Coursework_O2

September 8, 2019

1 Coursework Output 2

Instructions: In this coursework, you will show your domain of the data related Python skills required for business analytics. To do so, you will use packages such as *pandas* and *matplotlib*.

1.1 Importing and Visualising a Business Case in Python

Each student will be assigned a different dataset in a different format. The goal is to import the dataset into Python so that you can wrangle and visualise the data in better ways. This will allow you to get your own conclusions and start building up knowledge regarding on how you could potentially learn from data to predict or classify future instances.

You must create a program which allows you to select the following options: 1. Import your dataset into Python as a Pandas data frame. 2. Query and print an instance of the dataset (by row number or by row name). 3. Create a "reduced" dataset (with less columns) by indicating a list of columns to bring upon this newly created dataset. 4. Randomly split the original or the reduced dataset into two substes called *training* and *testing* according to a ratio specified by the user. 5. Visualise a dataset by means of a scatterplot which relates two variables/columns specified by the user. The plot has to show the x and y axis labels and use the *target/class* column (i.e. the last one) as the colour variable. 6. Exit the program.

1.2 Additional Considerations

- The program has to check that every input option added by the user is valid.
- No option can be executed until option 1 is executed first.
- Whenever a dataset or subdataset is imported or created, print it for the user to visually inspect it.
- In option 2, the row to query can be specified either using the row number or the row name.
- In option 3, the list of columns has to be specified by column name.
- Option 5 can receive as input either the original dataset or the training/testing ones (if these have been already created).

1.3 Submission Instructions

- Once that you have finished your program, run all cells and run the main program cell using the sequence of options 0-2-1-2-3-4-5-6 (option 0 is purposely created to verify that your program can handle the error).
- Then, without clearing the kernel, generate a html **OR** pdf file from the Jupyter notebook.

- Name both the Jupyter notebook and the html/pdf file with your id number and submit them to the corresponding Moodle's dropbox before **12th December, 2019**.

In [ ]:
```
## Use this cell to import all necessary packages
```

In [ ]:
```
## Use this cell to define the function corresponding to OPTION 1
def option1():
    '''With this function you import the dataset.'''
    return name
```

In [ ]:
```
## Use this cell to define the function corresponding to OPTION 2

def option2():
    '''This function queries and prints an instance of the dataset.'''
    return
```

In [ ]:
```
## Use this cell to define the function corresponding to OPTION 3

def option3():
    '''This function creates a new dataset by indicating which columns to include.'''
    return
```

In [ ]:
```
## Use this cell to define the function corresponding to OPTION 4

def option4():
    '''This function randomly splits the dataset into train and test.'''
    return
```

In [ ]:
```
## Use this cell to define the function corresponding to OPTION 5

def option5():
    '''This function visualises the dataset using a scatterplot.'''
    return
```

In [ ]:
```
## Use this cell to create the "main" part of your program
print('Welcome to ***NAME AND ID*** business case.')
```

1.4 Questions

Please answer the following questions to appraise your level of engagement with the content of the course. Use the Markdown cell corresponding to each question to write your answers.

1. Using any of the two continuous variables of your dataset, show an example of how a linear regression (implemented using an existing Python module) could be applied on the training data to predict the values of one column of the test data. Discuss if there is any metric that can be used to decide which two variables are most correlated.

ANSWER:

In [ ]:
```
# Use this cell to implement linear regression.
```

2. Using any clustering method available in literature and in a Python module (e.g. hierarchical, k-means), briefly describe the selected method and implement it to classify the data of the original dataset into clusters. How would you verify how accurate is your clustering algorithm with respect to the original dataset target/class?

ANSWER:

```
In [ ]: # Use this cell to implement clustering.
```