Assignment 5: Classify Math Publications

AI-2 Systems Project (Winter Semester 2024/2025)

Jan Frederik Schaefer

Friedrich-Alexander-Universität Erlangen-Nürnberg, Department Informatik

Topic:	Artificial neural networks
Due on:	April 5, 2025
Version from:	January 24, 2025
Author:	Jan Frederik Schaefer

Make sure you sign up before working on this assignment.[a]
Using someone else's solution code, even as inspiration, is not allowed!

[a] You can still decide to postpone the assignment. Signing up includes an elligibility check, which avoids situations where you invest work into an assignment that you are not supposed to take.

1 Task summary

Train a neural network to classify mathematical publications based on the formulas they contain.

Didactic objectives

1. Learn the basics of a modern machine learning framework,
2. get some hands-on experience with a classification task.

Prerequisites and useful methods

1. Basics of neural networks and machine learning.

2 Detailed task description

The goal of this assignment is to learn the basics of a modern machine learning framework, by solving a simple classification problem.

The assignment guide [AG] provides some suggestions on how to approach the task using keras, but you are free to use any framework and approach you like.

Copyright © 2021–2025 Jan Frederik Schaefer, Michael Kohlhase.

2.1 The dataset

The dataset for this assignment is based on arXiv [arxiv] publications that were converted to HTML using LaTeXML [latexml].

The formulas are represented using MathML [mathml], which is an XML-based language for representing mathematical notation that is part of the HTML5 standard. For example, the formula

$$\frac{x}{2} \leq 15$$

is represented as

```
<math>
  <mrow>
    <mfrac>
      <mi>x</mi>
      <mn>2</mn>
    </mfrac>
    <mo>≤</mo>
    <mn>15</mn>
  </mrow>
</math>
```

The <mrow> element is used to group elements of a formula (and arrange them horizontally). <mfrac> is used for fractions, <mi> for identifiers, etc.

Each line of the training dataset is a JSON object. It contains the following fields:

- "paper": a link to the publication on arXiv,
- "classification": the classification of the publication,
- "formulas": a list of the first 10 MathML formulas in the publication as strings (short formulas are skipped).

The test dataset does not contain the "classification" and "paper" fields. Instead, it has an "id" that contains a unique identifier for the publication.

The data is available in the assignment repository [AR].

2.2 Evaluation

The results should be stored as a JSON object, where the keys are the "id"s and the values are the predicted classifications.

The assignment repository has an "example test dataset" together with a result file that contains the correct classifications. Additionally, it contains a script that you can use to evaluate your solutions for the example test dataset. Your grade will be based on the your results for the "real test dataset", but it should be very similar to the results for the example test dataset.

The grade is simply based on the accuracy of your classifications, i.e. the fraction of correctly classified publications.

3 What to submit

Your solution should be submitted to your repository for this assignment. It should contain:

1. all your code,
2. the proposed classifications for the test dataset (see Section 2.2),
3. a solution summary (see [SoS] for more details – it should describe the main ideas, not document the code).
4. a README.md file explaining
   i. dependencies (programming language, version, external libraries and how to get them),
   ii. how to run your code on different environments,
   iii. the repository structure,
   iv. anything else we should know.

Please **do not commit large files** (e.g. a large model file or a Python virtual environment) to the repository. Rather, tell us how to produce them in the README.md file.

4 Points

How many points you get depends on the accuracy of your model. Concretely, you get

$$\left\lceil 215 \left(a - \frac{1}{4} \right) \right\rceil$$

points where a is the accuracy according to the test dataset. As long as you get at least 1 point for the accuracy of your model, you can get up to 20 points for the quality of the submission (README, solution summary, ...).

The official number of points for this assignment is 100. If you get more than that, the extra points will be counted as bonus points.

If the grading scheme does not seem to work well, we might adjust it later on (likely in your favor).

References

[AG] *Guide for "Assignment 5: Classify Math Publications"*. URL: https://kwarc.info/teaching/AISysProj/WS2425/assignment-2.5-guide.pdf.

[AR] *Repository for Assignment 5: Classify Math Publications*. URL: https://gitlab.rrze.fau.de/wrv/AISysProj/ws2425/a2.5-classify-math-publications/assignment.

[arxiv] *arXiv*. URL: https://arxiv.org/ (visited on 01/24/2025).

[latexml] *LaTeXML*. URL: https://dlmf.nist.gov/LaTeXML/ (visited on 01/24/2025).

[mathml] *MathML*. URL: https://www.w3.org/Math/ (visited on 01/24/2025).

[SoS] *Solution Summary*. URL: https://gitlab.rrze.fau.de/wrv/AISysProj/admin/general/-/blob/main/solution-summary.md.