Group Research Project

COMP/ENGN8535 Engineering Data Analytics - 2024

College of Engineering, Computing and Cybernetics
The Australian National University

Total Project Marks = 30%

Important Deadlines

1. Notify Course Convener of Group Members: 23:59 pm Monday 15 April 2024 (Start of Week 7).
2. Final Submission Due Date: 23:59 pm Sunday 26 May 2024 (End of Week 12).

Learning Objectives

One of the goals of COMP/ENGN8535 is to provide you with hands-on experience solving real-world problems using data analytics techniques, and prepare you for a career in data analysis, machine learning, artificial intelligence, mechatronics engineering, electrical engineering, signal processing, robotics, and related fields.

Task

You are to work in a group of ***at most 3 students*** to:

1. Identify a practical data analytics problem that can be solved using techniques we explore in ENGN/COMP 8535 including ranking (PageRank), recommendation (SVD), dimensionality reduction (PCA, Kernel PCA, MDS, IsoMap), clustering, or compressive sensing. Problems are provided below.
2. Design and implement a proposed solution to the problem in Python or MATLAB (preferably Python). You are allowed to use standard functions from common Python or MATLAB libraries (e.g., NumPy, SciPy, and SKiLearn) and do not have to implement your own functions from scratch. However, you should aim to demonstrate creativity and critical thinking when developing, implementing, justifying, analysing, and evaluating your method.
3. Test the proposed solution on a standard data set (or, as a last resort, simulated data if a standard data set is not available).
4. Compare the performance, computational effort or complexity, and difficulty of implementation with an alternative technique for solving the same problem (ideally a technique that has previously been proposed in a paper).
5. Write a report on the problem, proposed technique, the test performance of the proposed technique, and a comparison of the proposed technique to others in the literature in the form of a conference paper, as detailed below.

Group Requirements

You are to form your own groups of *at most 3 students*. You do not have to be enrolled in the same tutorials, computer labs, or degree. **One member from each group must email the course convener (Tim @ `timothy.molloy@anu.edu.au`) by the end of Monday 15 April 2024 with the names and uni-IDs of all group members.** If you have not found a group by Monday 15 April, you will be considered to be completing the project as an individual.

Project Report Requirements

The report is to be written in the NeurIPS conference paper template in either Word or LaTeX with a maximum of 7-pages A4 single column, inclusive of no more than half (1/2) a page of references. The templates are provided on Wattle.

Your final project report must be self-contained, as if it is a ready-to-submit conference paper. The content of the report should clearly describe, justify, and reflect on the project task. It should have a clear, logical technical report structure, and must contain the following sections (just like a regular research paper published in NeurIPS):

1. Project title, and the names (and uni-IDs) of all group members.
2. Introduction
    - Problem Statement, Motivation, and Background: What is the problem? Why is it important? How has it been solved in the past? Search for and discuss more than just the papers provided above.
    - Contribution Summary: What do you propose is the most appropriate analysis technique to solve the problem? Why not other approaches?
3. Problem Definition and Formulation - A mathematical description of the problem you are solving.
4. Method Description
    - Describe your proposed algorithm or method including its mathematical details.
    - Include key equations, pseudocode, discussion of key parameter, discussion of computational and memory complexity.
5. Experiments
    - Details of experimental setup (data used, any data preprocessing, key performance metrics that will be used to report results and their meaning, other algorithms compared to).
    - Details of experiments (how many were conducted, what does each experiment test or show).
    - Experimental results (quantitative presentation of results in graphs and tables)
    - Discussion of the experimental results (how your proposed algorithm performs in the experiments).
6. Conclusion (explain the point of your project and what conclusions you arrive at, and why or why not they make sense).
7. References (no more than half a page).

Also include a short abstract at the very start of the report summarising the problem, why it's important, which technique you investigated and why, and how it performed in your experiments. Your report should be free of errors in pagination, grammar and spelling.

Submission Instructions

- The report is to be submitted ***as a single pdf file*** through the Turnitin link on Wattle.
- All code for the simulations and algorithm implementations should be submitted as a single ZIP file through the link on Wattle.
- Please include the names of all group members on the report. Only one group member needs to submit the report and code.
- Late submissions are not permitted. Submissions after the due date without an extension will be awarded a mark of 0.

Marking Criteria

- Award level: already at the level of High Distinction, and more than that. The proposed method/algorithm is innovative, original, novel, or creative. The project report itself is almost ready to be published at a major national or international conference. The report formatting is already at NeurIPs level. OUTSTANDING ($\geq$90%).
- Something special; A real pleasure to read. A professional piece of work demonstrating obvious mastery of techniques. Almost flawless structure and implementation. Apparent that the supervisor would have gained a lot from the project. HIGH DISTINCTION (80%–90%).
- A worthy piece of work. The student has demonstrated an ability to manage, execute and document a significant piece of individual work. A good example of a research project report. Only flaws in this work are at the detail level. An easy read. DISTINCTION (70%–79%)
- Competently carried out a decent amount of technical work on a non-trivial project. Demonstrated some engineering nous in how things were done. Clearly structured report that conveys the work competently. CREDIT (60%–69%)
- Competently communicates what was done in the project. Reasonably achieved the basic goals of a basic project. No major conceptual flaws, but lots of room for improvement. Generally a poorly structured, poorly balanced (in terms of content) piece of work. PASS (50%–59%)
- Seems not to have understood the problem. Numerous glaring mistakes. No structure. A token effort. Unclear whether the author knew what this project was about. FAIL ($<$ 50%).

Reports longer than 7 pages or shorter than 6 pages will result in a 10-mark deduction (that is, your total mark will be reduced by 10).

Suggested Problems and Starting References

You may choose to work on one of the topics below.

Topic 1) BYO: Your own Data Analytics Project
You may propose to work on your own data analytics project (the topic must be on data analysis, or pattern recognition, or machine learning), however, in doing so, you must seek the course coordinator's written approval, and this approval must be obtained before the end of the second week of the teaching break (i.e. before Week 7 starts). This topic must also not be used for any of your other ANU courses.

Topic 2) MovieLens Movie Recommendation

- Project Idea:
Implement a system for movie recommendation via collaborative filtering techniques. The approach could be based on SVD or similar techniques such as nonnegative matrix factorisation (NMF). The performance of the proposed system should be evaluated using the MovieLens-100K dataset or similar.
  - If you want to implement an SVD-based recommendation, please read "Netflix Prize and SVD" by Stephen Gower at `http://buzzard.ups.edu/courses/2014spring/420projects/math420-UPS-spring-2014-gower-netflix-SVD.pdf`.
  - If you want to learn more about NMF, please read the following two papers: 1) "Algorithms for NMF. Algorithms for Non-negative Matrix Factorization" at `https://papers.nips.cc/paper_files/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html`, and 2) "Learning the parts of objects by non-negative matrix factorization" at `https://www.nature.com/articles/44565`

  After you complete the above movie recommendation task, you probably have gained a better understanding that certain movies can often be clustered together. You should then develop a data visualisation method to display a small subset (say 100 movies) of the 1700 movies on a 2-dimensional plane to visualise their similarity or dissimilarity. You may randomly select 100 movies of different genres, out of the 1700, for doing this experiment. To conduct this additional experiment, one (or more) of the following techniques may be needed: PCA, or MDS, or k-means. However, this is an open-ended task, there is no right or wrong solution as long as you project the 100 movies on a 2D plane, and convincingly visualise and explain their similarities.

- Dataset:
  The MovieLens-100k contains 100,000 ratings from 1000 users on 1700 movies. It can be found on Wattle.

Topic 3) IMDB Movie Recommendation System

- Project idea:
  As in Topic 2 above, your task is to implement a movie recommendation system and to produce additional visualisations. In contrast to Topic 2, you should use the IMDB dataset instead of the MovieLens-100K dataset.
- Dataset:
  The IMDB Dataset can be found on Wattle.

Topic 4) Image Segmentation via Spectral Clustering

- Project idea:
  Implement both:
  - Your own version of K-means clustering for colour-based K-class (K=4) image segmentation; and,
  - Your own version of Spectral clustering (aka. the Normalized-cut algorithm) for colour-based K-class (K=4) image segmentation.

  Test your segmentation algorithms on the horse, deer, and airplane images from the CIFAR-10 dataset and compare your segmentation results obtained with both algorithms.

  You may find the following materials useful:

  1 J. Shi and J. Malik, Normalized Cuts and Image Segmentation, Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 731-737, 1997.

  2 J. Shi and J. Malik, Normalized Cuts and Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, August 2000
- Dataset:
  CIFAR-10 dataset - available `https://www.cs.toronto.edu/~kriz/cifar.html`

Topic 5) Housing Price Prediction

- Project Idea:
  Develop a system for predicting the price of houses based on various factors like crime rate, number of rooms, etc.
- Dataset:
  A dataset of Boston house data is available at `https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`

Topic 6) Personality Prediction

- Project Idea:
  The Myers Briggs Type Indicator is a personality type system that divides a person into 16 distinct personalities based on introversion, intuition, thinking and perceiving capabilities. Develop a system for that classifies the personality of a person from the type of posts they put on social media.
- Dataset:
  A Personality Prediction Dataset is available at `https://www.kaggle.com/datasets/datasnaek/mbti-type`

Topic 7) Video Compression:

- Project Idea:
  Develop a novel (lossy) compression algorithm that takes in an image file and compresses it to at most 1% of its original size. Show (both mathematically and experimentally) that your proposed algorithm is optimal in the sense of minimising a suitable measure of reconstruction error.
- Dataset:
  Select a suitable subset of videos from the YouTube-UGC dataset at `https://media.withyoutube.com`