Lab Course Machine Learning

In-Class Exercise Sheet 1

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

HiWi: Harish Malik
Submission deadline : October 28, 2024

1 Variables

A company specializes in spherical segments and wants to automate the calculation of the lateral surface area, total surface area, and volume. Implement these calculations in Python using a **class**. The formulas are given as:

$$\begin{aligned}
\text{R} &= \sqrt{h(2r-h)} \\
\text{Lateral Surface Area} &= \pi R r \\
\text{Total Surface Area} &= \pi r(2h + \sqrt{h(2r-h)}) \\
\text{Volume} &= \frac{2}{3}\pi r^2 h
\end{aligned}$$

You should define a class that contains methods for each of the calculations (lateral surface area, total surface area, and volume). For the calculation, a variable is required for storing the result, as well as separate variables for the radius and the height. What is the result for the following values? Print the results using a print() statement.

$$\begin{aligned}
r &= 4.0 \\
h &= 2.0
\end{aligned}$$

Note: Not all functions are available in Python by default. Additional functionalities can be imported from packages. This applies, for example, to the constant π and the square root operator.
Use the 'math' package by adding '**import math**' at the beginning of your code. Afterward, all functions from this package can be used in the code. For example, the following holds:

$$\begin{aligned}
\sqrt{x} &= math.sqrt(x) \\
\pi &= math.pi \\
x^N &= x**N
\end{aligned}$$

2 Python Warmup

1. 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. 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. Use the provided data below and apply linear regression using the `LinearRegression` class from `scikit-learn`. You will fit a linear regression model and then generate a plot that shows both the original data points and the fitted curve on the same graph.

   Additionally, you are required to compute and report the following evaluation metrics for both the training and test data:

   - Root Mean Squared Error (RMSE)
   - Mean Absolute Error (MAE)
   - Mean Squared Error (MSE)

   Steps:

   a) Convert the provided training data to a logarithmic scale (both x and y).
   b) Use `LinearRegression` from the `scikit-learn` library to fit the linear regression model to the transformed training data.
   c) Extract the slope and intercept of the fitted line.
   d) Convert the intercept back to its original scale to get the coefficient b.
   e) Using additional test data, compute the RMSE, MAE, and MSE for both the training data and the test data using the original (non-logarithmic) values.
   f) Generate a plot that includes both the original training data points, the test data points, and the fitted curve based on the power function.

   Training Data:

x	0.5	2.4	3.2	4.9	6.5	7.8	9.1	10.3
y	0.8	9.3	37.9	68.2	155	198	305	425

   Test Data:

x	1.0	5.5	8.0	11.0
y	1.5	50	215	480

4. 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.