CSE 4554
Machine Learning Lab

Homework No: 1
Introduction to Python for Machine Learning

Hasan Mahmud, Ph.D.
Associate Professor, Department of CSE, IUT
Md. Shihab Shahriar
Lecturer, Department of CSE, IUT

August 24, 2023

Contents

1 Objectives

- To learn the basics of Python Programming which mainly focuses in ML
- Introducing Matrix Multiplication in Numpy
- Introducing the Python Library Pandas
- Using Pandas to load the IRIS dataset

2 Problem Discussion

2.1 Introduction to matrix multiplication and Python Library Pandas

In Python, the NumPy library provides support for multi-dimensional arrays and matrices. We can perform matrix multiplication using the dot() method.

```
import numpy as np

A = np.array([[1, 2],
              [3, 4]])

B = np.array([[5, 6],
              [7, 8]])

C = A.dot(B)

print(C)
```

This will print the result:

```
[[19 22]
 [43 50]]
```

Pandas is a popular Python library used for data analysis and manipulation. It provides easy-to-use data structures and data analysis tools to enable quick data munging and analysis.

To use Pandas, you first need to import it:

```
import pandas as pd
```

The main data structures in Pandas are Series and DataFrames.

A Series is a one-dimensional array-like structure that can hold data of any type (integers, strings, floats, etc). Here's an example Series:

```
data = [1, 2, 3, 4, 5]
ser = pd.Series(data)
```

A DataFrame is a two-dimensional tabular data structure with labeled columns that can hold different data types. You can create one from lists of data:

```
data = [[1, 2, 3], [4, 5, 6], [7, 8, 9]]
df = pd.DataFrame(data, columns=['Col1', 'Col2', 'Col3'])
```

Some key Pandas functions include:

- **df.head() -** returns first n rows of the DataFrame
- **df.tail() -** returns last n rows of the DataFrame
- **df.shape -** returns number of rows and columns in DataFrame

- **df.describe() -** generates statistics for numeric columns
- **df.mean() -** calculates mean of each column
- **df.groupby() -** groups DataFrame by one or more columns
- **df.sort_values() -** sorts DataFrame by one or more columns
- **df.plot() -** plots the DataFrame
- **df.loc -** selects data based on label
- **df.iloc -** selects data based on position

This covers some basic Pandas concepts and functions to help get you started!

3 Tasks

3.1 Matrix Multiplication

1. Generate two random 2x2 matrices A and B. Multiply them together and print the result.
2. Create two 3x3 matrices C and D. Manually calculate the matrix multiplication CD and verify it is the same as C.dot(D).
3. Given a matrix A, calculate $A^T A$ where ^{T} denotes the transpose.
4. Multiply a 4x3 matrix B with a 3x2 matrix A. Confirm the dimensions of the resulting matrix.
5. Given two matrices A and B, show that in general $A \cdot B \neq B \cdot A$.

3.2 Introduction to Python

1. Download the IRIS dataset as a CSV file and load the CSV file into a DataFrame. Examine the DataFrame using .head(), .tail(), .shape, .dtypes, etc.
2. Select a specific column from the DataFrame. Create a Plot to visualize the data in that column.
3. Filter the DataFrame to only show rows where a certain column value meets some criteria (e.g. only show rows where the 'Age' column is greater than 30).
4. Calculate summary statistics (mean, min, max, etc) for numerical columns in the DataFrame using .describe().
5. Group the DataFrame by one or more columns and calculate aggregates like count, mean, etc per group.
6. Sort the DataFrame values by a specific column in ascending or descending order.
7. Handle missing values in a DataFrame by dropping or filling.

4 Submission

Submit Your Google Colab notebook in the classroom with the following naming format <Student_id>_Homework<Homework_id>.