— GUIDE —

Assignment 5: Classify Math Publications

AI-2 Systems Project (Summer Semester 2024)

Jan Frederik Schaefer

Friedrich-Alexander-Universität Erlangen-Nürnberg, Department Informatik

This document is intended to help you solve the assignment "Assignment 5: Classify Math Publications" [AS]. You do not have to read it, but we do recommend to at least take a look at the tips and common issues.

1 A few tips

1. The pytorch tutorial for text classification is a good starting point [TC], but it might also be worth looking at the general introduction to pytorch [PT].
2. There are many parameters you can change and (especially without much experience) it is hard to tell, what effect an adjustment has. It can help to take notes on the performance of different configurations to inform future experiments.
3. Here are a few ideas how you can change the network structure:
   (a) Add one (or more) layers on top and try out different activation functions (can make a big difference).
   (b) Increase/decrease the size of the embedding layer.
   (c) Try out different cost functions.
   (d) Try out different batch sizes and learning rates.
4. Another important aspect is the text pre-processing. Here are a few things you can experiment with:
   (a) Replace rare words with a token `<UNK>`. Rare words could be words that occur less than n times in the training data (for some value of n).
   (b) Ignore frequent words that are very unspecific (e.g. `"the"`). Such words are called "stop words".
   (c) Do not distinguish different forms of a word (e.g. `"groups"` and `"group"` should be treated as the same word). A tool for reducing a word to its base form ("stem") is called a stemmer. Many libraries for stemming exist.

2 Common issues

1. Not submitting the solutions for the test dataset as specified in the *Evaluation* section.

References

[AS] *Assignment 5: Classify Math Publications*. URL: https://kwarc.info/teaching/AISysProj/SS24/assignment-2.5.pdf.

[PT] *PyTorch – Learn the Basics*. URL: https://pytorch.org/tutorials/beginner/basics/intro.html (visited on 08/01/2022).

[TC] *Text classification with the torchtext library*. URL: https://pytorch.org/tutorials/beginner/text_sentiment_ngrams_tutorial.html (visited on 08/01/2022).