COMP5318 - Machine Learning and Data Mining

Assignment 2

Due date is available on Canvas

This assignment is to be completed in groups of 2 students. It is worth 20% of your total mark. You have to register your group on Canvas. In case you cannot find a group, please contact the teaching team for an arrangement or you will be assigned randomly to any available groups.

1. Objective

The objective of this assignment is to apply machine learning and data mining methods to solve a practical problem. You are required to design and implement at least three methods.

2. Instructions

2.1 Datasets

In this assignment, you can choose **one of the following datasets** and build models to solve the specific given task of the dataset.

Classification:

1. EMNIST-ByClass, https://www.nist.gov/itl/products-and-services/emnist-dataset
2. Sentiment140, https://www.kaggle.com/datasets/kazanova/sentiment140
3. UNSW-NB15, https://research.unsw.edu.au/projects/unsw-nb15-dataset

Regression:

4. Wiki Face Dataset, https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/
5. S&P 500 stock, https://www.kaggle.com/datasets/camnugent/sandp500

Clustering:

6. Sales Transactions, https://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly

Note that if the datasets are too big to run, you can consider doing some pre-processing of the datasets or use part of them to train. However, they should be clearly explained in your report.

2.2 Assignment tasks

a) Choose a dataset from the list provided in Section 2.1. For example, if you choose the EMNIST-Byclass dataset, you can only use it for classification.

b) Try at least 3 different Machine Learning methods and compare their performance. You should experiment and clearly discuss your design decision to help you achieve a higher performance and speed. The design options should consider the following aspects:

   - Choosing an appropriate model and its complexity
   - Using pre-processing techniques on the datasets (e.g., clustering, feature extraction, etc.)
   - Computer infrastructure (e.g., parallelizing, speeding-up your code, etc.)
   - Ease of prototyping (e.g., implementation approaches, choice of algorithms and libraries)

c) You are expected to fine tune each algorithm and explain why one approach outperforms the others.

d) Since you are expected to use more complex models that may not been discussed in the lectures, you can use most external open-source libraries such as: scikit-learn, pandas, Keras, Tensorflow, PyTorch, Theano, Caffe2, or their equivalent in Python 3 to write your own classifiers. If you want to use any other external libraries, please post on Ed for confirmation.

e) It is not allowed to use pre-trained models from open source. You need to retrain and save your own models.

f) **You are only allowed to use Python 3 on Jupyter Notebook in this assignment.**

3. Report

The report must be organised in a similar way to research papers, and include the following:

- The **abstract** includes a self-contained, short, and powerful statement that describes your completed work.
- The **introduction** section should present the dataset (problem) that you chose, discuss its relevance in diverse applications (the importance of the problem), and give an overview of the methods you used and the results you achieved.
- The **previous work** section shows successful techniques utilised on the same or similar datasets and how they are different to yours.
- The **methodology** section discussed the method**s** that you use in this assignment. Explain the theory behind each of them and discuss your design choices. This section should present at least pre-processing approaches and machine learning techniques used for solving the tasks.
- The **experiment** section displays results and comparisons for the implemented algorithms. Include runtime, hardware and software specifications of the computer that you used for performance evaluations. You are then expected to include meaningful comments on the results of your experiments and reflect on your design choices.
- The **conclusion** section sums up your results and outlines potential directions for future works.
- The **references** section includes all references cited in your report, formatted in a consistent way.

3.1 Evaluation metrics

Classification: Using accuracy, precision, recall and confusion matrix.

Regression: Using Mean Square Error (MSE).

Clustering: Using internal clustering validation methods such as Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index, etc.

3.2 Report layout

It is allowed to have any format, just need to follow the above structure.

Length: maximum 25 pages (ideally 10 to 15 pages) with [-10] penalty each additional page after 25.

4. Submission

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

a) Report (a PDF file)

   The report should include your group ID and each member's details (student ID and name).

   You must include an appendix that provides detailed steps on how to successfully run your code, including any external libraries installation required to be able to execute your code.

b) Code (.ipynb files)

   Your code should be written as one or more **.ipynb files**. You should separate the code file containing the algorithm and parameters that yield the best result from all the other algorithms, so in this case there would be 2 code files to submit.

   Another alternative is to have one code file for each method / algorithm, i.e. 3 code files for 3 algorithms, 1 file for each one.

   Note: Do **NOT** submit the dataset and zip file.

c) Code (PDF files of .ipynb code)

   Every .ipynb code file must be saved as a PDF document and included in your submission e.g. if there are 2 .ipynb code files, you should also submit 2 PDF documents, one for each corresponding .ipynb file.

d) Your trained model (.h5 or .pt)

   Your trained model must be submitted along with Code and Report to save the marking time. You only need to submit the best trained model to Canvas. Note that it is your own trained model, it is not the existing pre-trained model from other sources.

4.2 Only one student in your group needs to submit all the files and they must be named using your group ID separated by underscores. For example,

- group1_report.pdf
- group1_best_algorithm1.ipynb
- group1_other_algorithms.ipynb
- group1_best_algorithm1.pdf
- group1_other_algorithms.pdf
- group1_pretrained_model.h5 or group1_pretrained_model.pt

4.3 Your submission should include report and all the code files. A plagiarism checker will be used.

4.4 Clearly provide instructions on how to run your code in the appendix of the report.

4.5 Provide hyperlinks of the datasets you used, any external open-source libraries you used for the experiments and analysis, and versions of the libraries e.g., PyTorch 1.6.

4.6 **Indicate the contribution of each group member. The contribution will be taken into consideration for adjusting the mark of each member accordingly**.

4.7 A penalty of MINUS 5 percent per each day after the due date. The maximum delay is 5 days, after which the assignment submission will no longer be accepted.

4.8 **Please refer to the *rubric* in Canvas (Canvas → Assignment2 → Rubric) for detailed marking scheme. The report and the code are to be submitted in Canvas by the due date.**

5. Inquiries after releasing the marking

After Assignment 2 marks come out, please submit your inquiries about marking within one week. All inquiries after that will be ignored.

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!