Programming Assignment 1

Write a program (in a group up to maximally 4 people) that implements a (batch) linear regression using the gradient descent method in *Python 3*. Use the following gradient calculation:

$$gradient = \sum_{i=1}^{N} \vec{x_i}(y_i - f(\vec{x_i}))$$

$$\vec{w} \leftarrow \vec{w} + \eta \cdot gradient$$

where $\vec{x_i}$ is one data point (with N being the size of the data set), η the learning rate, y_i is the target output and $f(\vec{x_i})$ is the linear function defined as $f(\vec{x}) = \vec{w}^T\vec{x}$ or equivalently $f(\vec{x}) = \sum_i w_i \cdot x_i$. Whereas $\vec{w}$ and $\vec{x}$ include the bias/intercept, i.e. w_0 and $x_0 = 1$. All weights should be initialized as 0.
Given are the two data sets[1] named *yacht* and *random* as csv files. Your program should be able to read both data sets and treat the last value of each line as the target output. Your task is to correctly implement the gradient descent method and return for each iteration the weights and sum of squared errors until a given threshold of change in the error is reached. The output of your algorithm **must** look like this:

```
iteration_number,weight0,weight1,weight2,...,weightN,sum_of_squared_errors
```

The solution (rounded to 4 decimals) for the *random* data set is given with a learning rate of 0.0001 and a threshold of 0.0001. With that, you can check the correctness of your solution. Please be reminded, that small rounding errors are normal and will be treated as correct. For each data set, you can acquire one point, if the solution of your program returns correct results. If the program fails, the data format is incorrect or I have to change source code, in order to make it work, you will get zero points.
Your program must accept the following parameters:

1. **threshold** - The threshold, that the change in error has to fall below, before the algorithm terminates.
2. **data** - The location of the data file (e.g. /media/data/yacht.csv).
3. **learningRate** - The learning rate of the gradient descent approach.

Therefore, I should be able to start your program like this:

```
python3 linearregr.py --data random.csv --learningRate 0.0001 --threshold 0.0001
```

The final program code must be sent via email until Sunday, 3rd of November 2019, 23:59 to your respective tutor. Please format your e-mail header as follows:

[1] http://wwwiti.cs.uni-magdeburg.de/iti_dke/Lehre/Materialien/WS2019_2020/ML/res/linreg.zip

[Exercise Group] ML Programming Assignment 1

Replace *Exercise Group* with the day and time of your exercise group. E.g for Monday from 13:00 to 15:00 it would be:

[Monday 13-15] ML Programming Assignment 1

Do not forget to include your names and matriculation numbers in the mail! Please also be prepared to present your solution shortly in front of the class. You will get one point for each data set, if the output of your regressor is correct.

2 points