University of Aberdeen
School of Natural and Computing Sciences
Department of Computing Science
MSc in Artificial Intelligence
2024 - 2025

Assessment 1

CS552J - Data Mining with Deep Learning

This assignment is assessed in pairs and accounts for 50% of your total mark for the course.

Learning outcomes On successful completion of this component a student will have demonstrated competence in the following areas:

- ➔ Understanding of goals relating to different types of data mining techniques, ability to identify appropriate goals for extracting information from data sets, and ability to apply all this in practice.
- ➔ Understanding of key models that support data mining, and ability to use appropriate models in practice.
- ➔ Using a non-trivial dataset, plan, execute and evaluate significant experimental investigations using multiple data mining, and deep learning strategies

Information for Plagiarism and Conduct: Your submitted report and source code may be submitted for plagiarism check (e.g., Turnitin). Please refer to the slides available at MyAberdeen for more information about avoiding plagiarism before you start working on the assessment. Please also read the following information provided by the university: https://www.abdn.ac.uk/sls/online-resources/avoiding-plagiarism/

In addition, please familiarize yourselves with the following document code of practice on student discipline (Academic)

Application Problem Definition: Detecting human activities from still images

Detecting human activities from still images remains a relatively challenging area for computer vision. Your task is to make use of a small portion of specifically selected images from MSCOCO, a large resource of labeled photos used to train and evaluate many models in computer vision.
Your aim is to build a classifier that can successfully classify the specific images with the IDs provided in MyAberdeen according to which human action is taking place in the image (e.g. sitting, standing).
The dataset includes multiple activity labels and multiple images. The image_ids are provided along with the image urls that you can download from the MSCOCO website (full instructions in template code provided).

No prior knowledge of the domain problem is needed or assumed to fulfill the requirements of this assessment.

Report Guidance & Requirements

Your report must conform to the below structure and include the required content as outlined in each section. Each subtask has its own marks allocated. You must supply a written report (pdf), along with the

corresponding source code written in python (single, well-documented jupyter notebook), containing all distinct sections/subtasks that provide a full critical and reflective account of the processes undertaken. **All details and results must be included in the report**, your code is only for evidence to support the information in the report. The report should be written in a formal manner, and results clearly presented and rationale described. Report should be created using the latex template provided, and follow the style guidelines within.
The following task requires you to expand and elaborate upon the principles of data mining, different components of machine learning and some aspects on how such techniques can be used in real-life problems such as in text mining. Use the dataset provided as an example

Programming Task Report: (~max 2000 words/ roughly 4 pages)

The report should contain the *description of data and methods* (1) as well as the *results of the task* (2), reported clearly and concisely and contain the elements outlined in this section as well as the information addressing the subtasks below. All reports should make use of the latex template provided, and be submitted in pdf form.

1. Description of Data and Methods (10 marks)
Using your own words, the lecture material and any other relevant sources, explain specifically:

1. What the dataset consists of and cite its source
2. What basic preprocessing steps you have used to work with this data be needed to work with this data (e.g. tokenization, embeddings etc) and which specific libraries are used
3. What models did you use? Add appropriate citations for method reproducibility and to justify choices.
4. What specific challenges and trade-offs did you consider and why? Add motivating examples and citations to justify choices (e.g. summarisation, making use of transformers, dataset size)
5. Evaluation & Error analysis: How did you evaluate your methods? how did you examine errors? (e.g. metrics of fluency, faithfulness, or use of confusion matrices, precision, recall, F1 score, evaluating performance in different subsets of data)

2. Programming Task: Detect activity from still images using neural networks

The problem we aim at tackling has been clearly described and defined earlier. This task includes *six subtasks*, each of which bears its own marks.
Subtasks:

1. Develop and compare performance of CNN and FNN models trained for image classification on only the images provided. Call these CNN_base and FNN_base (10 marks)
2. Explore generalization techniques, such as data augmentation, weight decay, early stopping, ensembles, and dropout best generalization. Report new CNN model trained on this augmented dataset as CNN_gen. (10 marks)
3. Transfer learning: use a small pretrained Image classification model on MSCOCO from huggingface to develop a binary classifier for two of the action classes. (10 marks)
4. Multimodal embeddings1. Create a simple neural classifier to classify CLIP embeddings of the images for their label. (5 marks)
5. Multimodal embeddings 2. Create a classifier that uses the cosine distance between the embedding of the labels e.g. “reading” as a feature or uses the embedding of the image to derive other useful features (5 marks)

3 Bonus – Optional

Should you decide to try a bonus task, there will be a reward of *5 marks*. The maximum overall mark for this assessment remains at 50/50; however, attempting the bonus exercise will a) make you practice with an alternative distributed library and b) enhance your chances of getting a higher mark overall. To gain these marks, you will need to show you have synthesized the data mining issues covered in practicals and lectures and are able to take them further: adapting a model, adopting advanced visualization, coming up with something that is novel.

Possible tasks:

- Include custom model: create a 6th model using your own custom features. Report the accuracy and justify your design
- Visualisation: use an appropriate interpretability library to visualize where your models go wrong. What insights does this give? Other error analysis techniques can be used.
- Be creative: pick some aspect of the course that you think can apply to this dataset and try it. Be sure to report the motivation and method clearly, as well as the outcome.

Only attempt to gain bonus marks once you are satisfied you have met the criteria for the rest of the assessment to the best of your ability.

Marking Criteria

- ***Quality of the report***, including structure, clarity, and brevity: is your writing specific and to the point? *- please report word count.*
- ***Reproducibility***. Can another MSc AI student repeat your work based on your report and code?
- ***Quality of your experiments***, including design and result presentation (use of figures and tables for better reporting) Configured to complete the task and the parameter tuning process (if needed)
- ***In-depth analysis of results*** generated, including critical evaluation, insights into data, and significant conclusions
- ***Quality of the source code***, including the documentation of the code

Submission Instructions

You should submit your work via MyAberdeen by **23:59 25/04/2025**. Both the report and the code should be submitted together in the form of a **zipped folder**. The naming convention for the files should be as follows: CS552J_Assessment2_Lastname_Firstname_StudentNumber.zip

Include within Zipped folder

- ➔ **Report** The name of the PDF file should have the same naming convention: For instance, if I was a student with ID number 4568985, my submission file name would be: CS552J_Assessment2_Sinclair_Arabella_4568985.pdf.
- ➔ **Latex source** You should also include the latex source code as evidence you used the template provided to create your pdf.
- ➔ **Python Notebook** submit supplementary material containing the source code of your implementation (as a python notebook ".ipynb"). Your script should use markdown to describe clearly what your code does. It should follow the same naming convention as the other files.

Please **do not submit any training data** on MyAberdeen. Please try to make your submission file less than 20MB as you may have issues when uploading large files to MyAberdeen.

Any questions pertaining to any aspects of this assessment, please address them to the course coordinator Arabella Sinclair (arabella.sinclair@abdn.ac.uk)