Lab Course Machine Learning

Exercise Sheet 1

Prof. Dr. Dr. Lars Schmidt-Thieme
Kiran Madhusudhanan

HiWi: Basharat Mubashir Ahmed
Sophia Damilola Lawal
Submission deadline : November 11, 2022

1 Python Warmup (10 points)

1. **[2 points]** In this part of the assignment, you have to write a word count program. Your program should read the provided text document on learnweb named *random text.txt* and then output the following stats:
   a) The number of unique non-stop words.
   (Hint: you can use "nltk" library to get a list of English language stop words.)
   b) The top 5 most frequent non-stop words.

2. **[2 points]** In a simple regression problem we fit a straight line $y = mx + b$ to a given data. However, not all problems in nature are by default linear. Given the data below see if a straight line is a good fit.

x	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	6.0	4.83	3.7	3.15	2.41	1.83	1.49	1.21	0.96	0.73	0.64

   In cases where the data does not follow a linear trend, one can transform the variables and then apply the linear regression technique to better fit the data. From the given choices, try which function would be a better representation for the data.

   a) Linear : $y = mx + b$
   b) Power : $y = bx^m$
   c) Exponential : $y = be^{mx}$
   d) Logarithmic : $y = m\log x + b$
   e) Reciprocal : $y = \frac{1}{mx+b}$

   Generate a 2 x 2 subplot with the following techniques, *plot, semilogx, semilogy, loglog*. Read about these plotting techniques. These plots will let you understand which of the above 5 choices will be the best fit. Plot the data points and the best fit curve in a well-formatted plot with axis labels, title and the legend.
   (Hint: you can use the *polyfit* function from *numpy* for this part.)

3. **[3 points]** In this part of the task, you will implement the linear regression from scratch. The task is to have a user-defined function that will fit data points to a power function of the form $y = bx^m$. The function will be named *powerfit(x,y)*. It accepts 2 arguments x and y and must return the coefficients b, m which are the constants of the fitted equation. Use your function on the data below and generate a plot that shows the data points and also the fitted function.

x	0.5	2.4	3.2	4.9	6.5	7.8
y	0.8	9.3	37.9	68.2	155	198

You are required to implement the algorithm below for the task.

Listing 1: Simple Linear Regression

$$
\begin{aligned}
&\texttt{learn-simple-linreg}\,(x, y):\\
&\quad \bar{x} := \tfrac{1}{N}\textstyle\sum_{n=1}^{N} x_n\\
&\quad \bar{y} := \tfrac{1}{N}\textstyle\sum_{n=1}^{N} y_n\\
&\quad \hat{\beta}_1 := \frac{\sum_{n=1}^{N}(x_n-\bar{x})(y_n-\bar{y})}{\sum_{n=1}^{N}(x_n-\bar{x})^2}\\
&\quad \hat{\beta}_0 := \bar{y} - \hat{\beta}_1\bar{x}\\
&\quad \texttt{return}\ \ (\hat{\beta}_0, \hat{\beta}_1)
\end{aligned}
$$

4. **[3 points]** In this part of the assignment, we will explore the 3D plotting capabilities in Python. Specifically, we will generate a 3D plot of an ice cream cone. The cone is 8 inches tall and has a base diameter of 4 inch. Furthermore, the top of the ice-cream is a hemisphere of 4-inch diameter. We define the following parametric equations for the cone:

$$x = r\,cos\theta, y = r\,sin\theta, z = 4r$$

with $\theta \in [0, 2\pi]$ and $r \in [0, 2]$
Also, for the top of ice-cream which is hemisphere, the equations are:

$$x = r\,cos\theta\,sin\phi, y = r\,sin\theta\,sin\phi, z = 8 + r\,cos\phi$$

with $\theta \in [0, 2\pi]$ and $\phi \in [0, \pi]$ Generate the 3D mesh plot for the ice-cream cone described by the equations. You are required to properly format the plot with axis labels, titles and grid. Specifically, you are required to learn how to produce well-formatted plots with symbols and latex formatting in the title and label. A poor formatting of plot will be penalized. Use *numpy* for all computations.

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.