CS-331 Introduction to Artificial Intelligence - *Spring 2023*

Programming Assignment 3

Made by: Muhammad Musa (23100004@lums.edu.pk)

Instructions:

- This project is on an group basis.
- Due date: **28 April 2023**. No submissions will be accepted after this date.
- You are required to submit **both** a .ipynb, and a .py file; both of them should be named: GroupNumber_PA3.
- Total Marks are 100. A breakdown of these marks is given with each part.
- **NO PLAGIARISM WILL BE TOLERATED** In case of finding any plagiarism, the case will be referred to the DC committee, and the student will be given a 0 in **all the projects**. Plagiarism is presenting someone else's work or ideas as your own, with or without their consent, by incorporating it into your work without full acknowledgment. All published and unpublished material, whether in manuscript, printed, or electronic form, is covered under this definition, along with ChatGPT. Plagiarism may be intentional or reckless, or unintentional. Under the regulations for examinations, intentional or reckless plagiarism is a disciplinary offense. If you are unsure about what counts as plagiarism and what doesn't, **PLEASE reach out** to the TAs so that they can guide you.

Part One [Decision Trees]:

In this part we are going to work on different kinds of regressions and work on their implementation. In this part, you are to implement a decision tree using the ID3 algorithm. You have already done a similar thing on paper in your homework. In this part, you have to convert that into code.

Dataset [10 marks]

The dataset we will use in this part is car.txt. This dataset is a car evaluation dataset that evaluates each car according to four class values:

(a) unacc (unacceptable)

(b) acc (acceptable)

(c) good (good)

(d) vgood (very good)

In the .txt file, each row represents one instance, and a comma separates each feature in each row. The feature names are not included, but their details are as follows:

- buying is the first value in each row. It can have the values:
  - vhigh
  - high
  - med
  - low
- maint is the second value in each row. It can have the values:
  - vhigh
  - high
  - med
  - low
- doors is the third value in each row. It can have the values:
  - 2
  - 3
  - 4
  - 5more

- persons is the fourth value in each row. It can have the values:
  - 2
  - 4
  - more
- lug boot is the fifth value in each row. It can have the values:
  - small
  - med
  - big
- safety is the sixth value in each row. It can have the values:
  - low
  - med
  - high
- car eval is the seventh and last value in each row, and its values have already been mentioned above.

There are some changes that you need to make to this dataset that will make it easier to implement the ID3 algorithm:

(a) As mentioned above, this dataset has four class values. You are to reduce these to two by using two classes: "unacc" and "good". All the instances where the class is "unacc" will remain "unacc", but in all the instances where the class is anything other than "unacc" ("acc", "good", or "vgood"), all of these are to be changed to "good" and come under the same class.

(b) The dataset at the moment is quite big. You will cut it down to roughly 500 instances in the training dataset and 100 instances in the test dataset.

(c) Lastly, we will change the categorical variables to numbers to make it easier to work with. The conversion would be:

- buying:
  - vhigh : 3
  - high: 2
  - med: 1
  - low: 0
- maint:
  - vhigh: 3
  - high: 2
  - med: 1
  - low: 0
- doors:
  - 2: 0
  - 3: 1
  - 4: 2

  -5more: 3
- persons:
  - 2: 0
  - 4: 1
  - more: 2
- lug boot:
  - small: 0
  - med: 1
  - big: 2
- safety:
  - low: 0
  - med: 1
  - high: 2
- car eval
  - unacc: 0
  - good: 1

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			

Figure 1: Example Dataset Information

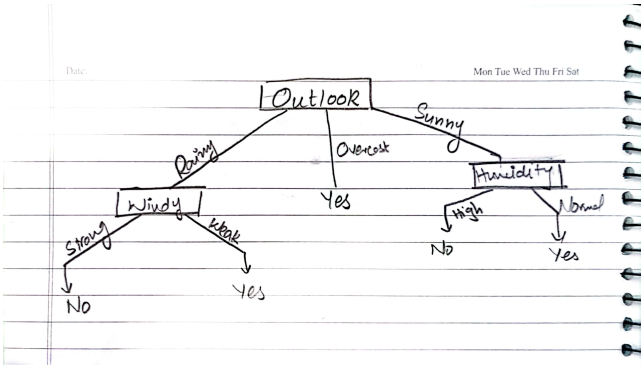

Figure 2: Example Decision Tree

Decision Tree [25 marks]

With this dataset, you are to implement the ID3 algorithm and use **information gain** and get the decision tree. You can represent it in any way you like. However, the easiest way would be to represent it as a dictionary of dictionaries. For example, for a dataset with the information shown in **Figure 1**, the decision tree that would be made would be as shown in **Figure 2**. To represent this in a dictionary of dictionaries, it would look as shown in **Figure 3**.
Note: To get your dictionary to match the representation style in Figure 3, you can and should use the inbuilt python library: ***pprint***.

After you are done with the decision tree, use it for the test dataset and report the accuracy in the notebook.

Part Two [Multi Layer Perceptron]:

In this part, you are simply going to use the sklern libraries to make a MLP classifier and run it on the provided dataset.

Dataset

The dataset for this is provided in the folder: "Dataset". This dataset contains images belonging to 6 different classes. Train and test data has already been split.

MLP [15 marks]

Your job is to code a network that trains on the train data provided and gives a good accuracy (50% at least). Skeleton code for preproocessing of the dataset and running the model has already been provided to you.

Bonus Part [10 marks]

If you are able to get an accuracy of above 55% in the first part, try to increase the the accuracy to parameters used ratio as much as you can in this part. You will be awarded marks out of 10 based on how much you improved from the first part and realtive to how other students performed.
Do keep in mind tho that this is a bonus part and not at ALL required to get full in this assignment.

```
{'Outlook': {'Overcast': 'Yes',
             'Rainy': {'Windy': {'Strong': 'No', 'Weak': 'Yes'}},
             'Sunny': {'Humidity': {'High': 'No', 'Normal': 'Yes'}}}}
```

Figure 3: Dictionary Representation

Part Three [Implementing NN from Scratch]:

This part is going to be tricky not going to lie, so make sure you give enough time to this and make sure your concepts surrounding neural networks, hidden and output layers, activation functions etc. are clear.

Dataset

You have been provided with a file called *part3.csv*, a simple dataset with two different classes. The dataset contains of 3 columns: x, y, and label, adn your goal for this part is going to be making a neural network **FROM SCRATCH WITHOUT USING ANY LIBRARIES** related to NNs.

VISUALISATION [7 MARKS]

Plot the dataset as a scatter plot. Each class's points should be shown in a different color. Ensure to include a legend, title, and x and y labels.

TRAINING DATA AND MAKING A NEURAL NETWORK [43 MARKS]

This is the open-ended part. Do a 70-30 split on your dataset and make a neural network from scratch. How you do this is up to you, you could either pre-process the data to get something linearly separable and run a simple neural network on that, or you can make a complex network using different layers and use it to train it on the current dataset.

BEST OF LUCK FOR YOUR ASSIGNMENT !!!