COMP4318/5318 - Machine Learning and Data Mining

Assignment 2

Deadline: 11:59pm, 14 October, 2024 (Monday week 11, Sydney time)

You are required to work in groups of two or three students. You must register a group on Canvas (People → A2-Group).

1. Summary

The objective of this assignment is to apply your machine learning and data mining skills to solve practical problems classification, regression, clustering.

The code for this assignment should be written in Python in the Jupyter Notebook environment. Your implementation of the algorithms should use the same suite of libraries that we have used in the tutorials, such as Keras, Scikit-learn, Transformer, NumPy, and Pandas. In exceptional cases, you may use alternative libraries like PyTorch, but you need to justify your choice. Other libraries may be utilised for minor functionality such as pre-processing, plotting; however, please specify any dependencies at the beginning of your code submission.

The total score for this assignment is allocated as follows:

1. Code: max 40 points
2. Report: max 60 points

Detailed assignment specifications and scoring criteria can be found on the assignment page on Canvas (Assignments → Assignment 2 - Specification). The sections below provide comprehensive information on the assignment tasks and guidelines for submission.

2. Instruction

2.1 Dataset catalogue

In this assignment, you are given a collection of datasets with corresponding problem categories (classification, regression, clustering):

- **Classification**:

1. EMNIST handwritten character dataset:

   Original dataset and description: https://www.nist.gov/itl/products-and-services/emnist-dataset

   For this assignment, we provide a smaller subset of the EMNIST-ByClass dataset:

   https://drive.google.com/file/d/1qEoWitIzaRUWRFJsy7BKWZdvKecuck4X/view?usp=sharing

2. Sentiment140:

   Original dataset and description: https://www.kaggle.com/datasets/kazanova/sentiment140

   For this assignment, we provide a smaller subset:
   https://drive.google.com/file/d/1w9vysV8MIl_6LF26XFZlfCbyG2MbDj--/view?usp=sharing

- **Regression**:

3. Wiki Face Dataset:

   Original dataset and description: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

   For this assignment, we provide a smaller subset: https://drive.google.com/file/d/1Hx5-D8ZgdOFCSKeUtpFYg379xYmR1hwN/view?usp=sharing

4. Electricity Consumption:

   Original dataset and description (UCI ML repository):
   https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014

   For this assignment, we provide a small and clean version based on the raw data:
   https://drive.google.com/file/d/1GGN2EivWIcRJ4UTm8bPz819bddkGuxxO/view?usp=sharing

- **Clustering:**

5. Sales Transactions:

   Original dataset and description (UCI ML repository):
   https://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly

2.2 Assignment Task

a) Select **ONE** dataset from the list provided in Section 2.1. Read the description of the dataset and the associated task. For the implementation, use the versions of the datasets we provide, if applicable.

b) Implement **THREE** different **Deep Learning** (or **Clustering**) methods to address the task. Note that you are **NOT ALLOWED** to use models with **pre-trained weights**.

c) When designing and implementing these methods, you should consider the following aspects:

- Choose methods that are suitable for the dataset and its associated problem.
- Choose appropriate pre-processing techniques for the dataset (e.g., data normalisation, dimensionality reduction, etc.)
- Choose appropriate sets of hyperparameters to search (if applicable to your chosen method)

d) Fine-tune models to obtain their best performance and improve their generalization ability.

e) Conduct experiments to evaluate and compare models' performance. Present the experimental results and provide meaningful discussion and reflection.

f) Recommended evaluation metrics for each task:
- Classification: Using accuracy, precision, recall, and confusion matrix.
- Regression: Using Mean Square Error (MSE).
- Clustering: Using appropriate clustering evaluation methods such as silhouette coefficient.

3. Requirements

3.1 Code

As mentioned above, your code submission for the assignment should be formatted as a .ipynb Jupyter notebook. It should be organised with the following structure:

- **Setup:** Install and import necessary packages and libraries used in your coding environment. It is recommended to specify their versions to ensure reproducibility. Define any necessary utility or helper functions (e.g., for plotting, optimization, etc.) if applicable.
- **Data loading, pre-processing, and exploration:** begin by loading the dataset, then conduct some exploration of the data to better understand its properties and the preprocessing that may be appropriate. You may like to investigate the dataset to uncover patterns, outliers, and relationships before building models. You should include anything you feel is relevant in this section.

  Based on your insight from the data exploration and/or with reference to other sources, you should apply appropriate preprocessing techniques. Different preprocessing techniques might be applied to the different algorithms.
- **Model Implementation:** You will be required to design and implement at least three different models covered in the course. These models must represent different architectures to effectively explore their strengths and weaknesses with the selected dataset. For instance, a combination of DNN, LSTM, and Transformer would be appropriate, whereas variations like a 1-layer DNN, 3-layer DNN, and 5-layer DNN are not acceptable as they do not sufficiently differ in architecture. Implement an instance of each model before tuning hyperparameters and set up any functions you may require tuning hyperparameters in the next section.
- **Hyperparameter tuning:** Conduct a search over relevant hyperparameters for each model using a chosen strategy (e.g., grid search, random search). Please preserve the output of these cells in your submission and keep these hyperparameter search cells independent from the other cells of your notebook to avoid needing to rerun them, as markers will not be able to run this code (it will take too long), i.e. ensure the later cells can be run if these grid search cells are skipped.
- **Evaluation and Comparison:** After selecting the best hyperparameters, train each model using these settings in separate cells, independent of the tuning process. Use these models to evaluate and compare their performance on the test set with appropriate evaluation metrics. Conduct visualizations to effectively present and analyse the experimental results.
- **Final models:** Conclude the best classifier based on the comparative analysis.

Important Notes:

- *Your code must be easily readable, well-commented, and accompanied by explanatory text to clarify the purpose of each code cell.*

3.2 Report

The report must be structured similarly to a research paper, with the following key sections:

- **Abstract:** include a self-contained, short summary of your work.
- **Introduction:** introduce the dataset and the problem you have chosen, discuss its relevance to real-world applications, outline the previous techniques and approaches that have been utilized to solve the problem and provide an overview of the methods you used and the results you obtained.
- **Data:** describe the dataset and pre-processing, including:

- **Data description and exploration**: Describe the data, including all its important characteristics, such as the number of samples, classes, dimensions, and the original source of the images. Discuss your data exploration, including characteristics/difficulties as described in the relevant section above and anything you consider relevant. Where relevant, you may wish to include some sample images to aid this discussion.
- **Pre-processing:** Justify your choice of pre-processing either through your insights from the data exploration or with reference to other sources. Explain how the preprocessing techniques work, their effect/purpose, and any choices in their application. If you have not performed pre-processing or have intentionally omitted possible preprocessing techniques after consideration, justify these decisions.

➢ **Methodology:** describe the models that you employed in this assignment, including:

- **Theory**: For each model, explain the main theoretical ideas (this will be useful as a framework for comparing them in the rest of the report). Explain why you chose those models.
- **Strengths and weaknesses**: Describe the relative strengths and weaknesses of the models from a theory perspective. Consider factors such as performance, overfitting, runtime, number of parameters, and interpretability. Explain the reasons, e.g. don't simply state that CNNs perform better on images but explain why this is the case.
- **Architecture and hyperparameters**: State and explain the chosen architectures or other relevant design choices you made in your implementation (e.g. this is the place to discuss your neural network configurations). Describe the hyperparameters you will tune over, the values included in the search, and outline your search method. Briefly explain what each hyperparameter controls and the expected effect on the algorithm. For example, consider the effects of changing the learning rate, or changing the stride of a convolutional layer. Justify why you have made these choices.

➢ **Experimental Results:** present the experimental setting (e.g., the details of the dataset, models, hardware, and software specifications of the computer used for performance evaluations). Provide the experimental results obtained from the algorithms you implemented in an intuitive way. Discuss and compare the performance of your models. Consider factors such as accuracy, runtime, number of hyperparameters, and interpretability. Employ high-quality plots, figures, and tables to visually support and enhance the discussion of these results. **Please do not include screenshots of raw code outputs when presenting your results.**

➢ **Conclusion:** summarize your main findings, mention any limitations, methods, and results, and suggest potential directions for future works. Write one or two paragraphs describing the most important thing that you have learned while completing the assignment.

➢ **References:** include the references cited in your report in a consistent format. You may choose any appropriate academic referencing style, such as IEEE.

Important Notes

- *The report must be in **PDF format**. It should follow the format specified in the provided template, which includes a single-column layout, Times New Roman font, and a font size of 11.*
- *The maximum length of the report is **10 pages** (excluding appendix and references)*
- *You must include an appendix that clearly provides instructions on how to set up the environment to run your code, especially the installation guide and version of any external packages and libraries used for implementation. In addition, you should include the hardware configurations used for the coding environment.*

4. Submission

4.1 Group Registration:

For this assignment, you can work in groups of two or three. Please register your group under *People → Group → A2-Group* on Canvas. **The group registration should be done by Friday, Sep 30th, 2024**

4.2 Proceed to Canvas and upload all files separately, as follows:

a) **Report (one PDF file)**

b) **Code (one *.ipynb* file):** a Jupyter Notebook containing all your implementation. You are only allowed to use one *.ipynb* file.

c) **Code (one PDF file of *.ipynb* code):** The *.ipynb* code must also be exported to a PDF version.

Important:

- *Only **ONE** group member needs to submit the assignment on behalf of the group.*
- *Do **NOT** submit the dataset or zip files to Canvas. We will copy the data folder to the same directory as yours.ipynb file to run your code. Please make sure your code is able to read the dataset from this folder.*
- ***BOTH CODE AND REPORT** will be checked for plagiarism.*

4.3 File Naming Conventions

The submission files should be named with your group ID and all student IDs separated by the underscore (_). For example,

- *a2_groupID_SID1_SID2_SID3.ipynb* (code)
- *a2_groupID_SID1_SID2_SID3.pdf* (pdf version of the code)
- *a2_groupID_SID1_SID2_SID3_report.pdf* (report)

where SID1, SID2 and SID3 are the SIDs of the three students.

In both the Jupyter Notebook and report, include only your SIDs and not your name. The marking is anonymous.

4.4 Late Submission Penalties

A penalty of MINUS 5 percent (-5%) for each day after the due date.

The maximum delay for assignment submission is 3 (three) days, after which the assignment will not be accepted. You should upload your assignment at least half a day or one day prior to the submission deadline to avoid network congestion.

Canvas may not be able to handle a large number of submissions happening at the same time. If you submit your assignment at a time close to the deadline, a submission error may occur causing your submission to be considered late. Penalty will be applied to late submission regardless of issues.

4.5 Marking Rubric

Please refer to the *rubric* on Canvas (Canvas → Assignment 2 → Rubric) for a detailed marking scheme.

5. Academic honesty

Please read the University policy on Academic Honesty very carefully:

https://sydney.edu.au/students/academic-integrity.html

Plagiarism (copying from another student, website or other sources), making your work available to another student to copy, engaging another person to complete the assignments instead of you (for payment or not) are all examples of academic dishonesty. Note that when there is copying between students, both students are penalised – the student who copies and the student who makes his/her work available for copying. The University penalties are severe and include:

* a permanent record of academic dishonesty on your student file,

* mark deduction, ranging from 0 for the assignment to Fail for the course

* expulsion from the University and cancelling of your student visa.

In addition, the Australian Government passed a new legislation last year (Prohibiting Academic Cheating Services Bill) that makes it a criminal offence to provide or advertise academic cheating services - the provision or undertaking of work for students which forms a substantial part of a student's assessment task. Do not confuse legitimate co-operation and cheating!