Machine Learning Workshop 1 – CSE 2793

MINOR ASSIGNMENT-7: NUMPY, MATPLOTLIB & PANDAS

Q 01 **Matrix Operations**

a. Create a matrix as follows.

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix}$$

b. Create a vector as follows.

$$y = [1, 2, 3, 4, 5, 6]$$

c. Select and display X_{13}, X_{22}, X_{32}.

d. Select and display the first three and the last three elements of y separately.

e. Find the size, number of elements and number of dimensions of both X and y.

f. Generate and display $X_{10} = X + 10$.

g. Find the maximum and minimum values of both X and y.

h. Calculate the and display Average, Variance, and Standard Deviation values of both X and y.

i. Reshape X as a 2×6 matrix.

j. Find and display the transpose of both X and y.

k. Flatten X and display.

l. Generate a 3×3 square matrix (P).

$$P = \begin{bmatrix} 1 & 3 & 6 \\ 1 & 4 & 5 \\ 2 & 2 & 7 \end{bmatrix}$$

m. Find and display the rank.

n. Find and display the determinant of P.

o. Find and display the diagonal of P.

p. Find and display the trace of P.

q. Find the eigen values and eigen vectors of P.

r. Generate $z = [2, 3, 6, 2, 5\ 2]$. Find the dot product between y and z.

s. Generate a 3×3 square matrix (Q).

$$Q = \begin{bmatrix} 4 & 6 & 7 \\ 5 & 5 & 2 \\ 3 & 4 & 6 \end{bmatrix}$$

Find and display $P + Q$, $P - Q$

t. Find and display element wise product of P and Q.

u. Find and display inverse of of P and Q.

v. Generate 10 no.s of random numbers of uniform, gaussian and logistic distributions.

Q 02 Search the internet for the following datasets, download and load the same.

a. Boston housing prices
b. Iris flowers
c. Handwritten digits

In each case

a. Display the first 5 rows.
b. Display the number of rows and columns of the data.
c. Display the descriptive statistics of all the columns.

Q 03 Generate a simulated dataset for regression application using NumPy, with the following properties.

a. Number of samples = 100

b. Number of features = 4
c. Number of targets = 1
d. Zero noise

Q 04 Generate a simulated dataset for classification application with the following properties.

a. Number of samples = 100
b. Number of features = 4
c. Number of classes = 2
d. Zero noise

Q 05 Perform the following tasks with pandas **`Series`**:

a) Create a **`Series`** from the list **`[7, 11, 13, 17]`**.
b) Create a **`Series`** with five elements that are all 100.0.
c) Create a **`Series`** with 20 elements that are all random numbers in the range 0 to Use method describe to produce the **`Series'`** basic descriptive statistics.
d) Create a **`Series`** called **`temperatures`** of the floating-point values 98.6, 98.9, 100.2 and 97.9. Using the **`index`** keyword argument, specify the custom indices **`'Julie', 'Charlie', 'Sam' and 'Andrea'`**.
e) Form a dictionary from the names and values in Part (d), then use it to initialize a Series.

Q 06 Consider Sample Python dictionary data and list labels:

```
exam_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James',
'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],
'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],
'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],
'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no',
'yes']}

labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
```

a) Write a program to create and display a DataFrame from the above specified dictionary data with the corresponding index labels.
b) Write a program to change the name **`'James'`** to **`'Suresh'`** in the name column of the DataFrame.
c) Write a program to insert a new column named **`'color'`** in the existing DataFrame and add corresponding color values.
d) Write a program to get list from DataFrame column headers.

Q 07 An NGO has participated in a three-week cultural festival. Using Pandas, store the sales (in Rs) made day wise for every week in a CSV file named **`"FestSales.csv"`** as shown in the table below.

Week 1	Week 2	Week 3
5000	4000	4000

5900	3000	5800
6500	5000	3500
3500	5500	2500
4000	3000	3000
5300	4300	5300
7900	5900	6000

Write a Python script to display the sales for the three weeks using a Line chart. It should have the following:

i. Chart title as "Festival Sales Report".
ii. X - axis label as Days.
iii. Y - axis label as "Sales in Rs".
iv. Line colours are red for week 1, blue for week 2 and brown for week 3.

Q 08 To the above **"FestSales.csv"** data, add a Day column that contains the different days of week, as shown below.

Day	Week 1	Week 2	Week 3
Monday	5000	4000	4000
Tuesday	5900	3000	5800
Wednesday	6500	5000	3500
Thursday	3500	5500	2500
Friday	4000	3000	3000
Saturday	5300	4300	5300
Sunday	7900	5900	6000

Write a python script to display Bar plot for the **"FestSales.csv"** file with column Day on x axis.

Q 09 Perform the following tasks using Python and associated libraries.

a) Download the data from the following link and keep in your working directory. https://people.sc.fsu.edu/~jburkardt/data/csv/trees.csv
b) Display the number of rows and columns in the data.
c) Display the first and last three data.
d) Generate the statistical overview of the dataset.
e) Generate the statistical overview of the dataset displaying only three digits after decimal point.
f) Find the correlation between the attributes. Comment on the results.
g) Since the data is spread over a wide range with different scales, it is not suitable to train models. Hence, bring the data into the range of [0 - 1].

Q 10 The **statsmodels** package (installed in the code cell above) includes built-in datasets. Execute the code below to download data from the **American National Election Studies of 1996** and print a detailed description of the schema.

```
import pandas as pd
import statsmodels.api as sm
import numpy as np

anes96 = sm.datasets.anes96
print(anes96.NOTE)
```

a) The DataFrame (df) contains data on registered voters in the United States, including demographic information and political preference. Using pandas, print the first 5 rows of the DataFrame to get a sense of what the data looks like.

b) Answer the following questions.
   i. How many observations are in the DataFrame?
   ii. How many variables are measured (how many columns)?
   iii. What is the age of the youngest person in the data? The oldest?
   iv. How many days a week does the average respondent watch TV news (round to the nearest tenth)?
   v. Check for missing values. Are there any?

c) We want to adjust the dataset for our use. Do the following:
   i. Rename the **`educ`** column as **`education`**.
   ii. Create a new column called **`party`** based on each respondent's answer to PID.
      - **`party`** should equal Democrat if the respondent selected either Strong Democrat or Weak Democrat.
      - **`party`** will equal Republican if the respondent selected Strong or Weak Republican for PID and
      - **`party`** will equal Independent if they selected anything else.
   iii. Create a new column called **`age_group`** that buckets respondents into the following categories based on their age: 18-24, 25-34, 35-44, 45-54, 55-64, and 65 and over.