DS5003 Data Engineering Lab

Midterm, Date: 13/09/ 2024

Timing: 2:00 to 4:30 PM Max Marks: 10

Instructions

1. Submit one .ipynb file containing all answers. The name should be **[student name]_midterm.ipynb**
2. Write questions in separate text blocks in Jupyter Notebook before the code blocks containing answers.
3. Read the questions carefully before answering. If a question asks to follow a particular approach or use a specific data structure, it must be followed.

---

1. Given an array of n elements, compute the number of movements (element shifts) required by Insertion Sort to

(a) Sort half of the elements in an array (Half of the elements are in their place) 1

(b) Sort the remaining elements to achieve a fully sorted array. 1

Input : [1,1,4,2,1,3]
The shifts over iterations= {0,0,0,1,2,1}
Output: a= 0, b = 4

2. You are provided with a time series dataset representing monthly temperatures in a certain city:

Temperatures = [32.4, 30.1, 29.0, 25.5, 22.3, 24.0, 27.9, 28.7, 29.5, 30.9, 32.0, 33.5]

(a) **Calculate the moving average** of the temperatures with a window size of user input. 1

Hint: A moving average helps smooth out fluctuations in data. For a window size of 3, the moving average at any point is the average of the current value and the two preceding values. For example:

- For the 3rd month (29.0), the moving average is (32.4+30.1+29.0)/3=30.5(32.4 + 30.1 + 29.0) / 3 = 30.5(32.4+30.1+29.0)/3=30.5.

(b) **Perform a Fourier Transform** on the temperature data to identify the most significant frequency components. 1

Formula: The Discrete Fourier Transform (DFT) of a time series $x[n]$ with N points is given by:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-i \cdot 2\pi \cdot k \cdot n/N}, \quad k = 0, 1, 2, \ldots, N-1$$

Where:

- $X[k]$ is the frequency component at frequency k,
- $x[n]$ is the temperature at time n,
- N is the total number of data points,
- i is the imaginary unit ($\sqrt{-1}$),
- $e^{-i \cdot 2\pi \cdot k \cdot n/N}$ is a complex exponential representing the oscillation at frequency k.

This formula transforms the temperature data from the time domain into the frequency domain, allowing you to identify periodic patterns such as seasonal cycles.

3. Image segmentation is the process of splitting an image into multiple parts where one part is coherent in terms of color and intensity of color. Find the image shared along the assignment. Do segmentation on the image. **4**

You can use the below code to import the image using openCV and visualize it.

```
# Install OpenCV if not already installed
!pip install opencv-python

#import libraries
import cv2
import matplotlib.pyplot as plt


image_path = 'image.jpg'
# Load the image from the specified file path using OpenCV
image = cv2.imread(image_path) #each entry in 'image' denotes the grayscale pixel values ranging[0-255]


# Display the image using Matplotlib
plt.imshow(image)
plt.axis('off')
plt.show()
```

4. (a). Plot histogram of arrays of random numbers drawn from a normal distribution with sizes 100 and 1000. 1

(b). Import the “insurance.csv” file and do the following task. Visualize the total insurance charges for smokers and non-smokers in each region, grouped by gender, using a bar plot. 1