Artificial Intelligence II
Deep Learning for Natural Language Processing
Fall Semester 2023
Homework 1
25% of the course mark
Announced: October 27, 2023
Due: November 19, 2023 before 23:59

Description

In this homework you have to develop a **sentiment classifier** using **logistic regression** for a **twitter dataset about the Greek general elections**, which is provided. Each row in the dataset contains an ID for each tweet, a sentiment label which can be *POSITIVE, NEUTRAL* or *NEGATIVE*, the text of the tweet, and the party the tweet is referring to. Your classifier should deal with 3 classes: *POSITIVE, NEUTRAL, NEGATIVE*.

Before you start the homework, make sure that you have studied the relevant slides of the course ("Introductory concepts of machine learning" and "Regression") and 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 you may find useful.

It is your responsibility to choose all the details of developing a good model (e.g., whether to do cross validation, whether to do regularization, which gradient-based training algorithm to use, how to choose the hyperparameters of the algorithm, how to make sure that your model does not underfit or overfit etc.).

Evaluation

You should plot learning curves that show that your models are not overfitting or underfitting. Also, you should use the toolkit Scikit-Learn (`https://scikit-learn.org/stable/`) and evaluate your classifier using *precision, recall* and *F-measure*.

Kaggle

You will submit your code (in the form of a Jupyter Notebook) through a Kaggle competition. Make sure to do the following:

- Your team name must be your academic identification number (Αριθμός Μητρώου - sdiXXYYYYY).
- Your solution must be submitted as a Notebook that outputs a result file named "submission.csv", **NOT AS A FILE UPLOAD**! The result file must follow the format specified in the provided "sample_submission.csv" file and must contain the predictions that your model makes over the test set.

- You must share your Notebook on Kaggle with the Teaching Assistant responsible for grading this assignment. **DON'T SHARE YOUR NOTEBOOK PUBLICLY**!

Data

You can view the data here. You should read your dataset from your kaggle notebook. No need to download/upload it.

Report

For this project, and the next ones, you are asked to create a detailed report. For this reason we provide you with a template in LaTeX. You may use Overleaf online editor. Find the template **here**. Open OverLeaf, create an account if you don't have one already, and then upload the zip file by selecting: New project; Upload project; Select a .zip file; (it uses a pdfLaTeX compiler).

If you are having any issues in writing with LaTeX, you can write it to word/docs following the template in LaTeX. However we are strongly advice you, to create it in LaTeX, as Overlead now provides you with many shortcuts and abilities making it easier for you.

Grading

Implementation: Code, kaggle submission [**Total 70%**]

- Data processing: [**10%**]
- Model creation: [**20%**]
- Experiments: [**30%**]
- Fine-tuning & Optimization: [**10%**]

Report: Analysis and Presentation [**Total 30%**]

- Experiments: [**10%**]
- Analysis: [**15%**]
- Plots: [**5%**]

Submission guides

We expect you to:

1. Submit your **Jupyter Notebook** (and make is available to supervisors) in **Kaggle** and **only**.*
2. Submit your report in a **.pdf** format from e-class. Name your report like: **[full-id].pdf** (e.g. ZZZZZZXXYYYYY.pdf if you are a bachelor student in this department).

**We won't accept code submissions from e-class/e-mails, etc.*

Support

Sergios - Anestis Kefalidis (s.kefalidis[at]di.uoa.gr) will be supervising this assignment. Please submit your questions on Piazza under the corresponding directory (hw1).