Deep Learning - Exercise 2

Fall 2024

To do these exercises, you will use Python 3 and the following packages:

- NumPy. This is highly recommended package to create various data transformation / generate data.
- Pandas. With this package you can import your data into a data frame similar to how it's done in R.
- Matplotlib. This package allows you to graph your data and data transformations.
- PyTorch. This package is a versatile deep learning framework that facilitates building and training of neural networks.
- Seaborn. This package is built on top of Matplotlib, providing a high-level interface for creating informative and aesthetically appealing statistical graphics in Python.

You are not strictly forced to use these packages, but it is highly recommended. Feel free to use other packages you think are necessary.

Warmup

1. NumPy
   a. Import NumPy in a python script.

```
a = np.full((2, 3), 4)
b = np.array([[1, 2, 3], [4, 5, 6]])
c = np.eye(2, 3)
d = a + b + c
```

   b. Think about which values are in the NumPy array ‘d’, then verify if you were correct.

```
a = np.array([[1,2,3,4,5],
              [5,4,3,2,1],
              [6,7,8,9,0],
              [0,9,8,7,6]])
```

   c. Sum the rows of ‘a’.
   d. Get the transpose of ‘a’.

2. Pandas
   a. Import pandas.
   b. Read the file ‘auto.csv’.
   c. Remove all rows with ‘mpg’ lower than 16.
   d. Get the first 7 rows of the columns’ weights’ and ‘acceleration’.
   e. Remove the rows in the ‘horsepower’ column that has the value ‘?’, and convert the column to an ‘int’ type instead of a ‘string’.
   f. Calculate the averages of every column, except for ‘name’.

3. PyTorch
4. Import Pytorch
    a. Create two random matrices using PyTorch’s (**torch.rand**) of size (3x3).
    b. Multiply the two matrices using PyTorch's matrix multiplication function (**torch.matmul**).

Exercise 2

The overall idea of this exercise is to predict the fuel consumption of cars (measured in miles-per gallon, mpg) for various cars based on a linear regression model. The dataset is available at the course website (auto.csv)

1. Load the **auto.csv** dataset again using the **pandas.read** function and remember to remove the missing values in the dataset, indicated by ‘?’, and then make sure the corresponding columns are casted to a numerical type.

2. Inspect the data. Plot the relationships between the different variables and mpg. Use for example the **matplotlib.pyplot** scatter plot. Do you already suspect what features might be helpful to regress the consumption? Save the graph.

3. Perform a linear regression using the OLS function from the statsmodels package. Use 'horsepower' as feature and regress the value 'mpg'. It is a good idea to look up the statsmodels documentation on OLS, to understand how to use it. Further, plot the results including your regression line.

4. Now extend the model using all features. How would you determine which features are important and which aren't? Try to find a good selection of features for your model.

5. Can you improve your regression performance by trying different transformations of the variables, such as $log(X)$, $\sqrt{X}$, $1/X$, X^2 and so on. For each transformation, which features are important and which aren’t?