COMP5318 - Machine Learning and Data Mining

Assignment 2

Deadline: 11:59pm, 19 May, 2023 (Friday week 12, Sydney time)

This assignment is worth 25% of your final mark. It consists of two components: code (10 marks) and report (15 marks).

Late submissions are allowed for up to 3 days late. A penalty of 5% per day late will apply. Assignments more than 3 days late will not be accepted (i.e., will get 0 marks). The day cut-off time is 11:59pm.

You are required to work in groups of two or three students. You must register a group on Canvas (People → Assignment 2).

1. Objective

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

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, NumPy, and Pandas. In exceptional cases, you may use alternative libraries like PyTorch instead of Keras, 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.

2. Instruction

2.1 Dataset

In this assignment, you are required to choose **one of the following datasets** and to implement **three Machine Learning algorithms** that we have studied in this course to solve the specific task of the dataset.

- **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

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

b) Implement three different Machine Learning methods to address the task.

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 set 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. 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.
- **Methodology:** describe the methods that you employed in this assignment, including pre-processing techniques and machine learning algorithms. Provide an explanation of the underlying theory behind each of them and discuss your design choices.
- **Experimental Results:** present the experimental setting (e.g., the details of 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, interpretability.
- **Conclusion:** summarize your main findings, mention any limitations methods and results and suggest potential directions for future works.
- **References:** include the references cited in your report in a consistent format.

- *The maximum length of the report is 15 pages.*
- *The report must be in PDF format. Make sure the report is well-structured, easy to read, and that it presents your findings in a logical and organized way.*

- *You must include an appendix that clearly provides the instructions on how to set up the environment to run your code (e.g., installation guide and version of any external packages and libraries used for implementation)*

4. Submission

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

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

b) **Code (one *.ipynb* file):** 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.

To generate a .pdf of your notebook for submission, please use File> Download as > PDF or Print Preview -> Save as PDF. Remember to submit both the .ipynb and the .pdf of your code.

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.*
- ***Both the code and report will be checked for plagiarism.***

4.2 File Naming Conventions

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

- *a2_groupID_SID1_SID2.ipynb* (code)
- *a2_groupID_SID1_SID2.pdf* (pdf version of the code)
- *a2_groupID_SID1_SID2_report.pdf* (report)

where SID1 and SID2 are the SIDs of the two students.

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

4.3 Marking Rubric

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

6. 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!