	Weak: 0	No: 0
Rainy: 1	Mild: 1	Normal: 1	Strong: 1	Yes: 1
Overcast: 2	Cool: 2			

Hence, if the row is: 0,0,0,0,0

That would mean that the outlook was sunny, temperature was hot, humidity was high, wind was weak, and the person did not play tennis.

You have also been provided encodings.txt. The dataset.txt contains integer labels for the string attribute values. And the encodings.txt contains information about which integer maps to which string label (this will be useful for naming the keys of your output tree).



You should interpret this file as follows:

- The first row of encodings.txt gives you the name of columns in the dataset.txt. For the first column name is outlook, 2nd one is Temperature, third is Humidity and so on.
- Then the second row gives you label mappings for attributes inside the first column e.g Outlook. The sunny maps to 0, Rainy maps to 1 and Overcast maps to 2.
- The third row gives label mappings for attributes inside 2nd column e.g Temperature. Hot maps to 0, Mild maps to 1, Cool maps to 2
- Onward rows follow the same pattern. In short each nth row except 1st one gives you labels for (n-1)th column. And each word inside that row maps to i where i is the index of that word in that row and i starts from 0.

You need to make a decision tree based on the values given in dataset.txt. You should use the ID3 algorithm to make the tree. You need to print your tree as a python dictionary. While

printing the tree you will need to convert integer labels back to their original string labels using the encodings given in the table above. Your output will look similar to this:

```
D:\>tree.py dataset.txt encodings.txt
{'Outlook': {'Overcast': 'Yes',
             'Rainy': {'Windy': {'Strong': 'No', 'Weak': 'Yes'}},
             'Sunny': {'Humidity': {'High': 'No', 'Normal': 'Yes'}}}}
```

Tips

You will need to use the numpy library for this assignment. Numpy is very similar to pytorch which you used in your assignment 2. You will find the following numpy functions helpful for this assignment (this list might be non-exhaustive depending on your implementation).

- `numpy.loadtxt()`
- `numpy.where()`
- `numpy.log2()`
- `numpy.count_nonzero()`

We have pretty-printed the dictionary using the `pprint.pprint()` function from the builtin python pprint library. You should also do this otherwise your dictionary will be printed in a single line and will be difficult to read.

Since your program should be general enough to work on any given dataset of the format described above with a yes/no output, you should consider making datasets yourself to test your code further. It will only take a couple of minutes to make the dataset (the one we have provided is from the example given in class).

How to run your code

Your data set should be present in the same directory as the python file, and your program should take as input the data set file from the command line. The command to run the program will be as follows:

```
python3 <RollNumber>.py <datasetName>.txt <encodings>.txt
```

(look up `sys.argv` on the internet to see how you can get the file name from the command line into your program)