Lab Course Machine Learning

Exercise Sheet 1

Prof. Dr. Dr. Lars Schmidt-Thieme
Jung Min Choi

HiWi: Harish Malik

Submission deadline : October 28, 2024

1 Advanced Python Challenges (10 points)

1. **[3 points]** You are given a dataset representing the weights and heights of individuals. Your tasks are:
   a) Download the dataset from *heights_ weights.csv*[1] . Compute the mean and standard deviation for both the weight and height columns.
   b) Plot histograms of the weights and heights to visualize their distributions.
   c) Compute the correlation between height and weight from scratch, without using any library functions for correlation (use only basic NumPy operations).

2. **[3 points]** You are given the following dataset representing the growth of bacteria over time:

Time (hours)	Bacteria Count
0	100
1	180
2	324
3	583.2
4	1049.76
5	1889.57

   a) Fit the following exponential growth model to the data:

$$y = ae^{bt} \tag{1}$$

   where:

   - y represents the bacteria count,
   - t represents the time (in hours),
   - a is the initial bacteria count, which you will determine by fitting the model,
   - b is the exponential growth rate, which you will also determine by fitting the model.

   b) Use nonlinear regression to fit the model to the data, and plot both the data points and the fitted curve. Compute the coefficient of determination (R^2) for the model.

   Hint: You may use the `curve_fit` function from `scipy.optimize` for nonlinear fitting, which will give you the values of a and b.

3. **[4 points]** In this problem, you will create an animated 3D plot of a torus rotating about an axis.
   a) The parametric equations of a torus are:

$$x(u,v) = (R + r\cos v)\cos u, \quad y(u,v) = (R + r\cos v)\sin u, \quad z(u,v) = r\sin v$$

   where $u \in [0, 2\pi]$, $v \in [0, 2\pi]$, R is the distance from the center of the tube to the center of the torus, and r is the radius of the tube.
   b) Set $R = 5$ and $r = 2$. Create a 3D mesh plot of the torus.
   c) Create an animation where the torus rotates about the z-axis. The animation should run for 10 seconds and display the rotating torus.

[1] https://www.kaggle.com/datasets/burnoutminer/heights-and-weights-dataset

d) Save the animation as a `.gif` file.

Instructions:

- You may refer to the following documentation for help with 3D plotting and animations in `Matplotlib`:
  - https://matplotlib.org/stable/api/toolkits/mplot3d.html
  - https://matplotlib.org/stable/api/animation_api.html
- Make sure to save your animated torus as a `.gif` file using `FuncAnimation` from `Matplotlib`.

2 Exploratory Analysis on Real-World Data (10 points)

1. **[6 points]** In this task you are required to explore a real-world dataset from the airport dataset named *task1.txt*. You are required to the following:
   - Load the dataset using pandas and display all necessary information contained in the file
   - You are tasked as a data scientist to create a story that is visually appealing from this data. Create plots using *matplotlib/seaborn* that will depict such interesting stories from flights that depart from and arrive in the Austin region. The figures should be annotated properly and also easily understandable on the first glance. A list of questions that can be explored/answered as reference are given below. Of course, you are free to explore any other possibilities.
     - Investigate what time of the day it is best to fly so as to have the least possible delays. Does this change with airlines?
     - Investigate what time of the year it is more suited to fly so as to have the delays minimum and does the destination affect this? You can lay insights on some popular destinations for the task.
     - Explore some airports that are bad to fly to. Does the time of day or year affect this?
     - Investigate on how the pattern of flights to various destinations alter over the course of year.
2. **[4 points]** In this part we will examine the data containing information on every Olympic medallist that is listed by participant count in top 20 sports, dating back to 1896. Load the dataset *task2.txt* and perform statistical analysis on the dataset. Specifically, do the following:
   - Compute the 95^{th} percentile of heights for the competitors in all Athletic events for gender Female. Note that sport refers to the broad sports (Athletics) and event is the specific event (100-meter sprint).
   - Find the single woman's event that depicts the highest variability in the height of the competitor across the entire history of Olympics. Use the standard deviation as the yardstick for this.
   - We wish to know how the average age of swimmers in Olympic has evolved with time. How has this changed over time? Does the trend for this differs from male to female? It will be easy to create a data frame that will allow one to visualise these trends with time. Plot a line graph that depicts separate line for male and female competitors. The plot must have a caption that is informative enough to answer the 2 questions that have been asked in this part.