ΥΣ19 **Artificial Intelligence II (Deep Learning for Natural Language Processing)**
Fall Semester 2021
Homework 1
20% of the course mark
Announced: November 5, 2021 Due: November 25, 2021 before 23:59)

1. (Linear algebra and calculus warm up). Prove the equation

$$\nabla_{\mathbf{w}} MSE(\mathbf{w}) = \frac{2}{m}(\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y}))$$

   which we presented on page 15 of the slides “Regression”.

   If you like math, you might also want to read the following in your spare time:

   - The paper “The Matrix Calculus You Need For Deep Learning” available at `https://arxiv.org/pdf/1802.01528.pdf`.
   - The book “Mathematics for Machine Learning” available at `https://mml-book.github.io/`.

   (1/10 marks)

2. In this exercise you will develop a vaccine sentiment classifier using softmax regression. The classifier will be trained on twitter data and will distinguish 3 classes (neutral, anti-vax and pro-vax). You will train your classifier using the datasets provided on e-class for this homework.[1] The datasets are CSV files with the tweet texts and the labels 0 (neutral), 1 (anti-vax) or 2 (pro-vax). The first dataset is the training dataset and the second one is the validation dataset. Your model will be evaluated by us on a test set that will not be provided to you, but will be of the same format as the training and validation datasets. Your code should be written in such a way so that the model can be evaluated on the test simply by changing the name of the file to load.

   Start by reading about softmax regression from the slides of the course, the relevant chapters 4 and 5 of the “Speech and Language Processing” book of Jurafsky and Martin (`http://web.stanford.edu/~jurafsky/slp3/`) or any other relevant literature which you may find.

[1] Thanks to Dr. Saptarshi Ghosh and Soham Poddar from the Department of Computer Science and Engineering, IIT Kharagpur, India for providing us these datasets.

You should use the toolkit Scikit-Learn for your implementation. You should plot learning curves that show that your models are not overfitting or underfitting. You should also evaluate your classifiers on the validation set using precision, recall and F-measure.

Finally, we would like to stress that the goal of this exercise is not to simply create a classifier that can achieve a good score, but to get acquainted with the methodology of an NLP project, to experiment with different options and to deliver a report that presents your solution along with a brief performance comparison between the different options that you have tried.

(9/10 marks)

Note: You should hand in: (i) a pdf with the answer to question 1 and a detailed report of your solution to question 2, including citations to relevant literature that you might have used in developing your solutions. (ii) one Colab notebook (ipynb files using `https://colab.research.google.com/`) containing your code for question 2. You should use Python 3.6 or a later version.