Exercise 1

Prof. Dr. Amr Alanwar

October 2024

Question 1: Pandas and Numpy

Task a:

Write a program that reads the provided text file and counts the occurrences of unique words. The program should include all words in the count.

Task b:

Modify the program to exclude common words such as 'the', 'a', 'an', and 'be' from the count. Finally, generate a histogram displaying the top 10 most frequent words.

Text file for tasks a & b: `https://raw.githubusercontent.com/aalanwar/Logical-Zonotope/refs/heads/main/README.md`

Task c:

Using Numpy for matrix operations: Create a matrix A with dimensions 100 × 20 (n = 100, m = 20). Initialize matrix A with random values. Then, create a vector v of size 20 × 1 and initialize it with values from a normal distribution, where the mean μ is 2 and the standard deviation σ is 0.01. Perform the following operations:

- Iteratively multiply each row of matrix A element-wise by vector v, and accumulate the results into a new vector c.
- Calculate the mean and standard deviation of vector c.
- Plot a histogram of vector c using 5 bins.

Question 2: Linear Regression

- Generate 3 sets of simple data. i.e. a matrix A with dimensions 100 × 2. Initialize it with normal distribution $\mu = 2$ and $\sigma = [0.01, 0.1, 1]$.
- Implement the "Learn Simple Linear Regression" algorithm and train it using matrix A to learn values of β_0 and β_1.
- Implement the "Predict using Simple Linear Regression" algorithm and calculate the points for each training example in matrix A.
- Plot the training points from matrix A and predicted values in the form of line graph.
- Comment on the effect that σ has on the line that is predicted.
- Put β_0 to zero and rerun the program to generate the predicted line. Comment on the change you see for the varying values of σ
- Put β_1 to zero and rerun the program to generate the predicted line. Comment on the change you see for the varying values of σ

- Use numpy.linalg.lstsq to replace step 2 for learning values of β_0 and β_1.
- Use sklearn.linear_model.LinearRegression to replace step 2 for learning values of β_0 and β_1.
- Comment and compare between the results, do you see differences? why?

Algorithm 1 Learn Simple Linear Regression

```
procedure
        (LEARN-SIMPLE-LINREG) $D^{\text{train}} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \in \mathbb{R} \times \mathbb{R}$
    $\bar{x} := \frac{1}{N} \sum_{n=1}^{N} x_n$
    $\bar{y} := \frac{1}{N} \sum_{n=1}^{N} y_n$
    $\hat{\beta}_1 := \frac{\sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N} (x_n - \bar{x})^2}$
    $\hat{\beta}_0 := \bar{y} - \hat{\beta}_1 \bar{x}$
    return $(\hat{\beta}_0, \hat{\beta}_1)$
end procedure
```

Algorithm 2 Predict using Simple Linear Regression

```
procedure PREDICT-SIMPLE-LINREG($x \in \mathbb{R}, \hat{\beta}_0, \hat{\beta}_1 \in \mathbb{R}$)
    $\hat{y} := \hat{\beta}_0 + \hat{\beta}_1 x$
    return $\hat{y}$
end procedure
```