04ex_OSEMN

March 14, 2020

1 OSEMN Exercises

```
In [1]: %matplotlib inline
        import pandas as pd
        import numpy as np
```

1. Create a random list of number and then save it to a text file named "simple_data.txt"

```
In [2]: N=10
        np.savetxt("simple_data.txt", np.random.rand(N))
```

2. Create a random matrix of 5x5 and then save it to a text file named "data.txt"

```
In [3]: N=3
        np.savetxt("data.txt", np.random.rand(N, N))
```

3. Load the saved txt file of point 2 and convert it to a csv file (by hand)

```
In [4]: import csv

        with open ("data.csv", "w") as outfile:
            writer = csv.writer(outfile,delimiter=',')
            data = np.loadtxt("data.txt")
            print(type(data))
            for row in data:
                print(row)
                writer.writerow(row)
```

```
<class 'numpy.ndarray'>
[0.00215232 0.58684945 0.06567951]
[0.66974597 0.51395297 0.53761241]
[0.51152316 0.339627   0.94137189]
```

4. load the binary file named *credit_card.dat* and convert the data into the real credit-card number. Each line correspond to a credit card number. Each character is composed by 6 bit (even the space) and the last 4 bit are just a padding

```
In [10]: import sys

         with open ("credit_card.dat", "rb") as infile:
             lines = infile.readlines()
             ctn = 0
             for line in lines:
                 if lines != line[-1]:
                     for i in range(19):
                         sys.stdout.write(chr(int(line[i*6:(i+1)*6],2)))
                 print("")
```

```
7648 5673 3775 2271
3257 8247 3354 2266
2722 0001 4011 6652
0661 3063 3742 3150
0432 1608 1462 4742
5827 2027 8785 7303
5774 8528 2087 1117
8140 1210 6352 2845
5764 1133 7301 7100
6456 1737 4126 6726
1228 8631 7382 0000
7051 0160 5374 3166
0618 3587 1630 6376
1545 5454 7444 5636
6735 3116 3202 6834
7287 5011 1547 8413
7033 2607 3328 4200
2568 5244 1874 5024
1684 2253 7570 7118
0672 2576 0575 6631
6332 8353 8787 1340
1813 3361 1175 4211
2477 6450 8840 2368
5512 3505 2563 1326
3083 7882 0621 0025
4521 5148 8045 0334
7563 3654 8713 5787
8324 2664 0476 5561
0565 2504 7168 3510
5107 5507 1767 0738
2462 1821 2448 1443
2788 0638 6861 6554
5851 5873 5474 0547
0670 1004 4013 2655
5874 5506 3048 0806
2805 5401 8462 1260
5083 8406 6310 1862
```

```
1076 1445 3013 2266
8440 4804 4844 5277
4758 6141 0686 1387
7586 0675 0315 2568
2544 1258 7432 5165
3474 5023 4434 5626
1410 0270 0434 5086
7315 4446 1104 4215
0224 7742 8300 0266
0170 2700 3145 0640
2006 2437 8054 1600
8142 4055 1776 0026
3026 7380 1241 1084
```

```
        ---------------------------------------------------------------------------

        ValueError                                Traceback (most recent call last)

        <ipython-input-10-9e6574f27091> in <module>
          7         if lines != line[-1]:
          8             for i in range(19):
    ----> 9                 sys.stdout.write(chr(int(line[i*6:(i+1)*6],2)))
         10         print("")


        ValueError: invalid literal for int() with base 2: b''
```

this error depends on the formato of the file end, the reading is correct

5. Load the file "user_data.json", filter the data by the "CreditCardType" field equals to "American Express". Than save the data a to CSV.

```
In [ ]: import csv
        import json

        dataset= json.load(open("user_data.json"))
        print(dataset[0].keys())

        for user in dataset:
            if user['CreditCardType']=='American Express':
                print(user.values())

        with open("Exercize5_OSEMN.dat", "w") as outfile:
            w= csv.writer(outfile, delimiter=',')
            w.writerows(dataset[0].keys())
            for user in dataset:
                if user['CreditCardType']=='AmericanExpress':
                    w.writerows(user.values())
```

6. Load the file from this url: https://www.dropbox.com/s/7u3lm737ogbqsg8/mushrooms_categorized.csv?dl=1 with Pandas. + Explore the data (see the info of the data) + Draw the istogram of the 'class' field. Decribe wath yuou see

```
In [ ]: import pandas as pd
        import io
        import requests
        url="https://www.dropbox.com/s/7u3lm737ogbqsg8/mushrooms_categorized.csv?dl=1"
        s=requests.get(url).content
        c=pd.read_csv(io.StringIO(s.decode('utf-8')))
```

```
In [ ]: c.hist(column="class")
        print("there are only 2 classes almost equivalently in numer")
```

7. Load the remote file https://www.dropbox.com/s/vkl89yce7xjdq4n/regression_generated.csv?dl=1 with Pandas and plot a scatter plot all possible combination of the following fields:

- features_1
- features_2
- features_3

```
In [ ]: import pandas as pd
        import io
        import requests
        url="https://www.dropbox.com/s/vkl89yce7xjdq4n/regression_generated.csv?dl=1"
        s=requests.get(url).content
        c=pd.read_csv(io.StringIO(s.decode('utf-8')))
```

```
In [ ]: c.plot.scatter("features_1","features_2")
        c.plot.scatter("features_1","features_3")
        c.plot.scatter("features_2","features_3")
```

8. Load the same file of point 6, and convert the file to json with Pandas.

```
In [ ]: import pandas as pd
        import io
        import requests
        url="https://www.dropbox.com/s/7u3lm737ogbqsg8/mushrooms_categorized.csv?dl=1"
        s=requests.get(url).content
        c=pd.read_csv(io.StringIO(s.decode('utf-8')))
        c.to_json("mushrooms.json")
```

```
In [ ]:
```