TASK 4

Text regression

You are in the group Deepforgetting consisting of junterhol (junterhol@student.ethz.ch (mailto://junterhol@student.ethz.ch)), merklec (merklec@student.ethz.ch (mailto://merklec@student.ethz.ch)) and ribadov (ribadov@student.ethz.ch (mailto://ribadov@student.ethz.ch)).

1. READ THE TASK DESCRIPTION

2. SUBMIT SOLUTIONS

3. HAND IN FINAL SOLUTION

1. TASK DESCRIPTION

GENERATING SCORES FROM USER REVIEWS

Welcome to task #4, the last project of this year's Introduction to Machine Learning course at ETH!

When customers shop online, they can leave reviews for the purchased product. In this task, given a dataset containing the review's title and its content, you will try to infer the score of the product, which is a continuous value between 0 and 10.

DATA DESCRIPTION

You can download the project material here:

Download (/static/task4_hr35z9.zip) (4.4 MB)

In the handout, you will find the following material:

- **train.csv** - a csv containing the columns "title", "sentence", and "score". The score is always within the range [0, 10].
- **test_no_score.csv** - csv containing the columns "title" and "sentence".
- **sample.txt** - a sample submission file.
- **template_solution.py** - a template file that will guide you through the implementation of the solution.
- **template_solution.ipynb** - the template solution in an interactive notebook form.

template_solution.py provides a starting template structure for how you can solve the task, by filling in the TODOs in the skeleton code. It is not mandatory to use this solution template but it is recommended since it should make getting started on the task easier. The template solution uses the PyTorch (https://pytorch.org/) deep learning framework.

SUGGESTED APPROACH

The recommended approach is to use the transformers (https://huggingface.co/docs/transformers/index) library, which contains the necessary network architectures and pretrained transformers. It is recommended to look into the following transformers:

- DistilBERT (https://huggingface.co/docs/transformers/en/model_doc/distilbert#transformers.DistilBertModel)
- ALBERT (https://huggingface.co/docs/transformers/en/model_doc/albert#transformers.AlbertModel)
- GPT-2 (https://huggingface.co/docs/transformers/en/model_doc/gpt2#transformers.GPT2Model)

Note that while traditional approaches using pretrained transformers requires fine-tuning them for the desired task, it is **NOT** required to fine-tune the transformers to pass the hard baseline. Using frozen weights are sufficient to pass the hard baseline. To do so, one can run the transformer separately and save the results to a stand alone file. When needed, the results are loaded without additional computation. It is recommended to refer to previous tasks for examples on how this is done. Interested students with sufficient computational resources are encouraged to fine-tune the models. This best mimics real world use cases of pretrained transformers.

Due to the heavy computational cost associated with transformer inference, it is recommended to use a modern GPU for this task. Students are invited to use Google Colab or the ETH Euler cluster. If no GPU can be obtained, it is possible to pass the hard baseline using CPU-based computation on the ETH Euler cluster within a reasonable time limit.

EVALUATION

We are interested in accurate prediction of the review score from the content. Your submitted predictinos will be evaluated by the *mean square error (MSE)* to the true score. Lower is better. The MSE is defined as

$$\mathrm{MSE}(\mathbf{y}^*, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i^* - y_i)^2$$

SUGGESTED MATERIAL (OPTIONAL)

You will learn all the details about transformers in the last few lectures of this course. However, if you are interested about the inner workings of transformer models the following will prove useful.

- Last year's IML material on transformers.
  - Lecture slides. (https://las.inf.ethz.ch/courses/introml-s23/slides/introml-24-transformers.pdf)
  - Lecture recording (https://video.ethz.ch/lectures/d-infk/2023/spring/252-0220-00L/530d20a9-d02a-45e1-a141-42c54d510eae.html)
  - Tutorial recordings (https://video.ethz.ch/lectures/d-infk/2023/spring/252-0220-00L/82638d96-1686-404a-9e11-ee01a7ed5595.html)
- Hugging face tutorials on transformers. (https://huggingface.co/docs/transformers/index)
- 3Blue1Brown. But what is a GPT? Visual intro to transformers. (https://www.youtube.com/watch?v=wjZofJX0v4M)
- Attention is all you need. (https://arxiv.org/abs/1706.03762)

GRADING

We provide you with **one test set** for which you have to compute predictions. We have partitioned this test set into two parts (of the same size) and use it to compute a *public* and a *private* score for each submission. You only receive feedback about your performance on the public part in the form of the public score, while the private leaderboard remains secret. The purpose of this division is to prevent overfitting to the public score. Your model should generalize well to the private part of the test set. When handing in the task, you need to select which of your submissions will get graded and provide a short description of your approach. This has to be done **individually by each member** of

the team. We will then compare your selected submission to our baselines. This project task is graded with grades between **2.0 - 6.0**. Your grade is calculated using a weighted sum of your private and public score. We do not share the weights used for the grading, again to prevent overfitting to the public score. The weights are selected such that:

- To achieve the best grade (6.0), you need to perform better than the hard baseline in both private and public score.
- To pass the project (grade: 4.0), you need to perform better than the easy baseline in both private and public score.

The medium baseline is only used for your reference and is not considered in the grading. In addition, for the grading, we consider the code and the description of your solution that you submitted. The following **non-binding** guidance provides you with an idea on what is expected to pass the project: If you hand in a properly-written description, your source code is runnable and reproduces your predictions, and your submission performs better than the baselines, you can expect to have passed the assignment.

⚠ Make sure that you properly hand in the task, otherwise you may obtain zero points for this task.

PLAGIARISM

The use of open-source libraries is allowed and encouraged. However, we do not allow copying the work of other groups / students outside the group (including work produced by students in previous versions of this course). Publishing project solutions online is not allowed and use of solutions from previous years in any capacity is considered plagiarism. Among the code and the reports, including those of previous years, we search for similar solutions / reports in order to detect plagiarism. Use of GPT3 Copilot or similar code/language generation tools in any capacity for writing code or reports will be considered and treated as plagiarism in the context of this course. Basic code autocompletion such as those used in the default setup of Sublime Text 3 are permitted. If we find strong evidence for plagiarism, we reserve the right to let the respective students or the entire group fail in the IML 2024 course and take further disciplinary actions. By submitting the solution, you agree to abide by the plagirism guidelines of IML 2024.

FREQUENTLY ASKED QUESTIONS

⦿ WHICH PROGRAMMING LANGUAGE AM I SUPPOSED TO USE? WHAT TOOLS AM I ALLOWED TO USE?

You are free to choose any programming language and use any software library. However, **we strongly encourage you to use Python**. You can use publicly available code, but you should specify the source as a comment in your code.

⦿ WHAT TO DO IF I CAN'T RUN THE CODE/SETUP AN ENVIRONMENT ON MY PC?

If you are having trouble running your solution locally, consider using the ETH Euler cluster to run your solution. Please follow the Euler guide (/static/euler-guide.md). The setup time of using the cluster means that this option is only worth doing if you really can't run your solution locally.

⦿ AM I ALLOWED TO USE MODELS THAT WERE NOT TAUGHT IN THE CLASS?

Yes. Nevertheless, the baselines were designed to be solvable based on the material taught in the class up to the second week of each task.

⦿ IN WHAT FORMAT SHOULD I SUBMIT THE CODE?

You can submit it as a single file (main.py, main.ipynb, etc.; you can compress multiple files into a .zip) having max. size of 1 MB. If you submit a zip, please make sure to name your main file as *main.py* (possibly with other extension corresponding to your chosen programming language, e.g. .ipynb).

⦿ IN WHAT FORMAT SHOULD I SUBMIT THE REPORT?

The handin page of the submission server contains a simple textbox in which you should insert your report. It should consist of a couple of sentences explaining the main ideas and concepts of your solution. Every student writes and submits the report independently.

⦿ WILL YOU CHECK / RUN MY CODE?

We will check your code and compare it with other submissions. We also reserve the right to run your code. Please make sure that your code is runnable and your predictions are reproducible (fix the random seeds, etc.). Provide a readme if necessary (e.g., for installing additional libraries).

⦿ SHOULD I INCLUDE THE DATA IN THE SUBMISSION?

No. You can assume the data will be available under the path that you specify in your code.

⦿ CAN YOU HELP ME SOLVE THE TASK? CAN YOU GIVE ME A HINT?

As the tasks are a graded part of the class, **we cannot help you solve them**. However, feel free to ask general questions about the course material during or after the exercise sessions.

⦿ CAN YOU GIVE ME A DEADLINE EXTENSION?

> ⚠ We do not grant any deadline extensions!

⦿ CAN I POST ON MOODLE AS SOON AS I HAVE A QUESTION?

This is highly discouraged. Remember that collaboration with other teams is prohibited. Instead,

- Read the details of the task thoroughly.
- Review the frequently asked questions.
- If there is another team that solved the task, spend more time thinking.
- Discuss it with your team-mates.

⦿ WHEN WILL I RECEIVE THE PRIVATE SCORES? AND THE PROJECT GRADES?

We will publish the private scores, and corresponding grades before the exam the latest.