Deep Learning Programming Exercise № 2

DM873 / DS809, Fall 2022 28/09/2022

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

- NumPy. This is often a good package to use, in order 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.
- Statsmodel. This package includes a good OLS function to perform your regression.

These can all be installed using pip, the python package manager. You are not strictly forced to use these packages, but it is highly recommended. Feel free to use other packages you think are necessary.

Installations

Run the following commands on terminal if not installed;

Listing 1: Installations of required Python packages

```
pip install numpy
pip install pandas
pip install matplotlib
pip install statsmodel
```

Problem

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 $\mathrm{auto.csv}$

What to do?

- Download the $\mathrm{auto.csv}$ from the course website and load it into python. Use the $\mathrm{pandas.read\_csv}$ function for importing the dataset. Be aware, there are some missing values in the dataset, indicated by $?$. You have to remove those lines and then make sure the corresponding columns are casted to a numerical type.
- Inspect the data. Plot the relationships between the different variables and mpg. Use for example the $\mathrm{matplotlib.pyplot}$ scatter plot. Do you already suspect what features might be helpful to regress the consumption? Save the graph.
- Perform a linear regression using the OLS function from the statsmodels package. Use $\mathrm{horsepower}$ as feature and regress the value mpg. It is a good idea to look up the statsmodels documentation on OLS, to under- stand how to use it. Further, plot the results including your regression line.
- 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.
- Can you improve your regression performance by trying different transformations of the variables, such as $\log x$, $\sqrt{x}$, $\frac{1}{x}$, x^2 and so on. Why are some transformations better?